Points à examiner avant de grouper des données provenant de deux cycles différents de l'Enquête sociale générale

Michael Wendt[1]
27 février 2007

Résumé : Depuis un certain temps déjà, une série d'enquêtes annuelles indépendantes menées auprès des ménages permettent de recueillir des données dans le cadre du programme de l'Enquête sociale générale. Ainsi, un riche ensemble de données est maintenant à la disposition des chercheurs en sciences sociales et dans d'autres domaines. L'intégration des données de deux ou plusieurs cycles pourrait donc être un outil utile. À cette fin, il faudrait utiliser les données de deux ou plusieurs cycles pour estimer les entités d'intérêt. Pour ce faire, il y a deux façons de procéder (qui, en général, donnent des réponses différentes). La première consiste à calculer des estimations distinctes selon le cycle et de les combiner. La deuxième consiste à grouper simplement les ensembles de données de différents cycles et à calculer les estimations à partir des données groupées. Dans le présent document, nous décrivons brièvement des situations dans lesquelles il convient d'utiliser la deuxième méthode, celle du groupement, et nous exposons certains éléments à prendre en considération par le chercheur qui souhaite avoir recours à cette méthode. En outre, nous présentons plusieurs options de base pratiques en ce qui concerne la marche à suivre par le chercheur pour procéder à l'intégration. La table des matières ci-dessous donne un aperçu général des points à examiner et des marches à suivre. Essentiellement, le chercheur doit décider si les deux ensembles de données *devraient* et *peuvent* être groupés aux fins de son projet d'analyse.

Table des matières :

[1] Division de la statistique sociale et autochtone, Statistique Canada, michael.wendt@statcan.ca

Introduction : L'Enquête sociale générale (ESG) de Statistique Canada est une enquête-ménages annuelle transversale qui permet de recueillir des données sociales sur les Canadiens adultes depuis 1985. La dernière diffusion de données, celle pour l'ESG-20, a eu lieu en juin 2007.

Les données complètes recueillies au cours de 20 cycles offrent une foule de renseignements qui peuvent être analysés de différentes façons pour tirer parti de ces cycles différents. Par exemple, un projet d'harmonisation des variables sur les 20 cycles de données est en cours. Le but visé est de créer une série de 20 ensembles de données indépendants au format de fichier cohérent, dans lesquels les noms et les formats des variables sont les mêmes et les définitions des variables sont comparables. En outre, les noms des variables de pondération seront harmonisés et des poids bootstrap seront fournis pour l'estimation de la variance. Cette série d'ensembles de données serait utile notamment en ce qu'elle aiderait les chercheurs à faire le suivi des caractéristiques des Canadiens au fil du temps en calculant des estimations individuelles (annuelles) et en observant les séries chronologiques.

Outre de tels projets à grande échelle, les chercheurs procèdent de plus en plus à des comparaisons de l'information provenant de deux ou plusieurs cycles différents pour divers types d'analyse. On pense immédiatement à deux types de situation. Dans le premier, les chercheurs utilisent des cycles dont les thèmes sont similaires (habituellement repris tous les cinq ans) pour créer des séries chronologiques *plus dispersées* d'estimations distinctes ou bien utilisent plusieurs cycles adjacents pour créer des séries chronologiques *plus courtes*. Il importe de souligner que différents cycles de l'ESG représentent différents moments de l'évolution de la population canadienne. Ainsi, dans de nombreux cas, la seule option possible consiste à calculer des estimations portant sur un seul cycle à la fois et à suivre une série chronologique ou à faire des comparaisons entre « alors » et « maintenant ». Le présent document ne porte pas sur ce type d'analyse.

Dans un deuxième type de situation, les chercheurs souhaitent combiner directement des données de deux ou plusieurs cycles de manière à produire une estimation *intégrée* en quelque sorte, fondée sur les cycles en question. Le plus souvent, les chercheurs souhaitent intégrer des cycles distincts pour accroître les tailles d'échantillon pour les petits domaines. Pour deux cycles, par exemple, on part du principe que petit échantillon + petit échantillon = échantillon suffisamment grand pour permettre une analyse utile. Si l'échantillon est suffisamment grand, alors les chercheurs normalement étudient les caractéristiques dans le domaine intégré. Prenons, par exemple, le cas de personnes qui déclarent être « membres d'une minorité visible ». Ces répondants peuvent être peu nombreux dans les échantillons des deux cycles distincts, de sorte que, pour des raisons de confidentialité ou de grande variabilité, les données sur les caractéristiques du groupe ne pourraient pas être publiées séparément. Cependant, en combinant les données, on espère obtenir un échantillon suffisamment grand pour permettre de procéder à certaines analyses. Toutefois, il faut d'abord déterminer s'il est logique de combiner les données. Si la taille ou la composition de la population de minorités visibles a changé d'une année à l'autre, quel est le sens d'une estimation combinée? D'ailleurs, si un

changement important s'est produit, sur quelle population de domaine cible l'estimation porte-t-elle? Ce sont là d'importantes considérations pour les chercheurs. Étant donné que le groupement des données, malgré ses limites, est une méthode utilisée couramment par les chercheurs aujourd'hui, nous donnons ici des conseils sur les mesures à prendre avant d'utiliser cette méthode.

Le présent document porte plus particulièrement sur la combinaison de données provenant de cycles distincts. Même si nous faisons une mise en garde en signalant que, dans de nombreux cas, on ne devrait pas procéder à cette intégration (voir les détails ci-dessous), nous supposons que le chercheur souhaite tâcher d'intégrer des données de deux[2] cycles différents afin d'accroître la taille de l'échantillon d'un petit domaine. Nous dressons une liste des éléments à prendre en considération par les chercheurs pour décider s'il convient d'utiliser une méthode d'intégration donnée, particuliers au contexte de l'ESG. Nous exposons plusieurs marches à suivre pour procéder à cette intégration si elle est appropriée, et nous formulons plusieurs suggestions si elle ne l'est pas. L'expression clé est celle utilisée dans le paragraphe précédent : ***analyse utile.*** C'est l'aspect qui oriente les conseils sur l'intégration des données.

De nombreux articles ont déjà été publiés sur la façon de procéder pour intégrer différentes enquêtes, notamment [Binder et Roberts], [Korn et Graubard] et [Thomas]. On trouvera une série d'exemples se rapportant aux analyses dans le domaine de la santé dans [Schenker et Raghunathan]. Dans le contexte de l'ESG, un aperçu de la méthode utilisée dans un cas particulier où les données de l'ESG-13 et de l'ESG-18 ont été combinées pour établir les taux de victimisation (criminalité) chez les minorités visibles est fourni dans [Marchand]. Le présent document ne vise pas à faire un survol des méthodes d'intégration; plutôt, il se veut d'utilité pratique pour le chercheur qui souhaite s'attaquer à une tâche donnée. Il s'agit d'exposer plusieurs marches à suivre simples, applicables dans de nombreux cas utiles dans le contexte de l'ESG.

Approche individuelle et approche groupée : Il y a deux façons de calculer les estimations pour une combinaison de deux cycles dont on suppose que la population cible est finie[3] (comme dans le cas des statistiques descriptives comme les moyennes et les proportions, par exemple). Premièrement, on pourrait calculer des estimations distinctes pour chaque cycle et les combiner ensuite par une moyenne pondérée (c'est ce qu'on appelle l'*approche individuelle*). La deuxième façon de combiner les cycles consiste à grouper les données, à ajuster les poids de sondage et à continuer comme si l'échantillon combiné était simplement un échantillon plus grand (il s'agit de l'*approche groupée*). Dans certains cas, l'une ou l'autre approche donnera une estimation non biaisée de la population mais, en général, l'approche individuelle et l'approche groupée donneront des estimations différentes et, éventuellement, des interprétations différentes. On conseille donc au chercheur d'examiner *d'abord* la possibilité d'adopter l'approche individuelle. On devrait adopter l'approche groupée *seulement* si l'on peut supposer que les

[2] Les éléments à prendre en considération au moment de combiner trois cycles ou plus sont semblables, mais il faut procéder avec plus de soin.

[3] Une autre situation se présente lorsqu'il s'agit d'estimer les paramètres d'un modèle; voir le point à examiner 1 ci-dessous.

caractéristiques ainsi que les domaines d'intérêt sont similaires d'un cycle à l'autre[4]. Ainsi, nous exposons dans le présent document de nombreux points à examiner pour déterminer si ces hypothèses sont vérifiées pour le projet que le chercheur entend mener.

Il y a de nombreux éléments à prendre en considération *avant* de procéder au groupement des données pour calculer des estimations groupées. Ils constituent en quelque sorte un processus itératif : déterminer pourquoi l'intégration est souhaitable, vérifier qu'elle est logique, intégrer les données, calculer les estimations ou procéder à des analyses, vérifier les résultats, est-il encore logique de procéder à l'intégration?, répéter.

On recommande au chercheur de consulter un méthodologiste d'enquête à ce sujet, si possible, puisque chaque projet de groupement de données peut soulever des questions méthodologiques différentes. La première tâche à accomplir dans un projet de groupement des données dépend du point de vue de l'intéressé. Un analyste peut commencer par poser des questions comme « mes échantillons sont-ils les mêmes? » et « les questionnaires sont-ils les mêmes, de sorte que mes variables ont le même sens? », tandis qu'un méthodologiste peut d'abord poser la question suivante : « nous voulons estimer la moyenne de population pour cette variable donnée; sur quelle population portera l'estimation? ». Nous tâchons de combiner les deux points de vue et c'est dans ce contexte que nous avons établi la liste des points à examiner et des marches à suivre.

Point à examiner 1 : Quel type d'estimation veut-on produire? Nous avons souligné qu'on souhaite produire une estimation ou un ensemble d'estimations *utile*. Essentiellement, le présent document porte sur le fait que le chercheur doit décider pourquoi il veut intégrer différents cycles et si les estimations calculées d'après les données combinées se prêtent à une interprétation.

En intégrant les données de deux ou plusieurs cycles afin d'accroître les tailles d'échantillon pour les petits domaines, les chercheurs généralement visent deux objectifs : estimer les moyennes de population ou les proportions[5] pour diverses caractéristiques et estimer des paramètres du modèle comme les coefficients dans une régression linéaire ou logistique. Nous appelons ces deux objectifs, respectivement, *analyse descriptive* et *modélisation*. Le type d'analyse souhaité détermine certains aspects méthodologiques, comme la population d'intérêt, mais les éléments à prendre en compte ci-dessous pourraient s'appliquer à l'un et à l'autre.

Souvent, aux fins d'un projet d'analyse, il faut établir en même temps de nombreuses estimations descriptives et de nombreuses estimations des paramètres du modèle. Toutefois, on avertit le chercheur qu'il devra peut-être procéder de façon différente pour produire différentes estimations[6]. En outre, il est recommandé de limiter le nombre de variables et la portée du projet, particulièrement étant donné le point à examiner 4 ci-dessous, qui expose les tâches détaillées à accomplir pour tâcher d'harmoniser les deux

[4] En outre, s'il s'agit de produire une estimation composite, il convient d'adopter l'approche individuelle.
[5] L'estimation du total d'un domaine fondé sur deux cycles ne se prête pas vraiment à une interprétation utile.
[6] Dans le cadre du processus itératif mentionné ci-dessus.

ensembles de variables. Le processus d'harmonisation habituellement prend le plus de temps.

Des cycles adjacents de l'ESG ne se chevauchent pas et ne sont pas complètement statistiquement indépendants. Pour réduire le fardeau du répondant, les numéros de téléphone sélectionnés dans un cycle sont exclus des cycles futurs pendant deux ans. Ainsi, le deuxième échantillon dépend du premier. Toutefois, la possibilité d'un chevauchement est extrêmement rare et il serait acceptable de procéder comme si les deux cycles étaient simplement deux échantillons indépendants de la population canadienne en différentes années. Dans le cas de cycles réalisés à plusieurs années d'intervalle, il est concevable qu'un répondant participe à l'un et l'autre cycle mais, de nouveau, cela est très peu probable[7]. L'hypothèse de données non chevauchantes permet d'utiliser une méthode beaucoup plus facile de combinaison de deux cycles et l'hypothèse de données indépendantes permet d'utiliser une méthode beaucoup plus facile de calcul d'estimations combinées. Il y a lieu de souligner, toutefois, que l'ESG-20 et l'ESG-21 constitueront des exceptions, puisque certains des répondants de l'ESG-21 ont également participé (sélectionnés à dessein) à l'ESG-20.

Lorsqu'il s'agit de produire des estimations descriptives, la population cible est considérée comme étant finie. Comme nous l'avons mentionné ci-dessus, l'approche groupée et l'approche individuelle donneront des réponses différentes. Il convient d'examiner la possibilité d'utiliser l'approche individuelle si la valeur des caractéristiques a changé d'un cycle à l'autre ou si la taille des domaines d'intérêt a beaucoup changé.

Selon l'approche groupée, toutefois, on établit les estimations groupées en concaténant d'abord les deux ensembles de données distincts, en ajustant les poids et en procédant au calcul comme si le grand échantillon unique représentait « la » population. D'autres détails sont fournis ci-dessous.

Une autre situation se présente lorsque le chercheur souhaite estimer les paramètres du modèle (les coefficients dans un modèle de régression linéaire, par exemple). En pareil cas, le modèle statistique décrit une population infinie et on peut supposer que le modèle a produit les valeurs (pour les variables en question) de chacune des populations finies ciblées dans les deux cycles. Lorsqu'on utilise des données groupées pour estimer des paramètres, il est utile d'ajouter un « effet du cycle » au modèle. Cela permet de vérifier la présence d'inégalités dans les paramètres chez les deux populations finies supposées avoir été générées au moyen du modèle. L'exemple ci-dessous éclaircit ce point :

> Exemple : Le modèle linéaire $Y = \beta_0 + \beta_1 X + \varepsilon$ décrit une relation (théorique) entre X et Y et suppose qu'on souhaite estimer β_0 et β_1. Nous supposons, pour le premier cycle, que le modèle appliqué à X_i, pour les personnes faisant partie de la première population, produirait Y_i (et nous pouvons procéder à une

[7] L'auteur a une fois exécuté un test pour déterminer si des numéros de téléphone avaient fait partie de deux échantillons séparés par plusieurs années et n'en a trouvé aucun. Le test n'était pas exhaustif, mais il a confirmé empiriquement la probabilité extrêmement faible qu'un répondant participe à deux cycles différents.

estimation fondée sur le premier échantillon). De même, pour le deuxième cycle, nous estimons les paramètres pour la deuxième population en utilisant le deuxième échantillon. Lorsque les données sont groupées, nous pouvons estimer le même modèle en utilisant l'échantillon plus grand, mais nous devrions d'abord examiner le modèle $Y = \beta_0 + \beta_1 X + \beta_2 I_{\text{cycle } 1} + \varepsilon$, où I est un indicateur pour le premier cycle, mettons, et tâcher de déterminer l'*effet du cycle*.

Le principal avantage de l'approche groupée tient à ce qu'une fois un ensemble de poids appropriés trouvé pour l'échantillon groupé, ces poids peuvent être appliqués à de nombreuses estimations différentes[8]. L'inconvénient est qu'il n'est pas toujours possible d'établir une variance minimale estimée pour toutes les estimations.

Point à examiner 2 : Quelle est la population cible? Le chercheur qui utilise des données d'enquête doit définir leur population cible. Des détails sont fournis dans la documentation de l'enquête. L'ESG est généralement réalisée auprès des Canadiens vivant dans les dix provinces, âgés de plus de 15 ans, et ne vivant pas en établissement, dans les réserves ou dans les bases militaires. Lorsqu'on combine les données groupées de différentes enquêtes, la première question qui se pose est la suivante : les échantillons étaient-ils censés représenter les mêmes populations cibles? Dans l'ESG, cela est presque toujours le cas[9]. Le chercheur doit ensuite se poser une question d'ordre *conceptuel* : quelle population cible l'échantillon groupé représente-t-il? Nous utilisons le terme « conceptuel » parce qu'il n'y a pas de réponse *statistique* ou *méthodologique* fixe. En outre, il est impossible de répondre vraiment à la question avant de tenir compte de certains des autres éléments à prendre en considération exposés ci-dessous. Nous rappelons ici au chercheur que, s'il constate que l'approche groupée n'est pas appropriée statistiquement, alors l'échantillon groupé ne représente pas vraiment une population concrète et il y a peut-être lieu d'avoir recours à une autre méthode ou à une autre source de données.

Malgré cette mise en garde, on pourrait toutefois considérer les deux échantillons comme constituant simplement deux occurrences du processus d'échantillonnage à deux points différents dans le temps. D'ailleurs, toute la question du groupement de différents cycles de l'ESG porte sur la question de savoir si le temps importe ou non. Au moment de combiner deux cycles adjacents, il semble approprié de supposer que la population *réelle* de Canadiens adultes n'a pas beaucoup changé d'une année à l'autre[10]. Toutefois, en ce qui concerne les caractéristiques étudiées *ou* les facteurs

[8] Comme nous l'avons indiqué ci-dessus, la plupart des chercheurs ne calculent pas une seule estimation ou même un seul type d'estimation. Par exemple, un chercheur peut vouloir estimer des totaux, des proportions et des coefficients de régression *dans le cadre* de la même analyse.

[9] Dans l'ESG-16 (et dans l'ESG-21, maintenant aux étapes finales de la collecte), la population cible était composée de personnes de 45 ans et plus, vivant dans les provinces, etc.

[10] En fait, dans les cycles récents de l'ESG, on a procédé à l'échantillonnage en vagues mensuelles au cours de la plus grande partie de l'année, de sorte que quelques semaines seulement peuvent s'écouler entre la fin de la collecte d'un cycle et le début de la collecte du cycle suivant. D'ailleurs, le processus de pondération (de l'enquête) suppose que les données de l'ESG sont recueillies dans le cadre de 12 (environ) enquêtes mensuelles indépendantes.

inconnus liés aux cycles réalisés à cinq ans d'intervalle, il faut procéder avec plus grand soin pour s'assurer que les éléments étudiés n'ont pas changé d'une façon critique. D'ailleurs, comme nous l'avons mentionné, cette *hypothèse doit être vérifiée de façon rigoureuse* et les détails sont précisés au point à examiner 5 ci-dessous.

Point à examiner 3 : Les échantillons sont-ils comparables? Ensuite, le chercheur doit poser la question suivante : est-il logique de *tâcher* de combiner ces ensembles de données? Dans le cas d'un projet de groupement des données de l'ESG, nous voulons dire par cela : les échantillons ont-ils été tirés de la même façon? Représentent-ils réellement la même population de la même façon, à différents points dans le temps seulement?

Dans le cas de l'ESG, la réponse est habituellement que les échantillons pour différents cycles sont tirés de la même façon et selon le même plan d'échantillonnage (ou un plan similaire) et que les données sont recueillies au moyen des mêmes méthodes de collecte. Toutefois, les méthodes ont évolué de façon incrémentielle au fil des ans. Par exemple, on fait au chercheur une mise en garde au sujet des divers échantillons supplémentaires de l'Enquête sur la population active utilisés pour certains cycles précédents. Autre exemple, les données de l'ESG-16 ont été recueillies d'une façon différente de celle utilisée pour d'autres cycles, pour lesquels on a habituellement recours à la méthode de composition aléatoire. De plus, les tailles d'échantillon de l'ESG sont passées d'environ 10 000 personnes et plus jusqu'à l'ESG-12 à environ 25 000 personnes depuis l'ESG-13.

Ces mises en garde prises en compte, les plans d'échantillonnage des différents cycles ne diffèrent pas beaucoup d'une manière qui aurait une incidence sur la plupart des estimations (d'ailleurs, nous savons par expérience que d'autres facteurs[11] ont une incidence beaucoup plus importante que celle du plan sur les estimations). Plus particulièrement, les poids de sondage et les poids bootstrap sont souvent assez comparables en ce qu'ils représentent un répondant de la même façon dans les populations respectives. Le lecteur trouvera des détails sur les poids et les tailles d'échantillon à l'annexe. En outre, il est utile de faire une lecture attentive de la section sur la méthodologie des guides de l'utilisateur du fichier de microdonnées à grande diffusion respectifs.

Il convient de se poser également une autre question : le groupement est-il logique du point de vue analytique? Autrement dit, les mêmes concepts ont-ils été mesurés de la même façon, globablement? Par exemple, dans le cadre de l'Enquête sociale générale, on recueille parfois de l'information sur des thèmes auprès de différentes personnes à intervalles de cinq ans. On prend grand soin de mesurer des thèmes comparables de façons semblables MAIS on prévient le chercheur que des changements ont eu lieu. De nouveau, une lecture attentive des deux guides de l'utilisateur du FMGD est importante; dans ce cas, il faut tenir compte des détails fournis au sujet des concepts.

[11] Par exemple, le point à examiner 4 ci-dessous porte sur la façon dont les concepts sont mesurés par les variables.

Le chercheur doit se poser une troisième question sur la comparabilité globale des ensembles de données, à savoir : le groupement est-il logique du point de vue pratique? Le nombre de variables comparables en commun sera-t-il suffisant pour permettre de faire l'analyse souhaitée? On a souvent recours au groupement pour accroître les tailles d'échantillon pour les petits domaines. Cela peut être utile aux fins d'une analyse de base, mais l'exécution de modèles multivariés complexes peut aller à l'encontre du but du groupement, puisque plus le nombre de paramètres à estimer est grand, plus il faut de degrés de liberté.

Pour le point à examiner 3, l'analyste doit faire preuve de bons sens et s'appuyer sur son expérience. Comme nous l'avons mentionné, les questions soulevées ici peuvent constituer un processus itératif comprenant les étapes suivantes :

Point à examiner 4 : Les deux ensembles de variables sont-ils similaires? Dans tout projet de groupement de données, l'harmonisation des variables est la tâche qui prend probablement le plus de temps. D'ailleurs, il faut vérifier toutes les variables devant être utilisées directement ou indirectement[12] dans l'analyse pour déterminer si elles mesurent le même élément et si elles ont été mesurées de la même façon. Il faut procéder à cette vérification variable par variable.

L'analyste doit d'abord s'assurer que, dans le cas des deux fichiers, les noms et les formats sont les mêmes. S'ils ne le sont pas, il faut les rapprocher. Les livres de codes respectifs sont un bon point de départ. Il faut s'assurer que chaque variable catégorique porte sur les mêmes catégories ou que les deux ensembles peuvent être regroupés en catégories similaires. On fait à l'analyste une mise en garde particulière au sujet des catégories « non déclaré » et « ne sais pas », puisqu'elles changent parfois au fil des ans. De même, pour chaque variable continue, l'analyste doit accorder une attention particulière aux codes en dehors de la fourchette normale de valeurs qui signifient, par exemple, « non déclaré », « ne sais pas » ou même « 10 ou plus ».

Parfois, des changements subtils ont été apportés à la formulation d'une question. Il faut examiner ces cas pour s'assurer que le concept mesuré dans les deux ensembles de données est comparable. Cela peut être assez complexe. Par exemple, la définition de ce qui constitue un crime avec violence a évolué au fil des ans en fonction des modifications apportées aux politiques. L'analyste qui étudie les crimes de violence doit s'assurer que les définitions conviennent à son travail.

Le questionnaire de l'ESG est assez complexe (ce qui est transparent pour le répondant, puisque les données sont maintenant recueillies au moyen d'une application d'interview téléphonique assistée par ordinateur). De temps à autre, il est nécessaire de modifier l'enchaînement ou la position de la question dans le questionnaire. Cela pourrait avoir un effet sur quelle question à poser et à quel répondant. Dans le livre de codes pour chaque cycle de l'ESG, on trouve au bas de chaque variable une brève description des répondants auxquels on a posé la question particulière.

[12] Cela veut dire les variables de poids, par exemple.

Enfin, l'analyste doit savoir que le thème d'un cycle donné peut avoir un effet sur la façon dont les répondants répondent. Par exemple, si le thème est « la santé » et si on vient de poser au répondant plusieurs questions détaillées au sujet de la santé et que le répondant pense à sa santé, il peut répondre de façon différente à la question « Comment décririez-vous votre état de santé général? » que s'il s'agissait d'un cycle portant sur « l'emploi du temps ».

Pour certains cycles portant sur le même thème, il existe un tableau détaillé des concordances des variables des cycles.

Une fois les variables harmonisées, une bonne façon de procéder à la vérification de l'*assurance de la qualité* consiste à calculer une totalisation croisée de la variable selon le cycle, pour les variables catégoriques, ou un résumé de cinq chiffres[13] selon le cycle, pour les variables continues. Des chiffres similaires pour une catégorie dans deux cycles ne garantissent pas la similarité des deux variables, mais des chiffres différents peuvent être un indicateur de problèmes éventuels et devraient être examinés.

La discussion ci-dessus peut être résumée comme suit :

> **Marche à suivre 1 : Liste de contrôle pour l'harmonisation des variables :** Ce qui suit est une liste de contrôle des éléments à vérifier ou à modifier pour faire en sorte que les deux variables soient comparables (ou pour déterminer si un facteur quelconque dans la collecte peut avoir une incidence sur l'analyse) :
>
> - les noms sont-ils les mêmes?
> - les formats sont-ils les mêmes?
> - les catégories sont-elles les mêmes ou peuvent-elles être regroupées en catégories similaires?
> - les valeurs « sans objet », « non déclarée » et « ne sais pas » sont-elles les mêmes?
> - les deux questions sont-elles les mêmes?
> - les deux enchaînements des questions sont-ils les mêmes?
> - la position de la question est-elle la même ou semblable?
> - y a-t-il d'autres éléments à prendre en considération, comme le type de thème?
> - comme vérification finale, procéder à une totalisation croisée (pondérée et non pondérée) ou à un résumé de cinq chiffres pour chaque cycle

Point à examiner 5 : Les deux ensembles d'estimations sont-ils similaires? Pour pouvoir combiner les variables en un ensemble de données groupées, les variables doivent être similaires, tel qu'indiqué au point à examiner 4 ci-dessus. TOUTEFOIS, afin de calculer une estimation combinée utile, les *estimations* des deux cycles doivent être similaires.

[13] En fait, on pourrait comparer la moyenne et les écarts-types ainsi que le résumé de cinq chiffres : minimum, premier quartile, médian, troisième quartile, maximum.

Les tableaux statistiques pondérés et non pondérés ou les résumés de cinq chiffres à la fin de la marche à suivre 2 ci-dessus peuvent également servir à une vérification rapide pour déterminer si les fréquences estimées (moyennes, etc.) s'apparentent.

En outre, le chercheur devrait procéder à un test d'hypothèse en bonne et due forme en utilisant chaque paire[14] d'estimations pour déterminer si les deux paramètres de *population* présentent ou ne présentent pas des différences statistiquement significatives. À cette étape, il est nécessaire de grouper les données.

> **Marche à suivre 2 : Comment combiner les données?** Le groupement des données de deux cycles différents de l'ESG est simple, une fois les variables comparables désignées. L'ensemble de données groupées se compose simplement des deux ensembles concaténés, l'un *au-dessus* de l'autre. Par exemple, en code SAS, cela pourrait être :
>
> ```
> data pooled_data;
>     set data_first_year(in = ina) data_second_year(in = inb);
>     indicator_first = 0;
>     indicator_second = 0;
>     if ina then indicator_first = 1;
>     if inb then indicator_second = 1;
>
> run;
> ```
>
> Il est utile de créer deux variables indicatrices, une pour la source de chaque ensemble de données. On peut obtenir les estimations individuelles en multipliant les poids par l'indicateur respectif.
>
> Nous supposons que les noms des variables sont exactement les mêmes. SAS ne se plaindra pas si les noms des variables sont différents d'un cycle à l'autre. Il créera simplement deux variables dont les valeurs seront manquantes dans les parties opposées. Toutefois, SAS se plaindra si les variables dans les deux cycles sont de formats différents.
>
> Les deux ensembles de données d'entrée devraient contenir toutes les variables d'intérêt et les poids, y compris les poids bootstrap pour l'un et l'autre ensemble de données. Les poids sont généralement comparables (il faut faire attention aux poids au niveau de la personne par opposition à ceux au niveau de l'incident ou du ménage; le poids correct dépend du type d'analyse) de même que les poids bootstrap, de sorte qu'il suffit souvent de les renommer. Toutefois, il faut faire une lecture attentive du processus de pondération dans la documentation du

[14] Strictement parlant, on pourrait également comparer la distribution bivariée de deux variables dans le premier cycle et la distribution bivariée des deux variables correspondantes dans le deuxième cycle (ainsi que, de façon plus générale, les distributions multivariées). À un moment donné, toutefois, le chercheur doit décider de ce qui est pratique pour une analyse, étant donné les tailles des cellules de plus en plus petites dans les totalisations croisées multidimensionnelles.

FMGD (particulièrement pour déterminer quels poids correspondent à quel concept).

Une fois les données groupées, on peut procéder à des tests de vérification d'hypothèses pour déterminer si les valeurs de population respectives sont ou ne sont pas statistiquement différentes. À cette fin, il faut procéder à une estimation de la variance, ce qu'on peut faire au moyen des poids bootstrap. Bootvar, Stata et SUDAAN sont tous capables de calculer des estimations de la variance utiles à l'aide de poids bootstrap pour de nombreux types d'estimations. Il y a de nombreuses façons d'effectuer des tests par paire dans SUDAAN. La marche à suivre ci-dessous est l'une de plusieurs méthodes. Pour les grands projets, on invite le chercheur à optimiser son code car le temps d'exécution pourrait être long.

Marche à suivre 3 : Comment comparer des estimations? Aux fins d'exemple, la variable_1 est supposée être catégorique et à trois catégories, soit value_1, value_2 et value_3. Le poids au niveau de la personne est supposé être wght_per et les poids bootstrap sont supposés être wtbs_001 à wtbs_200. Nous vérifierons si la valeur value_1 est ou n'est pas (significativement) différente selon le premier ensemble de données et selon le deuxième ensemble de données.

```
/* create an indicator variable for values of variable_1 = value_1 */
/* create a variable called one, which is constantly = 1, for denominator */
/* of ratio estimates */

data pooled_data;
        set pooled_data;
        indicator_value_1 = 0;
        if variable_1 = value_1 then indicator_value_1 = 1;

/* this computes two ratio estimates value_1 in data set 1 / one and */
/* value_1 in data set 2 / one */

proc ratio data = pooled_data design = brr;
        class indicator_first;
        weight wght_per;
        denom one;
        numer indicator_value_1;
        repwgt wtbs_001 - wtbs_200 / adjfay = 25;

run;

/* this produces a test statistic for */
/*  H0: ratio in population 1 = ratio in population 2 */

proc ratio data = pooled_data design = brr;
        class indicator_first;
```

```
        weight wght_per;
        denom one;
        numer indicator_value_1;
        repwgt wtbs_001 - wtbs_200 / adjfay = 25;
        contrast indicator_first = (-1 1);

run;
```

Proc ratio donne de bons résultats lorsqu'il s'agit de variables catégoriques (binaires). Pour tester une variable numérique au moyen de SUDAAN, on peut utiliser la procédure descript (il faut veiller à éliminer toutes les valeurs « non déclaré » et « ne sais pas ».

Après chaque test, le chercheur doit décider quelle sera l'incidence sur le projet de groupement. Si les deux paramètres de population respectifs ne sont pas statistiquement différents d'après les estimations, on peut calculer une estimation groupée utile. Toutefois, si le test échoue, le chercheur a deux options : soit éliminer l'estimation du projet d'analyse des données groupées, soit utiliser la variable (et la valeur) mais avec prudence (en ajoutant une mise en garde dans la documentation finale, par exemple). Le choix de cette deuxième option dépend de la valeur p de la statistique de test et la signification d'un test échoué pour le secteur spécialisé.

Point à examiner 6 : Groupement, pondération bootstrap et estimation de la variance Comme nous l'avons signalé ci-dessus, il existe diverses façons de calculer les estimations à partir de données groupées. La méthode la plus répandue consiste simplement à ajuster les poids et à procéder au calcul comme s'il n'y avait qu'un seul échantillon[15]. La façon de procéder pour ajuster les poids dépend de la population finale souhaitée. Comme notation, soit w_{1i} = le poids pour le $i^{ième}$ enregistrement dans le premier échantillon et w_{2j} = le poids pour le $j^{ième}$ enregistrement dans le deuxième échantillon. Supposons en outre que n_1 et n_2 sont les tailles respectives d'échantillon et que $N_1 = \sum w_{1i}$ et $N_2 = \sum w_{2j}$ sont les tailles de population estimatives respectives.

Pour obtenir la moyenne de N_1 et de N_2, procéder à l'ajustement :

$$w'_{1i} = w_{1i} \times 1/2 \text{ et } w'_{2j} = w_{2j} \times 1/2.$$

Pour obtenir N_2 comme le total de la *population groupée* (ici, le paradigme est que le premier cycle était simplement une collecte antérieure auprès de la deuxième population), procéder à l'ajustement :

$$w'_{1i} = w_{1i} \times N_2/(N_1 + N_2) \text{ et } w'_{2j} = w_{2j} \times N_2/(N_1 + N_2).$$

[15] On se rappellera que nous supposons que les deux cycles différents sont indépendants.

En général, si on multiplie les premiers poids par α et les deuxièmes poids par β, on obtient une population estimative totale de $\alpha \times N_1 + \beta \times N_2$. En théorie, on pourrait apporter un ajustement au poids pour chaque[16] estimation groupée à calculer et pour chaque domaine ou sous-population d'intérêt. Par ailleurs, on peut traiter des valeurs différentes de non-réponse à une question de cette façon, en procédant à différents ajustements. Dans la pratique, toutefois, il est probablement préférable de choisir un ajustement et de l'utiliser pour toutes les analyses. Les deux ajustements exposés ci-dessus semblent convenir le mieux à une large gamme de types d'analyses.

Quel que soit l'ajustement choisi, il doit être apporté aux poids finaux *et* à chacun des ensembles de poids bootstrap.

Marche à suivre 4 : Comment calculer une estimation groupée et sa variance connexe? En appliquant le premier paradigme ci-dessus (calcul de la moyenne des deux populations), le chercheur, essentiellement, a besoin de calculer seulement les nouveaux poids. Nous supposons les mêmes variables que celles dans la marche à suivre 3 ci-dessus. Dans SUDAAN, nous pourrions utiliser :

```
/* create new weights that are half the old ones */

data pooled_data;
        set pooled_data;
        wght_per_new = wght_per / 2;

        array w_old wtbs_001 – wtbs_200;
        array w_new wtbs_new_001 – wtbs_new_200;

        do i = 1 to 200;
                w_new(i) = w_old(i) / 2;
        end;

run;

/* compute ratio value_1 / one and its associated variance for the */
/* whole data set */

proc ratio data = pooled_data design = brr;
        weight wght_per_new;
        denom one;
        numer indicator_value_1;
        repwgt wtbs__new_001 - wtbs__new_200 / adjfay = 25;
```

[16] D'ailleurs, si on fait simplement la moyenne des poids, on n'obtient pas nécessairement les estimations les plus efficaces en termes de variance. Étant donné que la plupart des études contiennent de nombreuses variables, des méthodes de groupement plus efficaces pour une variable peuvent ne pas donner de bons résultats pour une autre. Le compromis le plus répandu est celui proposé dans le texte.

run;

Point à examiner 7 : Énoncer les conclusions et ajouter des mises en garde Lorsque le travail statistique est achevé, le chercheur doit décider de la validité et de l'applicabilité des résultats. Dans la présente section, nous fournissons une liste de mises en garde possibles. Une ou plusieurs peuvent être ajoutées au texte de tout projet de recherche ou servir de « liste de contrôle » pour les conclusions finales.

Le premier commentaire est général : pour les tests statistiques portant sur les paramètres de population, lorsque les valeurs p sont proches de 0,05, la valeur critique type, il faut user de prudence. En effet, dans un projet de groupement de nombreux facteurs pourraient accroître ou réduire la variance réelle d'un estimateur, qui a été estimée, mettons, par la variance groupée. L'auteur est d'avis que la plupart des hypothèses faites ci-dessus auraient peu d'effet incrémentiel sur les variances, mais il est souvent difficile, voire impossible, de prévoir l'effet d'un groupe quelconque d'hypothèses sur la vapeur p d'un test. Bref, le chercheur doit se rappeler que des valeurs p proches de 0,05 ne constituent pas une « forte preuve » en faveur du rejet ou non-rejet.

La liste des mises en garde suit plus ou moins le texte du document :

Marche à suivre 5 : Liste de mises en garde possibles : Voici une liste de base :

- mises en garde statistiques habituelles et hypothèses concernant la loi de distribution pour les modèles, indépendamment de l'utilisation de la méthode de groupement;

- mises en garde habituelles concernant l'ESG : seuls sont interviewés les adultes de 15 ans et plus vivant dans les provinces et ne vivant pas en établissement, et l'enquête par composition aléatoire exclut les personnes sans téléphone, etc.;

- deux cycles réalisés à un an d'intervalle[17] : « même si les enquêtes ont été réalisées en deux années adjacentes, nous pouvons considérer les deux échantillons comme représentant la même population de Canadiens, puisque peu de changements auront été observés »;

- deux cycles réalisés à plus d'un an d'intervalle[18] : « même si les deux enquêtes ont été réalisées à x années d'intervalle et si certains

[17] Dans ce cas, il suffirait probablement de faire simplement la moyenne des poids.

[18] Dans ce cas, le choix de α et de β, comme dans le point à examiner 6 qui met davantage l'accent sur l'échantillon plus récent, pourrait être plus approprié. Si le chercheur dispose de suffisamment de temps, il pourrait examiner différents choix MAIS le choix devrait être fait *a priori* et non de manière à correspondre à un ensemble souhaité de conclusions. Autrement dit, un choix de α et de β constituerait une mise en garde en soi.

changements sont survenus dans la population canadienne entre les deux points dans le temps, les essais exhaustifs des estimations que nous avons utilisées dans notre travail ont révélé qu'en ce qui a trait à nos analyses, la population canadienne était stable »;

- cycles adjacents non chevauchants : « pour réduire le fardeau du répondant, si un répondant participe à un cycle, il est exclu du cycle suivant; ainsi, les échantillons sont non chevauchants, de sorte que nous avons groupé les données des deux cycles en concaténant simplement les deux ensembles de données »;

- cycles non adjacents non chevauchants : « la possibilité qu'un répondant participe à deux cycles différents de l'ESG est extrêmement rare, de sorte que nous avons groupé les données des deux cycles en concaténant simplement les deux ensembles de données et en les considérant comme un seul plus grand ensemble de données »;

- cycles adjacents légèrement dépendants (non nécessaire dans le cas de cycles non adjacents) : « pour réduire le fardeau du répondant, si un répondant participe à un cycle, il est exclu du cycle suivant; ainsi, les échantillons ne sont pas statistiquement indépendants; toutefois, la possibilité d'un chevauchement est extrêmement rare, de sorte que l'effet est minime et nous pouvons donc supposer que les deux ensembles de données ont été tirés indépendamment de la même population »;

- mêmes méthodologie, variables, estimations : il est utile de faire mention des recherches et mises à l'essai exhaustives qui ont été nécessaires pour s'assurer que les ensembles de données étaient comparables, étaient les mêmes par paire ou pouvaient être rendus comparables, et que les estimations ne présentaient pas de différences statistiquement significatives par paire; dans les cas où on a constaté des différences, on peut ajouter des mises en garde détaillées du genre : « l'effet d'une différence x sur y serait peu important »;

- petites tailles d'échantillon pour divers domaines : enfin, il convient de souligner que le groupement n'est pas nécessairement une panacée; la somme de deux petits domaines peut encore être petite; en pareil cas, il faut ajouter les mises en garde habituelles au sujet des petites tailles d'échantillon (p. ex., variances accrues, préoccupations concernant la confidentialité).

Conclusions : Nous avons donné un aperçu de certains éléments à prendre en considération au moment d'intégrer des données de deux ou de plusieurs cycles de l'Enquête sociale générale. En outre, nous avons présenté plusieurs options en ce qui concerne la marche à suivre pour réaliser un tel projet. Le groupement des données peut être très utile aux chercheurs qui souhaitent faire le suivi de tendances sociales au fil du

temps ou accroître la taille de petits ensembles de données, et il est tout à fait faisable dans le contexte de l'ESG, même si cette dernière est une enquête transversale annuelle.

Annexe : Certains renseignements sur les poids pour divers cycles

Le tableau ci-dessous donne un bref aperçu des divers cycles et des considérations en matière de pondération (ce tableau sera également inclus dans le document d'ensemble décrivant la façon d'appliquer les poids de l'ESG qui paraîtra dans le Bulletin technique des Centres de données de recherche de Statistique Canada).

Cycle	Fichier	Principale(s) variable(s) de poids	Noms des poids bootstrap actuels	Moyenne	Commentaires
1	principal	wght	S.O.	S.O.	wght = 10000 x poids; projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
2	principal	fwght_os	S.O.	S.O.	Projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
2	sommaire	fwght_ms		S.O.	il s'agit du fichier sommaire d'emploi du temps, qui contient un enregistrement par répondant
2	épisode	fwght_ms	S.O.	S.O.	Contient un enregistrement par épisode d'emploi du temps; pour la façon d'utiliser les poids, voir le Guide de l'utilisateur du FMGD
3	principal	weight32, weight33, weight34	S.O.	S.O.	Weight32 = 10000 x poids, etc.; flag32 est un indicateur d'information au niveau de la personne; les épisodes d'accident peuvent être analysés séparément au moyen de l'indicateur flag33, les incidents criminels peuvent être analysés séparément au moyen de l'indicateur flag34; les poids correspondent à ces trois types d'enregistrements; projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
4	principal	pweight	S.O.	S.O.	pweight = 10000 x poids; projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
5	principal	pweight	S.O.	S.O.	pweight = 10000 x poids; projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
6	principal	finalwt	S.O.	S.O.	projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
7	principal	fwght	S.O.	S.O.	projet de post-pondération bootstrap en cours; 200 poids attendus bientôt

Cycle	Fichier	Principale(s) variable(s) de poids	Noms des poids bootstrap actuels	Moyenne	Commentaires
7	sommaire	timewgt	S.O.	S.O.	il s'agit du fichier sommaire d'emploi du temps, qui contient un enregistrement par répondant
7	épisode	timewgt	S.O.	S.O.	contient un enregistrement par épisode d'emploi du temps; pour la façon d'utiliser les poids, voir le Guide de l'utilisateur du FMGD
8	principal	wght_per	wpebs_001 - wpebs_200	25	
9	principal	perwght	S.O.	S.O.	projet de post-pondération bootstrap en cours; 200 poids attendus bientôt
10	principal	wghtfnl	bsw1 - bsw200	25	
10	enfant			S.O.	pas de poids au niveau de l'enfant
10	syndicat			S.O.	pas de poids au niveau du syndicat
11	principal	wght_fnl	wfin_001 - wfin_200	25	
12	principal	wghtfin	wfin_001 - wfin_200	25	
12	épisode			S.O.	
13	principal	wght_per	wpebs001 - wpebs200	25	
13	incident		wvcbs001 - wvcbs200	25	
14	principal	wght_per	wfin_001 - wfin_200	25	
15	principal	wght_per	wtbs_001 - wtbs_200	25	
15	enfant			S.O.	pas de poids au niveau de l'enfant
15	syndicat			S.O.	pas de poids au niveau du syndicat
16	principal	wght_per	wtbs_001 - wtbs_200	25	
16	soins reçus			S.O.	contient seulement les personnes de 65 ans et plus qui ont reçu des soins; pas de poids indiqué pour l'épisode
16	soins fournis 45-64			S.O.	contient seulement les personnes de 45 à 64 ans qui ont fourni des soins; pas de poids indiqué pour l'épisode
16	soins fournis 65 ans+			S.O.	contient seulement les personnes de 65 ans et plus qui ont fourni des soins; pas de poids indiqué pour l'épisode
17	principal	wght_per	wtbs_001 - wtbs_200	25	
18	principal	wght_per	wtbs_001 - wtbs_200	25	
18	incident	adjwtvic, wght_vic	wvcbs001 - wvcbs200	25	pour la différence entre les deux poids finaux, voir le Guide de l'utilisateur; les poids bootstrap correspondent à wght_vic
19	principal	wght_per	wtbs_001 - wtbs_500	25	

Cycle	Fichier	Principale(s) variable(s) de poids	Noms des poids bootstrap actuels	Moyenne	Commentaires
19	csp	wght_csp	wtcbs_001 - wtcbs_500	25	applicable aux questions à la section 10a
19	snt	wght_snt	wtsbs_001 - wtsbs_500	25	applicable aux questions à la section 10b, 11
19	épisode	wght_epi	wtbs_epi_001 - wtbs_epi_500	25	
20	principal	wght_per	wtbs_001 - wtbs_500	25	

Bibliographie :

Binder, D., Roberts, G. (2007) *Approaches for Analyzing Survey Data: a Discussion*, preprint, Statistics Canada.

Korn, E. L., Graubard, B.I. (1998), *Analysis of Health Surveys*, Wiley.

Marchand, I. (2007) *Combinaison des cycle 13 (1999) et cycle 18 (2004) de l'Enquête sociale générale pour dériver un profil de victimisation*, internal document, Statistics Canada.

Schenker, N., Raghunathan, T. (2007) *Combining information from multiple surveys to enhance estimation of measures of health*, Statistics in Medicine 2007; 26:1802-1811

Thomas, S. (2006) *Combining Cycles of the Canadian Community Health Survey*, Proceedings of the Statistics Canada Symposium, 2006.