N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

DÉPISTAGE ET AJUSTEMENT POUR LA NON-RÉPONSE DANS L'ENQUÊTE LONGITUDINALE AUPRÈS DES IMMIGRANTS DU CANADA

M. Simard, T. Leesti et J. Denis[1]

RÉSUMÉ

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) est conçue pour étudier le processus d'intégration des nouveaux immigrants à la société canadienne. L'enquête s'appuie sur un plan de sondage longitudinal, conformément auquel les immigrants sélectionnés sont interviewés à trois points dans le temps, c'est-à-dire environ six mois (cycle 1), deux ans (cycle 2) et quatre ans (cycle 3) après l'obtention du droit d'établissement au Canada. Comme la population de nouveaux immigrants est caractérisée par une très grande mobilité, le dépistage pour l'exécution du premier cycle a été particulièrement difficile. Comme la population cible venait d'arriver très récemment au Canada, les sources administratives conventionnelles n'ont été que d'une utilité limitée pour le dépistage. Par conséquent, la proportion de cas non résolus était généralement plus élevée pour le premier cycle de l'ELIC que pour d'autres enquêtes de Statistique Canada (environ 28 %). Le présent article décrit les travaux effectués dans le cadre de l'ELIC en ce qui concerne le dépistage des répondants, ainsi que l'étude de la non-réponse et la correction de cette dernière. Il met en relief les défis que pose le dépistage des nouveaux immigrants, ainsi que les stratégies mises en œuvre pour augmenter les taux de réponse. En outre, il décrit la méthode choisie pour corriger pour la non-réponse, c'est-à-dire une technique assistée par modèle fondée sur l'approche proposée par Eltinge et Yanseneh pour définir les classes d'ajustement de la pondération. Cette méthode intègre deux ajustements distincts, l'un pour les cas non dépistables et l'autre pour les non-répondants.

MOTS CLÉS : Assisté par modèle, dépistage, non-réponse, régression logistique.

1. INTRODUCTION

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) est conçue pour étudier la façon dont les nouveaux immigrants s'adaptent à la vie au Canada. Les nouveaux arrivants seront interviewés trois fois au cours des quatre premières années qu'ils passeront au Canada, environ six mois, deux ans et quatre ans après leur établissement, afin de brosser un tableau dynamique de leurs expériences. L'ELIC est la première enquête nationale réalisée auprès des nouveaux immigrants depuis 1970.

Parce qu'ils forment une population très mobile, les nouveaux immigrants sont difficiles à dépister et, subséquemment, à interviewer. Par conséquent, découvrir la méthode appropriée de pondération est un exercice délicat. Le présent article décrit la conception de l'enquête, les difficultés posées par le dépistage des nouveaux immigrants et les résultats de la collecte des données du premier cycle. Il décrit aussi dans les grandes lignes la stratégie utilisée pour faire la correction pour la non-réponse, c'est-à-dire une méthode de repondération en trois étapes fondée sur un modèle.

[1] M. Simard, T. Leesti et J. Denis, Statistique Canada, Pré Tunney, 120, avenue Parkdale, Ottawa, Ontario, Canada, K1A 0T6, johanne.denis@statcan.ca; michelle.simard@statcan.ca; tracey.leesti@statcan.ca.

2. CONTEXTE DE L'ENQUÊTE

2.1 Objectifs et contenu de l'enquête

Le besoin d'information sur les nouveaux immigrants du Canada, surtout leur processus d'adaptation, les facteurs qui influent sur l'adaptation et les services qu'ils utilisent pour faciliter le processus, devient de plus en plus pressant. Bien que l'adaptation complète puisse prendre plusieurs générations, l'ELIC est conçue pour examiner le processus durant la période critique des quatre premières années de l'établissement, durant laquelle les nouveaux venus nouent des liens économiques, sociaux et culturels avec la société canadienne.

On pose aux participants à l'enquête des questions sur divers aspects de leur vie, allant des raisons qui les ont poussés à venir s'installer au Canada aux problèmes qu'ils ont dû surmonter pour trouver un logement et un emploi et pour avoir accès aux études. L'enquête recueille aussi des renseignements sur le niveau de scolarité et la santé de leurs enfants. Des questions sur leur capacité d'accès aux services sont intégrées tout au long du questionnaire.

2.2 Conception de l'enquête

L'enquête est réalisée selon un plan de sondage longitudinal, conformément auquel les nouveaux immigrants sont interviewés trois fois, environ six mois, deux ans et quatre ans après l'obtention du droit d'établissement au Canada[2]. Le plan d'échantillonnage a été élaboré selon une approche monotone ou « en forme d'entonnoir ». Seuls les immigrants qui ont participé à l'interview du cycle précédent sont contactés lors du cycle suivant. Cette approche a été adoptée à cause de la nature de l'enquête et des objectifs analytiques.

L'enquête est conçue pour recueillir des renseignements sur les perceptions, les valeurs et les attitudes à des points particuliers dans le temps, afin d'évaluer l'intégration des immigrants durant les premières années de leur vie au Canada. Si les données n'étaient recueillies qu'une seule fois (c.-à-d. durant la quatrième année passée au Canada), des erreurs importantes de remémoration et de réponse pourraient se produire. En outre, pour faciliter la réalisation d'une étude complète de l'adaptation des immigrants, la gamme complète de données longitudinales doit être recueillie auprès de chaque répondant longitudinal.

2.3 Population cible et échantillon

La population cible de l'enquête comprend les immigrants qui satisfont aux critères suivants :

- immigrants reçus arrivant de l'extérieur du Canada;
- arrivés au Canada entre le 1er octobre 2000 et le 30 septembre 2001;
- âgés de 15 ans ou plus au moment de l'établissement au Canada.

Les immigrants qui ont fait une demande de droit d'établissement et l'ont obtenu alors qu'ils étaient au Canada sont exclus de l'enquête. Ces personnes pourraient avoir séjourné au Canada depuis un temps assez long avant d'obtenir officiellement « le droit d'établissement » et, par conséquent, manifesteraient probablement des caractéristiques d'adaptation différentes de celles des nouveaux immigrants au Canada. Environ 21 000 des 165 000 immigrants de 15 ans et plus qui se sont établis au Canada entre octobre 2000 et septembre 2001 ont été sélectionnés pour participer à l'enquête.

[2] Le premier cycle d'interviews a été réalisé d'avril 2001 à mars 2002. Le deuxième cycle a débuté en décembre 2002 et se poursuivra pendant un an environ. Le troisième cycle débutera en principe en octobre 2004.

3. PLAN D'ÉCHANTILLONNAGE

3.1 Base de sondage

La base de sondage de l'ELIC est le Système de soutien des opérations des bureaux locaux (SSOBL) de Citoyenneté et Immigration Canada, c'est-à-dire une base de données administratives contenant l'information sur toutes les personnes qui sont arrivées au Canada, y compris celles venues en tant qu'immigrant. Le SSOBL est une riche base de données contenant des renseignements sur diverses caractéristiques des immigrants qui peuvent être utilisées pour établir le plan de sondage et qui constituent aussi une excellente source d'information auxiliaire. Les variables d'intérêt importantes figurant dans le SSOBL incluent l'âge, le sexe, la langue maternelle, le pays d'origine, la connaissance de l'anglais et (ou) du français, la composante d'immigration, la date de l'établissement et la province prévue de destination au Canada.

Statistique Canada a reçu des renseignements détaillés tirés du SSOBL pour chaque immigrant qui s'est établi au Canada durant la période de référence de l'enquête (d'octobre 2000 à septembre 2001), deux mois après le mois de référence. Cela a permis de créer la base de sondage mois après mois en ajoutant simplement les nouveaux cas d'établissement survenus durant le mois.

3.2 Exigences d'échantillonnage

L'échantillon est subdivisé en deux composantes, à savoir l'échantillon de base et les échantillons supplémentaires. L'échantillon de base représente la population cible, tandis que les échantillons supplémentaires visent des sous-populations particulières. En s'appuyant sur les exigences de divers ministères des administrations fédérale et provinciales, ces sous-populations particulières ont été déterminées par analyse de la répartition prévue de l'échantillon au moment du troisième cycle. Les sous-groupes suivants ont été suréchantillonnés :

- réfugiés parrainés par le gouvernement;
- réfugiés autres que ceux parrainés par le gouvernement;
- entrepreneurs et investisseurs immigrants (« composante économique-gens d'affaires »);
- immigrants de la composante du regroupement familial en Colombie-Britannique;
- ensemble des immigrants en Alberta;
- immigrants de la composante économique au Québec (« économique-travailleurs qualifiés » et « économique-gens d'affaires ».

L'échantillon est créé selon une méthode d'échantillonnage stratifié à deux degrés. La première étape comporte la sélection des unités immigrantes (UI)[3] avec probabilité proportionnelle à la taille (PPT) de l'UI. La deuxième étape comporte la sélection d'un membre de l'UI dans chaque UI sélectionnée. Le membre sélectionné de l'UI est le répondant longitudinal (RL) qui participera à l'enquête. Seul le répondant longitudinal est suivi durant l'enquête et aucune interview n'est réalisée auprès des autres membres de l'unité immigrante.

Il a été établi que pour l'échantillon de base, la réalisation de 5 000 interviews complètes au moment du troisième cycle produirait des estimations fiables[4] au niveau national, pour les provinces où l'influx d'immigrants est le plus important (Québec, Ontario et Colombie-Britannique) et pour certaines composantes d'immigration (regroupement familial et économique). Il serait également possible d'obtenir des estimations fiables pour d'autres combinaisons de provinces et composantes d'immigration. En tenant compte des exigences relatives aux échantillons supplémentaires mentionnées antérieurement, le nombre global prévu d'interviews achevées lors du troisième cycle est de 5 755. Pour calculer la taille requise d'échantillon lors du premier cycle, on a appliqué une méthode à rebours, fondée sur des hypothèses d'érosion de l'échantillon établies d'après diverses sources. Selon cette méthode, la taille requise de l'échantillon au premier cycle était de 20 322 unités.

[3] Une unité immigrante comprend tous les individus qui font une demande d'immigration au Canada sur le même formulaire de demande de visa.

[4] Par estimation fiable, nous entendons être capable d'estimer une proportion minimale de 10 % avec un coefficient de variation de 16,5 %. Un nombre minimal de 450 immigrants répondants est nécessaire pour satisfaire à cette exigence.

4. MÉTHODES DE COLLECTE DES DONNÉES

Les interviews du premier cycle ont été menées d'avril 2001 à mars 2002. Étant donné les diverses contraintes opérationnelles (capacité sur le terrain, difficulté à dépister les immigrants), pour chaque mois d'échantillonnage, les opérations sur le terrain ont duré trois mois. Les interviews, dont la durée moyenne était d'environ 90 minutes, ont été réalisées dans l'une des 15 langues différentes proposées, à savoir l'anglais, le français, le mandarin, le cantonais, le punjabi, le farsi/dari, l'arabe, l'espagnol, le russe, le serbo-croate, l'urdu, le coréen, le tamil, le tagalog et le gujarati. Environ 70 % d'interviews ont été réalisées sur place et les 30 % restants, par téléphone. Toutes les interviews ont été réalisées par la technique d'interview assistée par ordinateur. Aucune interview n'a été réalisée au Yukon, dans les Territoires du Nord-Ouest ou au Nunavut étant donné les coûts élevés de collecte et le fait qu'un fort petit nombre d'immigrants s'établissent dans ces régions. Les interviews ont été réalisées dans toutes les régions métropolitaines de recensement et toutes les régions non éloignées du Canada.

5. DÉPISTAGE

Le dépistage des nouveaux immigrants et la prise de contact avec ces derniers sont des activités particulièrement compliquées, car les nouveaux immigrants ont tendance à former une population très mobile, point qui a été illustré durant l'essai pilote de l'ELIC réalisé au printemps 1997. Durant cet essai, on a constaté que presque la moitié de la population cible avait déménagé au moins une fois au cours des six premiers mois passés au Canada.

5.1 Défis du dépistage

L'une des plus grandes difficultés qu'il a fallu contourner durant les opérations en vue de repérer les répondants potentiels pour le premier cycle de l'enquête était le manque initial de renseignements sur les coordonnées des nouveaux immigrants dans la base de sondage. Au moment de l'arrivée au Canada, les immigrants sont tenus uniquement d'indiquer la province dans laquelle ils ont l'intention de s'installer. Pour certains, il s'agissait de la seule information disponible et, parfois, il s'est avéré que la province prévue de destination n'était pas celle où l'immigrant s'était finalement installé.

En outre, au moment de l'interview du premier cycle, la population cible de l'ELIC n'avait vécu au Canada que six mois. Par conséquent, les sources administratives habituelles qu'on peut demander de consulter pour le dépistage dans le cas d'autres enquêtes, comme les fichiers de formulaires T1 et T4, les fichiers des prestations fiscales pour enfants, les fichiers d'assurance-chômage ou le registre des adresses de Revenu Canada, n'étaient pas disponibles pour le premier cycle de dépistage, puisque les nouveaux immigrants n'étaient pas encore inclus dans ces fichiers.

5.2 Stratégies de dépistage

Plusieurs stratégies ont été mises en œuvre afin de surmonter certaines difficultés posées par le dépistage des répondants. Ainsi, dès le début de l'enquête, les activités mensuelles de dépistage dans chaque région ont été surveillées de près. Tout problème éventuel a été cerné immédiatement et des solutions appropriées ont été proposées.

En outre, des équipes spécialisées de dépistage s'occupant uniquement de l'ELIC ont été mises en place dans certains bureaux régionaux. Cette coordination des activités de dépistage a donné plus de temps aux intervieweurs pour mener les interviews, et a permis à l'unité spécialisée d'acquérir de l'expérience dans le dépistage de cette population particulière. Et, si la plupart du dépistage a été effectuée par l'unité spécialisée, il a été reconnu qu'un certain degré de dépistage devait encore avoir lieu sur le terrain et être traité par des intervieweurs particuliers. Afin de compenser pour cette situation, la période de collecte mensuelle a été étendue pour donner du temps supplémentaire pour le dépistage de répondants éventuels et la prise de contact avec ces derniers.

Enfin, comme nous l'avons mentionné plus haut, les renseignements sur l'adresse de contact enregistrée dans la base de sondage n'étaient pas toujours complets. Au Bureau central, une activité importante a consisté à mettre à jour le fichier de l'échantillon en recueillant autant de renseignements que possible sur les adresses au moyen de sources de dépistage à jour pertinentes pour la population cible et susceptibles de fournir certaines coordonnées pertinentes; toutes ces activités ont eu lieu avant que le fichier soit transmis aux intervieweurs sur le terrain aux fins de la collecte des données. Ces sources incluent le questionnaire de prise de contact de l'ELIC, les fichiers d'adresses des ministères provinciaux de la Santé (pour toutes les provinces sauf une), les fichiers administratifs de Citoyenneté et Immigration Canada et les fichiers téléphoniques.

Le questionnaire de prise de contact de l'ELIC a été conçu pour faciliter le repérage des répondants potentiels après leur arrivée au Canada. Ce questionnaire a été remis à chaque immigrant au moment de l'émission outre-mer du visa leur donnant droit à l'établissement et a été recueilli par les agents de l'immigration au port d'entrée au Canada. Ce questionnaire demandait aux immigrants de donner une adresse probable au Canada (si elle était connue), ainsi que l'adresse d'une personne-ressource au Canada.

Durant le développement de l'ELIC, il a été confirmé que les fichiers d'adresses des ministères provinciaux de la Santé seraient la meilleure source de dépistage, car les nouveaux immigrants peuvent faire la demande d'une carte d'assurance-maladie dans les trois mois de leur arrivée. L'accès à ce fichier n'est, cependant, accordé qu'avec le consentement de l'immigrant. Par conséquent, une question concernant le consentement a été ajoutée au questionnaire de prise de contact de l'ELIC demandant aux immigrants d'autoriser Statistique Canada à consulter l'information détenue par les organismes provinciaux de Santé aux fins du dépistage uniquement.

Alors que le taux de consentement des immigrants a été élevé (79 %), le taux de retour des questionnaires a été assez faible et, par conséquent, les adresses correspondant à la carte d'assurance-maladie n'ont été obtenues que pour 35 % environ de l'échantillon. Cependant, lors de l'analyse subséquente de la qualité des sources de dépistage, les fichiers d'assurance-santé se sont avérés être l'une des sources les plus fiables, 77 % des cas donnant des informations pertinentes concernant l'adresse.

Après avoir compilé toutes les informations de dépistage éventuel, on a attribué un ordre de priorité à chaque source de dépistage obtenue pour l'immigrant sélectionné et on a envoyé le fichier au bureau régional aux fins de la collecte. Des renseignements permettant une prise de contact éventuelle ont été fournis pour environ 75 % de l'échantillon, avec, en moyenne, cinq points de contact pour chaque immigrant.

6. RÉSULTAT DE LA COLLECTE

6.1 Résultats et observations

En bout de ligne, 59,2 % de l'échantillon original ont reçu un code d'unité répondante, résultat un peu meilleur que les 50 % estimés à l'étape de la planification. Le tableau 1 donne les résultats possibles de la collecte des données tels qu'ils étaient prévus et tels qu'ils ont été observés lors du premier cycle. Durant la collecte, tout immigrant sélectionné pouvait être classé dans l'une de quatre catégories, à savoir *répondant*, *non-répondant*, *hors du champ de l'enquête* et *cas non résolu*. Les trois premières catégories ont été définies comme étant des *cas résolus* car la situation de l'immigrant était connue. Elles ont abouti à une prise de contact avec l'immigrant ou avec une personne capable de confirmer sa situation, c'est-à-dire dans ou hors du champ de l'enquête. Certains exemples de cas hors du champ de l'enquête sont le décès, la date de naissance incorrecte dans le SSOBL ou l'unité ne résidant plus au Canada. Le quatrième résultat possible de la collecte est celui des *cas non résolus* ou non dépistables. Ces cas sont ceux pour lesquels aucun contact n'a plus être établi et qui sont demeurés non résolus. Aucun renseignement sur l'endroit où ces personnes se trouvaient au Canada n'était disponible. Le tableau montre que le taux observé de *cas non résolus* (environ 28 %) est plus élevé que le taux prévu (23 %). Inversement, le taux observé de *cas résolus hors du champ de l'enquête* est plus faible que prévu (2,8 % contre 7,8 %).

Tableau 1 : Résultats de la collecte du premier cycle

Résultat	Prévu	Observé
Cas résolus	**77,0**	**71,7**
Répondants dans le champ de l'enquête	65,3	82,7
Non-répondants dans le champ de l'enquête	26,9	14,5
Hors du champ de l'enquête	7,8	2,8
Cas non résolus	**23**	**28,3**

Ces résultats combinés donnent à penser que les cas non résolus pourraient compter une proportion significativement plus élevée d'unités hors du champ de l'enquête. Par conséquent, on a entrepris un examen complet des résultats de la collecte et une étude exhaustive des profils de réponse. En plus de surveiller de près les codes de résultat de la collecte et d'examiner les notes des intervieweurs, on a consulté des experts de l'immigration pour mieux comprendre les caractéristiques de la population cible. Bien qu'une partie des cas non résolus puissent être attribués au manque d'information de dépistage, certains sont demeurés non résolus malgré l'existence de très bons renseignements de dépistage. Il était raisonnable de penser qu'une bonne partie de ces cas non résolus correspondaient à des immigrants ayant quitté le pays. Durant le processus d'élaboration de l'enquête, il est devenu évident que ces résultats devaient être intégrés dans la stratégie de repondération pour corriger comme il convient les profils de non-réponse.

6.2 Mécanismes de réponse

Les profils et les mécanismes de réponse ont dû être étudiés avec soin pour apporter la correction appropriée pour la non-réponse. Les mécanismes de réponse classiques sont les suivants :

i) uniforme [ou réponse manquant entièrement au hasard (MCAR, missing completely at random)*]* : la probabilité de réponse est complètement indépendante du processus de mesure et est constante sur l'ensemble de la population;
ii) dépendant ignorable [ou réponse manquant au hasard (MAR, missing at random)*]* : la probabilité de réponse est conditionnellement indépendante des mesures non observées, connaissant les mesures observées, autrement dit, le mécanisme de réponse dépend de certaines données auxiliaires ou des variables disponibles pour toutes les unités mesurées;
iii) dépendant non ignorable [ou réponse ne manquant pas au hasard (NMAR, not missing at random)*]* : la probabilité de réponse dépend de la variable d'intérêt, donc, n'est pas disponible pour toutes les unités mesurées.

Souvent, on corrige pour la non-réponse par repondération à l'intérieur de certains groupes ou cellules qui sont définis d'après i) le mécanisme de non-réponse sous-jacent et ii) le repérage d'unités d'échantillonnage ayant des probabilités de réponse comparables, c'est-à-dire l'application de la théorie des groupes à réponse homogène (GRH). Puis, à l'intérieur de ces groupes, les poids appliqués aux unités répondantes sont rajustés pour tenir compte des non-répondants. La méthode utilisée pour définir les groupes est le sujet de la section 8.2. Pour l'ELIC, non seulement le mécanisme de réponse ne paraît pas uniforme, mais les résultats de la collecte semblent indiquer que plus d'un mécanisme intervient dans la production de l'échantillon effectif.

Les tableaux 2 et 3 fournissent certaines précisions au sujet de la collecte des données. Le tableau 2 présente la répartition de l'échantillon par catégorie de réponse selon la composante d'immigration et le tableau 3, selon le groupe d'âge. Ces deux tableaux fournissent des preuves préliminaires que des profils différents se dégagent de l'échantillon. Ainsi, au tableau 2, la composante économique affiche un taux de réponse plus élevé que la composante de la famille, mais un taux de cas résolus nettement plus faible, ce qui indique que les immigrants de la composante économique ont été plus difficiles à dépister, mais qu'une fois rejoints, ils ont répondu à l'enquête. Bien qu'elle affiche le taux de cas résolus le plus élevé, la composante de la famille a le taux de non-réponse le plus élevé. Le résultat est de nouveau que certains répondants potentiels n'ont pas répondu au questionnaire, mais le processus de non-réponse diffère de celui observé pour la composante économique. Les réfugiés sont les immigrants pour

lesquels les taux de réponse et de cas résolus sont les plus élevés. Enfin, les taux de cas hors du champ de l'enquête varient selon la composante d'immigration. Des conclusions similaires peuvent être tirées si l'on examine les résultats selon le groupe d'âge, tels que présentés au tableau 3.

Tableau 2 : Résultats de la collecte du premier cycle selon la composante d'immigration

Résultat	Économique	Famille	Réfugié	Total
Cas résolus	**66,5**	**79,9**	**80,0**	**71,7**
Répondants dans le champ de l'enquête	83,2	78,9	88,4	82,7
Non-répondants dans le champ de l'enquête	13,3	18,7	10,6	14,5
Hors du champ de l'enquête	3,5	2,3	1,0	2,8
Cas non résolus	**33,5**	**20,1**	**20,0**	**28,3**

Les cas appartenant aux groupes d'âge moyen, à savoir les 25 à 34 ans et les 35 à 44 ans, semblent plus difficiles à résoudre, mais une fois dépistés, ces immigrants ont tendance à répondre plus fréquemment (environ 85 % pour chaque groupe d'âge). Le pourcentage de non-répondants dans le champ de l'enquête est aussi particulièrement élevé pour le groupe des 65 ans et plus.

Tableau 3 : Résultats de la collecte du premier cycle selon le groupe d'âge

Résultat	15 à 24 ans	25 à 34 ans	35 à 44 ans	45 à 64 ans	65 ans et plus	Total
Cas résolus	**74,7**	**67,9**	**70,6**	**77,5**	**85,7**	**71,7**
Répondants dans le champ de l'enquête	82,8	84,9	84,0	77,9	69,2	82,6
Non-répondants dans le champ de l'enquête	14,6	12,5	13,1	19,0	27,2	14,5
Hors du champ de l'enquête	2,6	2,6	3,0	3,1	3,6	2,8
Cas non résolus	**25,3**	**32,1**	**29,4**	**22,5**	**14,3**	**28,3**

La foule de renseignements disponibles dans la base de sondage a permis de faire des comparaisons avec d'autres variables qui, toutes, donnent des résultats similaires. L'étude de la répartition de l'échantillon est concluante : les cas répondants et non répondants diffèrent, les cas non résolus et non répondants diffèrent et les cas non résolus semblent afficher le même profil que les cas résolus hors du champ de l'enquête. Certaines hypothèses ont été émises en se fondant sur ces observations. Premièrement, plusieurs mécanismes non uniformes donnent naissance à l'échantillon résultant. Deuxièmement, la non-réponse comprend deux composantes, à savoir les unités non répondantes dans le champ de l'enquête provenant des cas résolus, d'une part et les cas non résolus, d'autre part. Comme il est raisonnable de penser qu'une bonne part des cas non résolus sont imputables à des immigrants qui ont quitté le Canada, la question qui se pose est celle de savoir comment il faudra en tenir compte dans la stratégie de pondération. Cette question renvoie directement à celle de savoir quelles personnes sont représentées par les poids finaux, ce qui est le sujet de la section suivante.

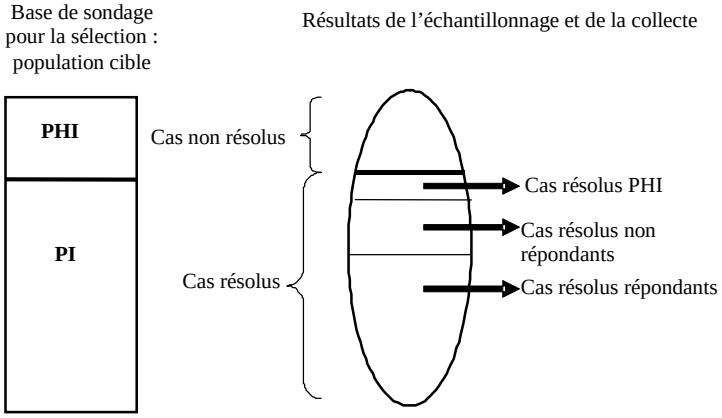
7. PROBLÈME DE PONDÉRATION – LE POIDS FINAL

Comme pour toute enquête probabiliste, l'échantillon de l'ELIC est sélectionné de façon à représenter une population de référence, à savoir la population d'immigrants, à un point particulier dans le temps. Chaque unité d'échantillonnage doit, par conséquent, représenter un certain nombre d'unités de la population. Si l'on arrivait à dépister, prendre contact et interviewer toutes les unités sélectionnées dans l'échantillon et si la base de sondage utilisée était parfaite (c.-à-d. couvrait exactement la population d'intérêt), alors le poids de sondage appliqué à

chaque unité représentait exactement et précisément le nombre d'immigrants dans la population cible. Dans cette situation, la pondération produirait des estimations sans biais. Cependant, il n'en est pas ainsi lorsque l'enquête comporte des cas de non-réponse, des cas non résolus ou non dépistables et une base de sondage imparfaite.

Pour la plupart des enquêtes, la somme des poids finaux représente les dénombrements estimés de la population cible, laquelle coïncide habituellement avec la population d'intérêt. Rappelons que la base de sondage couvre la population cible, c'est-à-dire les immigrants qui satisfont à chacun des trois critères décrits à la section 2.3. Toutefois, il a été confirmé que certains de ces immigrants se sont établis au Canada, mais n'y ont résidé que pour une courte période avant de retourner dans leur pays d'origine ou d'immigrer dans un autre pays. Comme l'objectif principal de l'enquête est de comprendre le processus d'intégration des immigrants qui sont arrivés récemment au Canada et que ceux qui sont partis n'ont pas les mêmes caractéristiques d'adaptation que ceux qui résident de façon permanente au Canada, regrouper les immigrants qui ont quitté le Canada et ceux qui y résident encore pourrait introduire un biais dans la correction de la pondération. Par conséquent, on a établi les concepts de population d'intérêt (PI) et de population hors du champ d'intérêt (PHI). La **population d'intérêt** comprend les immigrants qui répondent à chacun de trois critères décrits dans la définition de la population cible (section 2.3) ET qui ont vécu au Canada depuis plus de six mois durant une année de collecte des données particulière. La **population hors du champ d'intérêt** comprend les immigrants qui ne vivent plus au Canada, c'est-à-dire ceux qui en sont partis après avoir été admis comme immigrants. Le diagramme 1 illustre les concepts associés à la population, à l'échantillon et aux résultats de la collecte.

Diagramme 1 : Vue d'ensemble des concepts de l'enquête en rapport avec la pondération



8. AJUSTEMENT EN TROIS ÉTAPES ASSISTÉ PAR MODÈLE

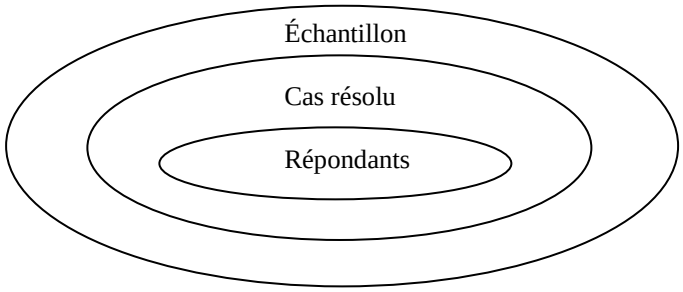
8.1 Principe général

La stratégie de pondération de l'ELIC est fondée sur une série d'ajustements en cascade intégrant les caractéristiques de toutes les unités considérées comme faisant partie de la *population d'intérêt* uniquement[5]. D'après les deux mécanismes distincts de non-réponse (c.-à-d. la non-réponse dans les parties résolue et non résolue de l'échantillon), deux ajustements sont appliqués consécutivement au poids de sondage des unités répondantes de la population d'intérêt. Le premier ajustement a pour objet de redistribuer les poids des unités non répondantes de la partie résolue de l'échantillon entre les poids des unités répondantes comprises dans la partie résolue. Le résultat de cette première étape est une estimation sans biais du nombre total de cas résolus dans la population d'intérêt. La deuxième ajustement vise à incorporer les poids des cas non résolus aux poids des cas résolus faisant partie de la population

[5] Le présent article couvre uniquement la pondération finale des unités faisant partie de la population d'intérêt (PI). Il existe aussi un poids final appliqué aux unités de la population hors du champ d'intérêt (PHI), mais il n'est pas décrit ici. Pour des détails sur les poids PHI, consultez le guide de l'utilisateur du fichier de microdonnées de l'ELIC (2003).

d'intérêt. Comme le montre le diagramme 2, ces rajustements correspondent à une approche en deux étapes. Rappelons toutefois que, comme on ne dispose d'aucune information sur la situation de ces cas non résolus, c'est-à-dire compris dans la PI ou dans la PHI, celle-ci sera prédite. Le deuxième ajustement consiste, en fait, à intégrer les unités classées par prédiction dans la PI dans la partie non résolue de l'échantillon.

Diagramme 2 : Ajustement en deux phases pour les unités faisant partie de la population d'intérêt de l'ELIC



8.2 Prédiction de la situation des cas non résolus

Commençons par présenter la notation. Après la collecte des données, chaque unité peut être classée dans la partie résolue ou non résolue de l'échantillon.

Échantillon : $S = U + R$ où U est l'ensemble de cas non résolus et R, l'ensemble de cas résolus. La partie résolue de l'échantillon peut être décomposée comme suit :

$$\mathrm{R} = \mathrm{RR} + \mathrm{RN} + \mathrm{RH}$$

où RR représente les unités répondantes, RN représente les unités non répondantes, RH représente les unités non comprises dans la population d'intérêt ou PI, formant la population hors du champ d'intérêt ou PHI[6] et $RPI = RR + RN$ représente les cas résolus dans la PI.

Conceptuellement, l'ensemble de cas non résolus (U) est composé d'unités de la PI et de la PHI. Le défi consiste à prédire la répartition entre ces deux sous-populations. Une option consiste à supposer que la répartition est la même que dans la partie résolue de l'échantillon. Cependant, il existe des preuves convaincantes que la proportion d'unités PHI est plus forte dans la partie non résolue que dans la partie résolue de l'échantillon. De surcroît, les taux de cas non résolus diffèrent selon certaines caractéristiques, ce qui donne à penser que le mécanisme de réponse n'est pas uniforme[7]. La solution consistait, par conséquent, à prédire la probabilité qu'une unité soit dans la PI d'après diverses caractéristiques sociodémographiques.

Soit X'_k l'ensemble de variables auxiliaires disponibles pour tous les *cas résolus*. X'_k est utilisé pour ajuster le modèle de régression logistique qui suit pour obtenir le paramètre de régression estimé $\hat{\alpha}$:

$$P\left[RPI_k = 1 \middle| X'_k\right] = \left(1 + \exp\left[-\left(\alpha_0 + \alpha_k X'_k\right)\right]\right)^{-1}$$

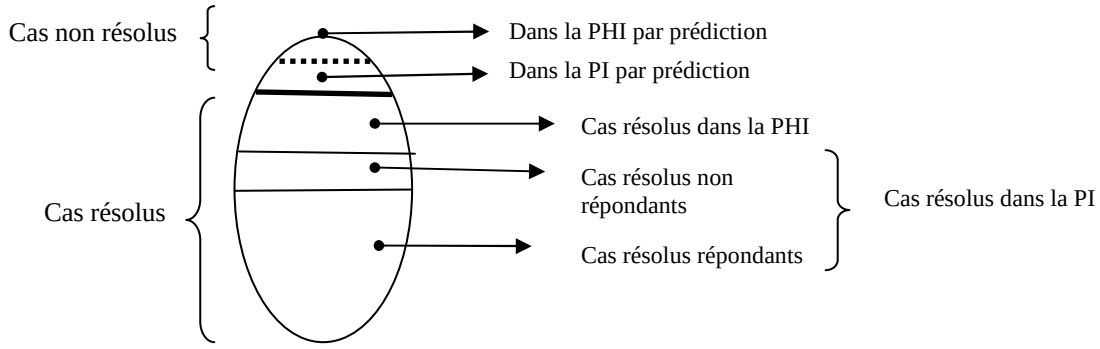
$$où\ RPI_k = \begin{cases} 1\ si\ k \in RPI \\ 0\ si\ k \in RH. \end{cases}$$

[6] Par souci de simplicité, nous définissons ces unités comme étant hors du champ d'intérêt par opposition à la population d'intérêt. La catégorie hors du champ de l'enquête habituelle est l'une des composantes de la catégorie « population hors du champ d'intérêt ».

[7] En fait, on soupçonne fortement que les immigrants qui n'ont pu être dépistés avaient quitté le Canada, ce qui équivaut à un mécanisme dépendant non ignorable. Toutefois, le petit nombre de cas disponibles a rendu presque impossible l'estimation robuste des paramètres de « non-ignorabilité » dans le modèle conjoint utilisé dans ce genre de situation.

L'étape suivante consiste à estimer $\hat{\eta}'_k = \left(1+\exp\left[-\left(\hat{\alpha}_0 + \hat{\alpha}_k X'_k\right)\right]\right)^{-1}$, pour $k \in U$, c'est-à-dire la probabilité d'être comprise dans la PI pour chaque unité de la partie non résolue de l'échantillon. La situation prévue, c'est-à-dire dans la PI ou dans la PHI, est alors obtenue au moyen d'un essai de Bernoulli ($\hat{\eta}'_k$). Comme le montre le diagramme 3, après cette première étape, on a attribué à toutes les unités une situation indiquant si elle faisait partie de la PI ou de la PHI. Il convient de souligner que les prédicteurs les plus significatifs d'être un immigrant faisant partie de la population résidant au Canada sont, par ordre d'importance, le *niveau de scolarité*, le *groupe d'âge* et la *composante d'immigration*.

Diagramme 3 : Situation finale pour toutes les unités d'échantillonnage



8.3 Création de classes de pondération fondées sur les modèles prédictifs

Comme nous le mentionnons à la section 6.2, nous construisons des classes d'ajustement de la pondération en émettant l'hypothèse qu'elles doivent être homogènes par rapport à la correction apportée (et sont donc appelées groupes homogènes ou GH). Les classes de repondération pour la non-réponse de l'ELIC sont construites en se fondant sur les unités ayant la même propension à répondre. Les classes de repondération pour la non-résolution des cas sont construites en se fondant sur la même propriété d'homogénéité, à part le fait qu'elles comportent des unités ayant « la même propension à être résolue ». Plusieurs méthodes peuvent être appliquées pour former les groupes homogènes. Ainsi, les algorithmes d'arbres de décision, comme CHAID dans le logiciel Knowledge Seeker (Kass, 1980; Angoss Software, 1995) et les modèles de régression logistique ont été utilisés à grande échelle pour créer les groupes homogènes. Une autre méthode, proposée récemment par Eltinge et Yansaneh (1997) produit des classes contenant des unités ayant la même probabilité de réponse. Ces probabilités sont les valeurs prévues d'après un modèle de régression logistique. Pour l'ELIC, les groupes homogènes pour l'ajustement pour la non-réponse et pour la non-résolution des cas sont définis selon la méthode d'Eltinge-Yanseneh.

Classes d'ajustement de la pondération pour la non-réponse

Soit X''_j l'ensemble de variables auxiliaires disponibles pour tous les *cas résolus dans la PI* (c.-à-d. les répondants et les non-répondants). X''_j est utilisé pour ajuster le modèle de régression logistique suivant :

$$P\left[RR_j = 1 \middle| X''_j\right] = \left(1+\exp\left[-\left(\alpha_0 + \alpha_j X''_j\right)\right]\right)^{-1}$$

où $RR_j = \begin{cases} 1 \text{ si j} & \text{RR} \\ 0 \text{ si j} & \text{RN.} \end{cases}$

D'après le modèle, nous obtenons $\hat{\eta}''_j = \left(1+\exp\left[-\left(\hat{\alpha}_0 + \hat{\alpha}_j X''_j\right)\right]\right)^{-1}$, pour $j \in \text{RPI}$. Autrement dit, nous obtenons la probabilité d'être un répondant pour tous les cas résolus dans la population d'intérêt. Nous trions les unités $j \in RPI$ d'après le $\hat{\eta}''_j$ associé. Puis, nous suivons les mêmes étapes que celles décrites dans la méthode d'Eltinge-Yanseneh, à savoir :

Étape 1 : Construire C classes d'ajustement de la pondération contenant chacune le même nombre d'unités. Nous choisissons la méthode des quantiles égaux, en raison du contrôle sur le nombre prévu de répondants dans chaque cellule. Calculer les estimations de la moyenne corrigée d'après les C classes. Six variables ont été utilisées pour l'ELIC.
Étape 2 : Répéter l'étape 1 avec C+1 classes, jusqu'à ce que toutes les estimations paraissent constantes.

Les moyennes estimées des six variables utilisées pour les diagnostics ont convergé peu après que le nombre de classes, C, soit supérieur à 11. Notons que les prédicteurs les plus significatifs de l'état de répondant sont, par ordre d'importance, la *composante d'immigration*, le *groupe d'âge*, le *niveau de scolarité*, la *connaissance des langues officielles* et la *langue maternelle*.

Classes d'ajustement de la pondération pour les cas non résolus

Une approche comparable est suivie pour définir les classes d'ajustement de la pondération pour la partie non résolue de l'échantillon. Soit X_i''' l'ensemble de variables auxiliaires disponibles pour toutes les *unités dans la population d'intérêt* (c.-à-d. les cas résolus-PI et les cas prévus-PI). X_i''' est alors utilisé pour ajuster le modèle de régression logistique suivant[8]:

$$P\left[RES_i = 1 \middle| X_i'''\right] = \left(1 + \exp\left[-\left(\alpha_0 + \alpha_i X_i'''\right)\right]\right)^{-1}$$

$$\text{où } RES_i = \begin{cases} 1 \text{ si } i \in RPI \\ 0 \text{ si } i \in UPI. \end{cases}$$

D'après le modèle, nous obtenons $\hat{\eta}_i''' = \left(1 + \exp\left[-\left(\hat{\alpha}_0 + \hat{\alpha}_i X_i'''\right)\right]\right)^{-1}$, pour $i \in$ PI, pour $PI = \text{UPI} + \text{RPI}$. Autrement dit, nous obtenons la probabilité d'être un cas résolu pour toutes les unités comprises dans la population d'intérêt. Nous trions les unités $i \in PI$ d'après le $\hat{\eta}_i'''$ associé. Les mêmes étapes 1 et 2 que celles décrites plus haut sont suivies. La convergence des moyennes ajustées estimées a eu lieu lorsque C a été supérieur à 12. Notons que les prédicteurs les plus significatifs d'être un cas résolu sont, par ordre d'importance, la *qualité de la source de données de dépistage*, le *mois de référence* et le *nombre d'années de scolarité*. Le *niveau de scolarité*, le *groupe d'âge* et la *composante d'immigration* ont également été inclus pour tenir compte implicitement des unités PI prévues provenant de la partie non résolue de l'échantillon.

8.4 Ajustement de la pondération pour la non-réponse et la non-résolution des cas

La procédure finale d'ajustement du poids de chaque unité répondante comprise dans la partie des cas résolus PI de l'échantillon est la suivante :

i) Ajustement dans chacune des 13 classes d'ajustement pour la non-réponse :

$$\frac{\text{somme pondérée des non - répondants + somme pondérée des répondants}}{\text{somme pondérée des répondants}}$$

ii) Ajustement dans chacune des 12 classes d'ajustement pour la non-résolution des cas:

$$\frac{\text{somme pondérée des cas non résolus prévus + somme pondérée des cas résolus}}{\text{somme pondérée des cas résolus}}$$

[8] Il convient de souligner que, puisque ce modèle comprend implicitement les unités PI prévues, nous incluons également dans le modèle l'information auxiliaire provenant du premier modèle décrit à la section 8.2.

9. CONCLUSION

Durant l'élaboration des méthodes d'enquête, il est devenu évident que l'ajustement de la pondération de l'ELIC serait difficile et nécessiterait des ajustements non conventionnels. Cependant, sans données empiriques, établir une nouvelle stratégie non biaisée n'a pas été une tâche simple. Il a fallu intégrer autant d'information que possible sur les mécanismes de réponse en veillant à ne pas créer des profils artificiels dans les données ni un trop grand nombre de cellules d'ajustement. L'utilisation de la méthode d'Eltinge-Yanseneh assure non seulement d'obtenir des classes homogènes et de taille égale, mais permet aussi d'utiliser toutes les variables explicatives possibles sans qu'elles n'aient d'effet sur l'analyse. La pondération de l'ELIC est un bon exemple de technique fondée sur un modèle utilisant toutes les données auxiliaires pertinentes. Il s'agit aussi d'un bon exemple d'étude minutieuse des profils de réponse afin de bien comprendre les mécanismes sous-jacents.

Durant les prochains cycles, les résultats de la collecte continueront d'être surveillés de près. Comme les taux de réponse à l'ELIC devraient, en principe, être plus faibles que ceux observés pour d'autres enquêtes, au début du projet, il a été décidé de procéder en conséquence à un suréchantillonnage pour le premier cycle, afin de s'assurer de pouvoir produire des estimations fiables à la fin des deuxième et troisième cycles. Si tout se passe comme prévu, la taille effective de l'échantillon sera plus grande qu'on ne s'y attendait. De nouvelles sources d'information de dépistage sont continuellement recherchées et un certain nombre seront introduites pour le troisième cycle.

RÉFÉRENCES

Angoss Software (1995), "Knowledge Seeker IV for Windows – User's Guide." Angoss Software International Limited.

Kass, G. V. (1980), "An explanatory technique for investigating large quantities of categorical data." *Applied Statistics*, Vol. 29, No.2, pp.119-127.

Eltinge, J.L. et Yansaneh, I.S. (1997), "Diagnostics for formation of non-response adjustment cells with an application to income non-response in the U.S. Consumer Expenditure Survey." *Survey Methodology*, June 1997, Vol. 23, No. 1, pp. 33-40.

Little, R.J.A. (1982), "Models for nonresponse in sample surveys." *Journal of the American Statistical Association*, Vol. 77, pp. 237-250.

LSIC microdata user guide (2003), *available upon request.*