Guide de l'utilisateur des microdonnées

ENQUÊTE LONGITUDINALE AUPRÈS DES IMMIGRANTS DU CANADA

VAGUE 3

Statistique Canada Statistics Canada

Canada

Table des matières

1.0 Introduction

L'Enquête longitudinale auprès des immigrants du Canada (ELIC), qui est effectuée conjointement par Statistique Canada et Citoyenneté et Immigration Canada dans le cadre du Projet de recherche sur les politiques, est une enquête détaillée qui vise à étudier le processus d'adaptation des nouveaux immigrants à la société canadienne.

Le présent guide a été élaboré afin de faciliter la manipulation du fichier de microdonnées des résultats de la troisième vague de l'ELIC. Ce fichier contient les données des trois vagues de collecte de l'enquête. La collecte a été menée par Statistique Canada, la première vague s'étant déroulée entre avril 2001 et mai 2002, la deuxième entre décembre 2002 et décembre 2003 et la troisième entre novembre 2004 et novembre 2005.

Toutes les questions concernant le fichier de données ou son utilisation devraient être adressées à :

Statistique Canada
Services à la clientèle
Division des enquêtes spéciales
Téléphone : (613) 951-3321 ou appelez sans frais : 1 800 461-9050
Télécopieur : (613) 951-4527
Courriel : des@statcan.ca

2.0 Contexte

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) est une enquête détaillée qui vise à étudier le processus grâce auquel les nouveaux immigrants s'adaptent ou s'intègrent à la société canadienne. Lors de leur adaptation à la vie au Canada, de nombreux immigrants font face à divers défis : trouver un logement approprié, apprendre ou mieux parler une des langues officielles du Canada ou les deux, participer au marché du travail ou avoir accès aux études et à la formation. Les résultats de cette enquête fournissent une indication de la manière dont les immigrants relèvent ces défis ainsi que du genre de ressources qui facilitent leur établissement au Canada. L'enquête examine également la façon dont les caractéristiques socio-économiques des immigrants influent sur le processus d'établissement.

Les sujets abordés dans l'enquête comprennent les compétences linguistiques, le logement, la scolarité, la reconnaissance des titres de compétences acquis à l'étranger, l'emploi, la santé, les valeurs et attitudes, la création et l'utilisation de réseaux sociaux, la citoyenneté, le revenu et les impressions sur la vie au Canada. Les questions portent sur la situation du répondant avant et après son arrivée au Canada.

Exception faite du module sur le revenu, où on demande à la personne la mieux renseignée de répondre, aucune interview par procuration n'est acceptée. L'unité d'analyse principale est toujours l'immigrant sélectionné, qu'on appelle le répondant longitudinal (RL), et ce, même si certaines questions ont trait à l'expérience d'autres membres du ménage tels que le/la conjoint(e) ou les enfants.

3.0 Objectifs

Il y a un besoin grandissant d'information sur les nouveaux immigrants du Canada. Bien que la pleine intégration puisse être l'affaire de plusieurs générations, l'Enquête longitudinale auprès des immigrants du Canada (ELIC) s'intéresse au processus d'établissement durant les quatre premières années suivant l'arrivée au pays, période cruciale durant laquelle les nouveaux arrivants nouent des liens économiques, sociaux et culturels avec la société canadienne. À cet égard, le but de l'enquête est double :

- examiner comment les nouveaux immigrants s'adaptent à la vie au Canada au fil du temps; et
- fournir des renseignements sur les facteurs susceptibles de favoriser ou d'entraver cette adaptation.

3.1 Avantages des enquêtes longitudinales

Par sa nature longitudinale, l'ELIC présente certains avantages. Ainsi, la collecte de données auprès d'une même cohorte d'immigrants lors d'occasions successives permet de mesurer directement, et de façon plus efficace, le processus d'établissement que si un échantillon différent de la même population d'immigrants était tiré à chaque interview. Le gain d'efficacité a toutefois un prix. Dans certaines situations, les hypothèses sur lesquelles reposent les modèles analytiques conventionnels ne sont plus valides, en particulier, lorsqu'une réponse dépendante du temps — par exemple, une période sans emploi — et non pas le changement lui-même nous intéresse. Dans de tels cas, un seul immigrant peut contribuer plus qu'une observation à l'analyse (mesures répétées). De plus, en raison des complexités du plan de sondage et de la pondération de l'ELIC, if faut considérer d'autres facteurs dans l'analyse des données.

Les données recueillies lors de la troisième vague d'entrevues permettront aux chercheurs d'examiner les changements qui ont eu lieu dans la vie des immigrants de l'ELIC pendant leurs quatre premières années au Canada, et d'étudier l'impact de ces changements sur le processus d'établissement. Par exemple, la reconnaissance des titres de compétences ou des diplômes, la poursuite d'études et l'expérience professionnelle pourraient être utilisées afin d'examiner le succès sur le marché du travail.

Plusieurs types d'analyses sont possibles en utilisant les données longitudinales. Par exemple, une simple analyse descriptive pourrait estimer le nombre d'immigrants dont le plus haut niveau scolaire atteint a changé entre les vagues de l'enquête. Les changements individuels dans le plus haut niveau scolaire atteint pourraient ensuite être combinés à d'autres données socio-démographiques et économiques afin de modéliser la probabilité que, durant les quatre premières années au Canada, l'immigrant ait pu se trouver un emploi. En profitant de l'historique des emplois dans la Liste d'emplois, ces mêmes informations pourraient également être utilisées afin d'examiner la durée de temps, en semaines, pour se trouver un emploi (souvent appeler analyse de données de survie, analyse de durée ou analyse historique d'événements).

Il y a une grande variété de documentation traitant de l'analyse de données longitudinales. Diggle et al (2002) examinent certains enjeux reliés à l'analyse des données longitudinales. Korn et Graubard (1999) discutent l'analyse des données d'enquêtes, en présentant des exemples et des points à considérer lors de l'analyse de données longitudinales provenant d'enquêtes à plan complexe. Allison (1999) donne un guide pratique pour l'ajustement des modèles de survie en utilisant le système SAS. Le développement des modèles de survie dans le cas des données d'enquêtes à plan complexe est examiné par Lawless et Boudreau (2002). Ces textes et articles ne constituent pas une liste complète, mais fournissent au lecteur de bonnes références initiales.

3.2 Analyse longitudinale avec les fichiers de la vague 3

Les fichiers de données de la vague 3 ont été structurés pour faciliter l'analyse longitudinale de façon similaire à ceux de la deuxième vague. L'utilisateur n'est pas requis de fusionner les informations des trois vagues : les profils de réponse complets de la vague 1, de la vague 2 et de la vague 3 sont fournis pour chaque immigrant qui a répondu à l'interview de la vague 3. De plus, chaque variable est nommée de manière à ce qu'il soit aisé d'établir la référence à une vague spécifique, ce qui facilite l'identification des questions et du contenu communs aux trois vagues. Des renseignements supplémentaires concernant la structure de la base de données sont présentés au chapitre 5.0; l'attribution des noms de variables est présentée à la section 5.4. En outre, le chapitre 6.0 traite de la comparabilité longitudinale des concepts mesurés aux trois vagues.

Pour chaque répondant, il y a une seule variable de poids longitudinal, soit WT3L (qu'on retrouve dans l'entité répondant longitudinal (RL)), qui devrait être utilisée pour toutes analyses menées sur les données longitudinales de la vague 3. On peut interpréter ce poids comme le nombre d'immigrants appartenant à la population d'intérêt de la vague 3 représenté par l'immigrant répondant. La population d'intérêt comprend tout immigrant inclus dans la cohorte de L'ELIC qui habitait toujours au Canada au moment de l'interview de la vague 3. La méthode d'obtention des poids est présentée au chapitre 12.0; l'utilisation des poids est discutée au chapitre 14.0.

En raison de la complexité du plan d'échantillonnage et des ajustements des poids, les formules traditionnelles de variance employées dans certains logiciels d'analyse ne sont pas appropriées. Des méthodes et outils spéciaux sont ainsi recommandés lors de l'analyse des données de l'ELIC; ceux-ci sont présentés au chapitre 15.0.

Références

Allison, P.D. (1999). Survival analysis using the SAS system: a practical guide, SAS Institute, Cary, N.C.

Diggle, P., Heagerty, P., Liang, K., Zeger, S. (2002). Analysis of Longitudinal Data, Oxford University Press, New York.

Korn and Graubard (1999). Analysis of Health Surveys, Wiley, New York.

Lawless, J.F., Boudreau, C. (2002). Modélisation et analyse des données sur la durée provenant d'enquêtes longitudinales. Recueil du Symposium 2002 de Statistique Canada, Statistics Canada, 11-522-XCB.

4.0 Concepts et définitions

De nombreux concepts et variables revêtent une importance cruciale du point de vue de l'analyse des données de l'Enquête longitudinale auprès des immigrants du Canada (ELIC). Les principaux concepts qui sous-tendent l'ELIC sont expliqués ci-après.

Agent d'immigration : Fonctionnaire canadien qui traite la demande de l'immigrant au moment de son arrivée au Canada.

Attestation d'études : Le plus haut niveau de scolarité lorsque supérieur à un diplôme d'études secondaires, les diplômes d'études professionnelles ou techniques et tout autre diplômes ou certificats d'études ou de formation obtenus à l'extérieur du Canada constituent des attestations d'études.

***Entièrement reconnue* :** L'employeur ou l'établissement reconnaît l'attestation d'études comme légitime selon certaines normes.

***Partiellement reconnue* :** L'employeur ou l'établissement reconnaît en partie l'attestation d'études comme étant légitime selon certaines normes.

***Non reconnue* :** L'attestation d'études n'est pas reconnue comme étant légitime selon certaines normes.

Catégories d'immigration :

***Catégorie économique* :** Catégorie d'immigrants sélectionnés en fonction de leurs compétences ou d'autres atouts permettant de contribuer à l'économie canadienne (comprends les travailleurs qualifiés, les investisseurs, les entrepreneurs et les travailleurs autonomes).

Catégorie de la famille : Catégorie d'immigrants parrainés par de proches parents ou des membres de la famille vivant au Canada.

Catégorie des immigrants indépendants : Catégorie d'immigrants qui possèdent les compétences nécessaires à l'exercice de certains emplois ou qui représentent un atout important pour le Canada. Ces personnes présentent une demande de leur propre chef ou ont des parents éloignés vivant au Canada.

Catégorie des réfugiés : Catégorie de personnes qui demandent la protection du Canada.

Citoyenneté : Qualité de citoyen – né au pays ou naturalisé – qui confère à un individu les mêmes droits, privilèges et responsabilités que ceux des autres individus.

Conseiller en immigration : Professionnel qui donne des conseils ou fournit des services relativement à des questions touchant l'immigration.

Discrimination : Traitement défavorable d'une personne en raison de ses caractéristiques personnelles telles que la race ou la couleur de sa peau, l'origine ethnique ou culturelle, la langue ou l'accent, la religion, etc.

Emploi à temps partiel : Emploi occupé par une personne qui travaille habituellement moins de 30 heures par semaine à son emploi principal ou à son unique emploi.

Emploi à temps plein : Emploi occupé par une personne qui travaille habituellement 30 heures ou plus par semaine à son emploi principal ou à son unique emploi.

Famille de recensement : Couple marié (avec ou sans enfants des deux conjoints ou de l'un d'eux), couple vivant en union de fait (avec ou sans enfants des deux partenaires ou de l'un d'eux) ou parent seul (peu importe son état matrimonial) demeurant avec au moins un enfant dans le même logement. Un couple vivant en union libre peut être de sexe opposé ou de même sexe. Les « enfants » dans une famille de recensement incluent les petits-enfants vivant dans le ménage d'au moins un de leurs grands-parents en l'absence des parents. Dans l'enquête, la famille de recensement est aussi désignée sous le nom de « famille immédiate ».

Famille économique : Groupe de deux personnes ou plus qui vivent dans le même logement et qui sont unis par le sang, par alliance, par union libre ou par adoption.

Groupe de population : Groupe de population auquel le répondant appartient. Il comprend les minorités visibles (voir la définition ci-dessous) ainsi que les Autochtones et les personnes qui sont de race blanche ou qui ont la peau blanche.

Groupe ethnique ou culturel : Groupe de personnes ayant en commun une culture distincte. Le terme « groupe ethnique ou culturel » implique que les valeurs, les normes, le comportement et la langue, *et pas nécessairement l'apparence physique*, sont les caractéristiques distinctives importantes.

Intégration (ou établissement) : Processus grâce auquel les nouveaux arrivants participent à la vie communautaire au Canada et à son façonnement.

Minorité visible : Font partie des minorités visibles les personnes, autres que les Autochtones, qui ne sont pas de race blanche ou qui n'ont pas la peau blanche.

Nouveau membre :

> **Vague 1 :** Personne qui vit au sein du ménage du répondant, mais qui ne faisait pas partie de l'unité immigrante du répondant longitudinal (RL). Il peut s'agir de personnes qui vivaient déjà au Canada au moment de l'arrivée du RL.
>
> **Vagues 2 et 3 :** Personne qui vit au sein du ménage du répondant, mais qui ne faisait pas partie du ménage du RL à la vague précédente. Il peut s'agir de personnes qui vivaient déjà au Canada au moment de l'arrivée du RL.

Organisme d'aide aux immigrants ou réfugiés : Corps constitué qui aide les immigrants ou réfugiés à répondre à leurs besoins.

Parrain : Citoyen canadien ou résident permanent âgé de 19 ans ou plus vivant au Canada qui s'engage à fournir à l'immigrant parrainé l'aide essentielle sous forme de logement, de vêtements, de nourriture et d'aide à l'établissement au cours d'une période donnée.

Partenaire en union libre : Personne qui n'est pas légalement mariée au répondant, mais qui vit au sein du ménage comme conjoint. Cette personne peut être de même sexe ou de sexe opposé.

Période de référence : C'est le point ou la période dans le temps dans lequel s'inscrit une question dans l'Enquête. En d'autres mots, c'est la date et la durée à laquelle une question est circonscrite (c.-à-d. la période de temps couverte par une question). Par exemple : la période entre les deux premières entrevues. Les périodes de référence peuvent changer d'une vague à l'autre pour la même question.

Personne ayant déménagé : Personne qui faisait partie de l'unité immigrante du répondant longitudinal, mais qui ne vivait pas au sein du même ménage au moment de l'interview.

PMR : Personne la mieux renseignée sur un sujet précis. Les seules questions de l'ELIC qui sont posées à la PMR sont celles du module Revenu qui ont trait au revenu familial. Si la PMR n'est pas disponible, on pose ces questions au RL.

Programme d'accueil : Programme en vertu duquel on jumelle le nouvel arrivant à un bénévole qui connaît bien les coutumes canadiennes et qui peut le renseigner sur les services offerts, le mettre en relation avec des personnes, l'aider à trouver du travail ou un logement, etc. Ce programme vise à faciliter l'intégration des nouveaux arrivants.

Programme d'établissement et d'adaptation des immigrants (PEAI) : Programme permettant d'allouer des fonds pour offrir aux nouveaux arrivants des services directs et essentiels : accueil et orientation, traduction et interprétation, aiguillage vers des services communautaires, counselling professionnel, renseignements d'ordre général, services reliés à l'emploi, etc.

Répondant longitudinal (RL) : Personne sélectionnée pour répondre aux questions de l'ELIC dans le cadre de chacune des trois vagues.

SSOBL : Cet acronyme désigne le « Système de soutien des opérations des bureaux locaux », c'est-à-dire la base de données administratives de Citoyenneté et Immigration Canada. Le SSOBL a servi de base de sondage pour l'ELIC.

Unité immigrante : Groupe de personnes qui ont présenté une demande pour venir au Canada en vertu du même visa et, aux fins de l'enquête, qui sont arrivés au pays en compagnie du répondant longitudinal ou trois mois avant ou après ce dernier.

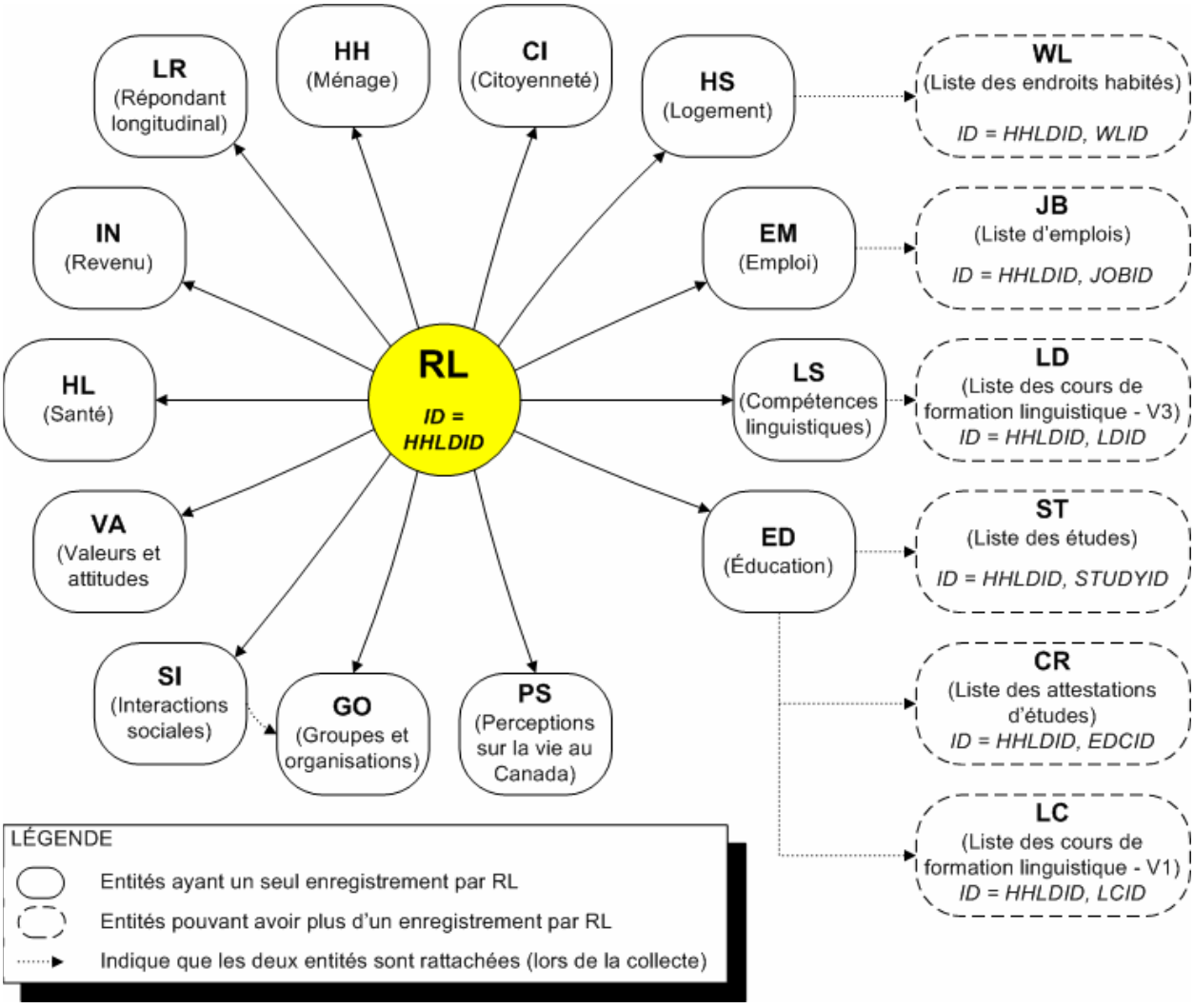
5.0 Structure et contenu des fichiers

Le fichier de la troisième vague de l'Enquête longitudinale auprès des immigrants du Canada (ELIC) contient les données de toutes les vagues de collecte de l'Enquête. Il contient tous les enregistrements relatifs aux répondants qui ont été retracés et ont accepté de répondre aux trois vagues, soit 7,716 répondants. Le fait d'avoir les données des trois vagues fusionnées dans le même fichier facilite les analyses longitudinales.

5.1 Le modèle de données

Les données de l'ELIC ont été réparties en différentes bases de données plus petites, appelées entités. Cette structure, que l'on appelle le modèle de données, constitue une façon intuitive et pratique de stocker des données. Chaque entité englobe les variables relatives à des concepts généraux qui imitent sensiblement la structure des modules du questionnaire. La figure suivante présente les entités de l'ELIC et leur organisation structurelle :

Figure 5.1 Structure des données



L'identificateur du répondant longitudinal (RL), la variable « HHLDID », est la variable qui lie toutes les entités. Pour la majorité des entités, il n'y a qu'un seul enregistrement par répondant longitudinal (RL). Certaines entités ont plus d'un enregistrement par RL, il s'agit des entités-listes, ou fichiers d'évènements (WL, JB, CR, ST LC, LD). Dans ces fichiers, le nombre d'enregistrement par RL va de zéro à plusieurs. Chacune des entités-listes était rattachée à un module particulier lors de la collecte. Cela a un impact au niveau de l'imputation notamment : si le module ou le fichier d'évènement s'avère incomplet, les deux seront imputés (voir le chapitre 11.0 sur l'imputation).

Tableau 5.1 Liste des entités du modèle de données et de leur contenu

Entité	Identificateurs - clés uniques	Nom du fichier		Provenance du contenu recueilli ou dérivé		
		Format texte	Format SAS	Questionnaire Vague 1	Questionnaire Vague 2	Questionnaire Vague 3
CI Citoyenneté	HHLDID	MAIN	CI	Module antécédent : BG_Q06 à BG_Q09B, BG_Q16 et BG_Q17 Module Valeurs et Attitudes : VAS_Q01 à VAS_Q04A	Module Citoyenneté	Module Citoyenneté
CR Liste des attestations d'études	HHLDID, EDCID	CR	CR	Attestation d'études (sous-module du module Éducation)	Attestation d'études (sous-module du module Éducation)	Attestation d'études (sous-module du module Éducation)
ED Éducation	HHLDID	MAIN	ED	Module Éducation	Module Éducation	Module Éducation
EM Emploi	HHLDID	MAIN	EM	Module Emploi	Module Emploi	Module Emploi
GO Groupes et organisations	HHLDID	MAIN	GO	Groupes et organisations (sous-module du module Interactions sociales)	Groupes et organisations (sous-module du module Interactions sociales)	Groupes et organisations (sous-module du module Interactions sociales)
HH Ménage	HHLDID	HH	HH	Composante entrée (y compris la matrice de relation entre les membres du ménage)	Composante entrée (y compris la matrice de relation entre les membres du ménage)	Composante entrée (y compris la matrice de relation entre les membres du ménage)
HL Santé	HHLDID	MAIN	HL	Module Santé	Module Santé	Module Santé
HS Logement	HHLDID	MAIN	HS	Module Logement Module Antécédents : BG_Q14 et BG_Q15	Module Logement	Module Logement
IN Revenu	HHLDID	MAIN	IN	Module Revenu	Module Revenu	Module Revenu
JB Liste d'emplois	HHLDID, JOBID	JB	JB	Détails sur l'emploi et Liste d'emplois (sous-modules du module Emploi)	Détails sur l'emploi et Liste d'emplois (sous-modules du module Emploi)	Détails sur l'emploi et Liste d'emplois (sous-modules du module Emploi)

Entité	Identificateurs - clés uniques	Nom du fichier		Provenance du contenu recueilli ou dérivé		
		Format texte	Format SAS	Questionnaire Vague 1	Questionnaire Vague 2	Questionnaire Vague 3
LC Liste des cours de langue seconde de Vague 1	HHLDID, LCID	LC	LC	Liste d'éducation (sous-modules du module Éducation) pour ce qui concerne les cours de formation linguistique seulement.	Seul un suivi est fait en ce qui concerne les cours de formation linguistique ayant débuté à la vague 1 à travers les modules suivants : Liste précédente d'éducation, Renseignements détaillés et Liste d'éducation (sous-modules du module Éducation). *Les cours de formation linguistique ne sont pas recueillis à la Vague 2.*	Certaines informations du module Compétences linguistiques sont utilisées pour combler de l'information manquante en ce qui concerne les cours de formation linguistique ayant débutés à la vague 1 et se continuant jusqu'à la vague 3. *Voir l'entité LD pour les cours de formation linguistique suivis à la Vague 3.* Note : Des dates de fin manquantes pour les cours de formation linguistique ayant débuté à la vague 1 et s'étant terminés à la vague 2 sont recueillis par l'entremise du module *Liste précédente d'éducation vague 1* du questionnaire de vague 3.
LD Liste des cours de langue seconde de Vague 3	HHLDID, LDID	LD	LD	Sans objet	Sans objet	Renseignements détaillés sur la formation linguistique (sous-module du module Compétences linguistiques)
LR Répondant longitudinal	HHLDID	MAIN	LR	Composante entrée et module antécédents : BG_Q01 à BG_Q05 et BG_Q18 à BG_Q20, en plus de certaines variables provenant de la base de données administrative de Citoyenneté et Immigration Canada.	Composante entrée	Composante entrée
LS Compétences linguistiques	HHLDID	MAIN	LS	Module Compétences linguistiques, excluant les questions sur les tests linguistiques : LS_Q11E à LS_Q16E et LS_Q11F à LS_Q16F	Module Compétences linguistiques	Module Compétences linguistiques

Entité	Identificateurs - clés uniques	Nom du fichier		Provenance du contenu recueilli ou dérivé		
		Format texte	Format SAS	Questionnaire Vague 1	Questionnaire Vague 2	Questionnaire Vague 3
PS Impressions sur la vie au Canada	HHLDID	MAIN	PS	Module Impressions sur la vie au Canada	Module Impressions sur la vie au Canada	Module Impressions sur la vie au Canada
SI Interactions sociales	HHLDID	MAIN	SI	Module Interactions sociales	Module Interactions sociales	Module Interactions sociales
ST Liste des études	HHLDID, STUDYID	ST	ST	Renseignements détaillés et Liste d'éducation (sous-modules du module Éducation) pour ce qui concerne tous les types de cours à l'exception de la formation linguistique	Liste précédente d'éducation, Renseignements détaillés et Liste d'éducation (sous-modules du module Éducation) pour ce qui concerne tous les types de cours à l'exception de la formation linguistique	Liste précédente d'éducation, Renseignements détaillés et Liste d'éducation (sous-modules du module Éducation) pour ce qui concerne tous les types de cours à l'exception de la formation linguistique. Note : Des dates de fin manquantes pour les cours ayant débuté à la vague 1 et s'étant terminés à la vague 2 sont recueillis par l'entremise du module *Liste précédente d'éducation vague 1* du questionnaire de vague 3.
VA Valeurs et attitudes	HHLDID	MAIN	VA	Module Valeurs et attitudes, à l'exclusion de VAS_Q01 à VAS_Q04A	Module Valeurs et attitudes	Module Valeurs et attitudes
WL Liste des endroits habités	HHLDID, WLID	WL	WL	Endroit habité (sous-module du module Logement)	Endroit habité (sous-module du module Logement)	Endroit habité (sous-module du module Logement)

5.2 Format des fichiers

Les fichiers de l'ELIC sont disponibles sous deux formes :

1. **Les fichiers textes (en format ASCII)**
   Toutes les entités à l'exception de l'entité contenant l'information sur le ménage du répondant (HH) et les entités-listes (CR, JB, ST, WL, LC, LD) sont contenues dans un grand fichier texte (MAIN). Les autres entités ont leur propre fichier texte séparément. Il existe des cartes de syntaxe SAS et SPSS qui permettent le formatage de ces fichiers (dont le nom se termine par SASE et SPSSE pour les cartes de syntaxe anglaises et par SASF et SPSSF pour les cartes de syntaxe françaises).

2. **Les fichiers en format SAS**
   Chaque entité constitue un fichier individuel, tel que décrit au tableau 5.1. Tous les fichiers de l'ELIC comportent un identificateur clé unique qu'on appelle identificateur de ménage (nom de variable HHLDID) et qui est propre au répondant longitudinal. Tous les

fichiers peuvent être fusionnés à l'aide de cette variable clé. Les entités-listes comportent en plus d'autres identificateurs afin que chaque enregistrement soit unique.

5.3 Contenu des fichiers

Toutes les entités à l'exception des entités-listes contiennent un enregistrement par répondant longitudinal ayant fourni des réponses aux trois vagues de l'ELIC, soit 7 716 enregistrements. Les variables de chaque vague ont des noms uniques et sont combinées pour former les fichiers de vague 3. En d'autres mots, chaque entité contient les variables de la vague 1, de la vague 2 et de la vague 3.

Les entités-listes, soit Liste des attestations d'études (CR), liste d'emploi (JB), liste d'études (ST), liste des endroits habités (WL), liste des cours de langue seconde de Vague 1 (LC) et liste des cours de langue seconde de Vague 3 (LD) peuvent contenir plus d'un enregistrement par répondant. Dans ces entités, le nombre minimal d'enregistrements pour un répondant est de zéro et le maximum recueilli varie selon l'entité (CR = 7, JB = 13, ST = 12, WL = 5, LC = 3, LD = 4).

Il est important de noter que pour produire les estimations, le poids final ne doit être utilisé directement que sur les enregistrements des RL. Aucune estimation pondérée ne peut être produite directement à partir des enregistrements des entités-listes. Pour plus de détails, veuillez vous référer au chapitre 12.0 traitant de la pondération.

5.4 Structure des variables

Afin de faciliter l'interprétation des données par les utilisateurs, l'attribution des noms de variable et des valeurs est régie par certaines règles dans le système de documentation du fichier de microdonnées de l'ELIC. Tout d'abord, chaque nom de variable contient un identificateur (identificateur de l'item – voir le chapitre 6.0 traitant de la comparabilité longitudinale) qui permet d'identifier les variables liées longitudinalement d'une vague à l'autre, c'est-à-dire mesurant des phénomènes similaires mais à des périodes différentes.

Tous les noms de variable comptent au plus huit caractères (la plupart en compte sept), ce qui permet d'utiliser facilement ces noms avec des progiciels d'analyse comme SAS ou SPSS.

Voici une description de la structure du nom des variables :

- Les **deux premiers** caractères constituent l'acronyme de l'entité à laquelle l'élément appartient. Voir le tableau 5.1 pour les descriptions.

- Le **troisième** caractère du nom de la variable désigne la vague de l'ELIC :

  « 1 » indique qu'il s'agit de la première vague,
  « 2 » indique qu'il s'agit de la deuxième vague, et
  « 3 » indique qu'il s'agit de la troisième vague.

- Le **quatrième** caractère indique le type de variable. On distingue six types de variable :

  c Variable codée : variable qui est codée en se fondant sur des listes exhaustives de codes standards (CTP91 – Classification type des professions, SCIAN – Système de classification des industries de l'Amérique du Nord, CPE – Classification des programmes d'enseignement, et la Liste des codes de pays du recensement).

  d Variable dérivée : variable créée à partir d'une ou plus d'une variable recueillies ou codées (p. ex., taille du ménage, situation vis-à-vis de l'activité, etc.).

l Variable dérivée longitudinale : variable créée à partir d'une ou plus d'une variable combinant des données de plus d'une vague (p. ex, nombre de semaine au travail depuis l'arrivée au Canada (cumulatif), plus haut niveau de scolarité (information pouvant être changée à chaque vague), etc.).

g Variable de combinaison : variables recueillies, codées ou dérivées dont les catégories ont été regroupées (p. ex., groupes d'âge, régions du monde, etc.).

i Indicateur d'imputation : indique que la valeur d'une variable pour un répondant a été imputée (imputation par champs) ou que l'entité au complet a fait l'objet d'une imputation (dite massive). Les variables indicatrices d'imputation par champs suivent immédiatement les questions imputées. Par exemple, la variable indicatrice d'imputation pour IN1Q003 est INI004.

q Variable recueillie : variable contenant à une question qui a été posée directement au répondant.

z Variables obtenues par voie de couplage avec les dossiers administratifs de Citoyenneté et Immigration Canada.

- Les **cinquième, sixième et septième** caractères constituent un numéro séquentiel (commençant à 001) attribué à chaque variable du fichier. Au sein d'une même entité, ce numéro reste le même d'une vague à l'autre pour toute variable liée longitudinalement. L'ordre des variables dans le fichier ne correspond cependant pas à ce numéro séquentiel, mais reflète davantage une suite logique basée selon les thèmes et sur l'ordre des questions dans le questionnaire. Les modifications apportées au questionnaire aux vagues 2 et 3 ont modifié sensiblement l'ordre des questions initialement utilisé à la vague 1.

- Le **huitième** et dernier caractère (une lettre) sert à indiquer que des modifications importantes apportées à une variable entre deux vagues pouvant influer sur la comparabilité de deux variables. Si une modification a pour effet de modifier le sens d'une question ou les valeurs qui s'y rattachent, la variable est traitée comme nouvelle et est affublée d'un « x » ou d'un « y » (« x » si le nouvel item a été créé à la vague 2 et « y » s'il a été créé à la vague 3). La question à savoir quand une nouvelle variable doit être créée (renommée) est discutée au chapitre 6.0.

Tableau 5.3 Exemples de noms de variable

Exemple 1: Variable CI1Q002	
CI	Variable tirée de l'entité Citoyenneté
1	Variable de la première vague
Q	Tirée directement d'une question (faisant partie du questionnaire)
002	Variable numéro 002 de l'entité Citoyenneté

Exemple 2: Variable HL2D004x	
HL	Variable tirée de l'entité Emploi
2	Variable de la deuxième vague
D	Variable dérivée
004	Variable numéro 004 de l'entité Emploi
x	Signifie que cette variable est semblable, mais non tout à fait identique à la variable HL1Q004. Dans ce cas-ci, la formulation a été modifiée dans le questionnaire de la deuxième vague. Cette modification a été jugée assez importante pour qu'une nouvelle variable soit créée. Notez que cette modification est notée dans la Table de concordance (voir la section 6.5 pour plus d'information au sujet de la table de concordance).

Exemple 3: Variable LS3Q093y	
LS	Variable tirée de l'entité Compétences linguistiques
3	Variable de la troisième vague
Q	Tirée directement d'une question (faisant partie du questionnaire)
093	Variable numéro 093 de l'entité Compétences linguistiques
y	Signifie que cette variable est semblable, mais non tout à fait identique à la variable LS2Q093. Dans ce cas-ci, il y a eu un changement important dans les catégories de réponse. Cette modification a été jugée assez importante pour qu'une nouvelle variable soit créée. Notez que cette modification est notée dans la Table de concordance (voir la section 6.5 pour plus d'information au sujet de la table de concordance).

6.0 La comparabilité longitudinale

6.1 Les changements dans le questionnaire

Dans les enquêtes longitudinales, une règle générale est que les questions restent identiques d'une vague à une autre, à l'exception de la période de référence. Ainsi, les variables créées à chaque vague mesurent le même phénomène, mais à des temps différents, ce qui permet aux utilisateurs d'effectuer des analyses longitudinales.

Toutefois, de nombreuses révisions ont eu lieu dans le questionnaire de l'Enquête longitudinale auprès des immigrants du Canada (ELIC) de la deuxième vague et aussi, de façon moins importante, dans celui de la troisième vague. Bien que ces changements aient généralement pour but d'apporter des améliorations dans la compréhension des questions, ils sont susceptibles d'affecter la comparabilité longitudinale.

6.2 La nomenclature : les items et les variables

Par souci de clarté, on a développé une nomenclature qui s'avère fort pratique pour saisir la perspective longitudinale des variables. On fait une distinction entre ce que l'on appelle les « items » et les « variables ». Un *item* réfère à un phénomène précis, mesuré de façon spécifique, chez un ensemble donné de répondants et pour une période de référence spécifique. Quant à la *variable*, elle est la représentation, ou la mesure d'un *item* à une vague.

Par ailleurs, l'identificateur de l'*item* est imbriqué dans le nom de la *variable*, ce qui permet de les relier de façon intuitive. En effet, les deux premiers caractères (qui identifient l'entité) lorsque combinés aux cinquièmes, sixièmes et septièmes caractères (qui identifient la *variable* au sein de l'entité), se trouvent en fait à identifier l'*item*, et resteront toujours invariables d'une vague à l'autre. Par exemple, la *variable* HH1Q009 mesure le nombre de personnes dans le ménage tel que rapporté à la vague 1, tandis que la *variable* HH2Q009 mesure la même chose, mais tel que rapporté à la vague 2. On parlera donc des *variables* HH1Q009 et HH2Q009, et de façon plus générale de l'*item* HH_009, qui représente le nombre de personnes dans le ménage à une vague.

6.3 L'importance des changements

Les modifications apportées au questionnaire peuvent compromettre la comparabilité longitudinale. Des changements mineurs affectent légèrement la façon dont un *item* est mesuré comparativement à la vague précédente. Il peut s'agir de mots utilisés dans le questionnaire ou de directives additionnelles données à l'intervieweur. Afin de permettre à l'utilisateur de juger de l'impact des modifications au questionnaire pour son analyse, les changements, mêmes mineurs, sont indiqués dans la table de concordance (voir la section 6.5 concernant la table de concordance).

Toutefois, certaines modifications importantes dans le questionnaire font en sorte que l'on ne peut plus associer les *variables* qui en découlent à des *items* déjà existants à une vague précédente. Certains *items* deviennent alors obsolètes et il a est devenu nécessaire de créer de nouveaux *items*. On procède ainsi afin que les utilisateurs puissent aisément reconnaître les comparaisons longitudinales qui peuvent s'avérer suspectes, c'est-à-dire celles qui comparent des *variables* mesurant des *items* différents.

Lorsqu'un nouvel *item* similaire à un item déjà existant doit être créé, il garde un identificateur similaire. En effet, le nouvel *item* prend alors le même identificateur que l'*item* similaire qu'il remplace, à l'exception qu'un « x » ou un « y » a été ajouté à la fin, c'est-à-dire au huitième caractère du nom de la variable. Dans ce contexte, un « x » ou un « y » à la fin d'une variable

peut-être vu comme un indicateur qu'un changement important est survenu dans le questionnaire. Un « x » ou un « y » désignera un *item* nouveau, mais similaire (sans être identique) à un *item* déjà existant, et qui, le plus souvent, a dû être créé pour tenir compte d'un changement dans le questionnaire. Par exemple, la variable EM1Q049 est basée sur la question EM_Q19 du questionnaire de la vague 1, et mesure l'activité principale du répondant à la vague 1. À la vague 2, l'activité principale est mesurée par la question EM_Q02. Cependant, la question EM_Q02 de vague 2 diffère de façon importante de la question EM_Q19 de vague 1 car de nouvelles catégories de réponse ont été ajoutées. C'est pourquoi on dit que ces deux variables mesurent deux *items* différents, mais qui se ressemblent : EM_049 à la vague 1 et EM_049x à la vague 2.

Cette convention permet de préserver le lien intuitif entre deux *items* similaires, mais non identiques. Un utilisateur avisé devant se servir d'une variable arborant un « x » ou un « y » au huitième caractère, cherchera à comprendre la nature des modifications apportées. La table de concordance devrait constituer un premier arrêt, qui l'informera sur les différences entre les *items* (voir la section suivante).

Les entités constituées de listes d'évènement, soit Liste des endroits habités (WL), Liste des études (ST), Liste des attestations d'études (CR), Liste d'emplois (JB), Liste des cours de langue seconde de vague 1 (LC) et Liste des cours de langue seconde de vague 3 (LD) constituent des exceptions du fait que le huitième caractère de la variable est souvent utilisé pour identifier les évènements lorsque l'on crée un fichier plat (un enregistrement par répondant). Pour ces entités donc, les nouvelles variables ont reçu un nouvel identificateur d'item. Toutefois, dans ces cas, une note dans la table de concordance indique le lien avec une variable de la vague précédente.

6.4 La nature des changements

Conceptuellement parlant, six types de changements dans les questions ou dans les façons de poser les questions d'une vague à l'autre sont susceptibles d'affecter la comparabilité longitudinale:

1) La formulation ou le concept
Parfois, les changements dans le texte des questions ont pour but de mesurer un phénomène similaire, mais pas tout à fait identique à ce qui était mesuré à la vague précédente. D'autres fois, il s'agit simplement de clarifier des questions qui peuvent avoir créé des difficultés aux répondants lors de la collecte. Cependant, des nuances dans le texte des questions sont susceptibles d'introduire des différences dans la compréhension du phénomène et dans les comportements de réponse.

2) Les directives aux intervieweurs
Certaines questions comportent des directives spécifiques destinées aux intervieweurs. Un changement de directive peut entraîner des comportements de réponse différents. Dans la plupart des cas, il s'agit de questions où l'intervieweur doit lire les choix de réponse aux répondants pour une question à une vague, mais non à une autre.

3) Les catégories de réponse
Ce genre de modification touche seulement les questions où les répondants ne peuvent fournir qu'une seule réponse parmi un choix de réponses multiples. En général, les catégories de réponse restent identiques dans les questionnaires d'une vague à l'autre. Toutefois, certaines questions ont subi des modifications. La plupart du temps, il s'agit d'ajout de catégories, ce qui permet d'obtenir plus de détails. Lorsqu'il n'était pas possible de recréer un item en manipulant les catégories de réponse, un nouvel item a été créé.

4) L'univers

L'univers, ou la couverture, est l'ensemble des immigrants à qui les données d'une variable s'applique. Il est donc sensible aux enchaînements dans le questionnaire. Une modification de l'univers donne presque toujours lieu à la création d'un nouvel item.

5) La structure des questions

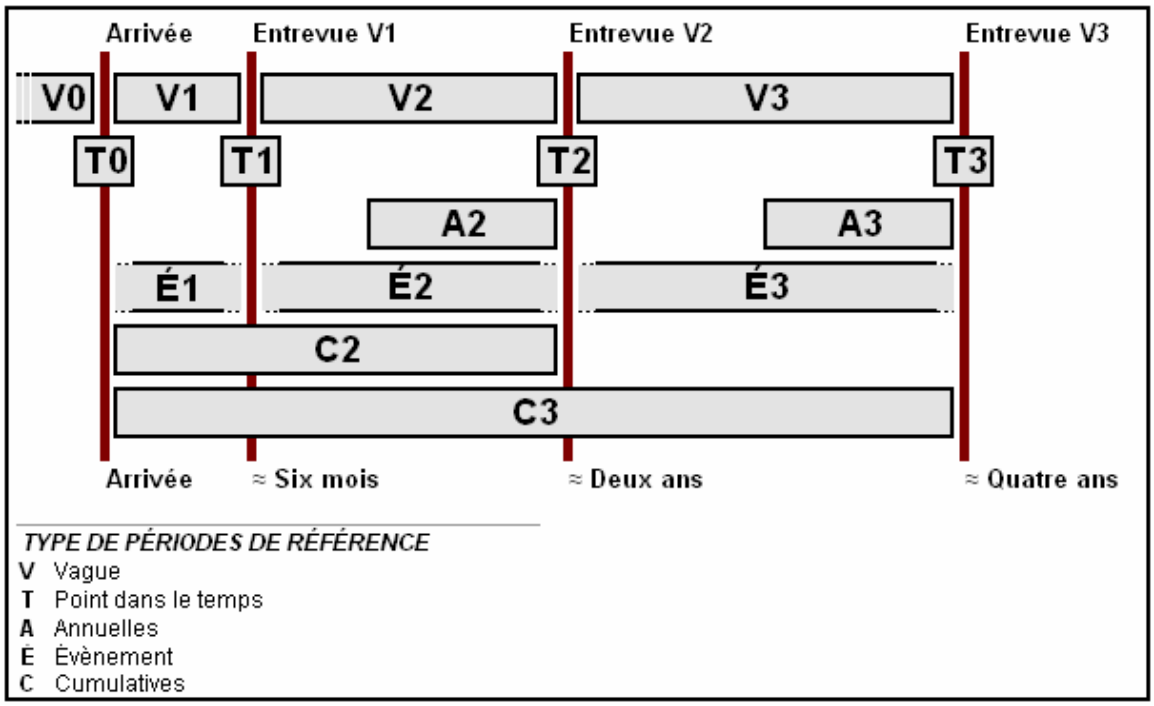
Parfois, une information recueillie à une vague peut être recueillie à la vague ultérieure de façon fort différente sans qu'il ne s'agisse d'un changement de concept ou d'univers. Il s'agit souvent de modifications dans la structure des questions.

Un exemple de cela est la variable HL2D005 : pour assurer la comparabilité longitudinale, plusieurs questions de la deuxième vague ont été combinées afin de créer une variable dérivée mesurant l'item HL_005 existant à la première vague (mesuré par la variable HL1Q005). L'information est comparable, mais la structure des questions est différente.

6) Le type de période de référence

Il est important de comprendre la dimension de temps qui se trouve au coeur des enquêtes longitudinales. Les variables de l'ELIC peuvent être catégorisées en cinq types, selon leur rapport au temps : les variables de vague, les variables annuelles, les variables se référant à un point dans le temps, les variables d'évènement et les variables cumulatives.

Figure 6.1 Types de périodes de référence utilisées dans l'ELIC



Deux variables mesurant le même item à des vagues différentes auront des périodes de référence différentes. Par contre, le type de période de référence des deux variables reste le même. En d'autres mots, pour un même item, la période de référence change pour chaque variable, au gré des vagues, mais le type de période de référence reste inchangé. Si, à cause de changements dans le questionnaire, il n'est plus possible de mesurer un item pour un même type de période de référence, un nouvel item est créé.

Les changements de type de périodes de référence sont assez rares. Les deux exemples les plus importants de changements de type de période de référence ont lieu dans le module Revenu à la vague 2 et dans le module Santé à la vague 3. Voici une description des différents types de période de référence ainsi qu'un exemple de question pour chacun :

Les variables de vague
Elles couvrent entièrement la période entre la date d'arrivée au Canada et l'entrevue de la vague 1, ou entre deux entrevues consécutives. Notez que la durée de la période de référence est inégale d'une vague à l'autre, passant d'environ six mois à la vague 1, à un an et demi à la vague 2 et trois ans à la vague 3. Par exemple : *Depuis votre dernière entrevue, vous êtes-vous fait de nouveaux amis au Canada?*

Les variables annuelles
Les variables annuelles couvrent une période de 12 mois précédant la date d'entrevue. Il est à noter qu'il n'y avait pas de variables de ce type à la première vague. Par exemple : *Au cours des 12 derniers mois, avez-vous reçu un revenu provenant de sources de l'intérieur ou de l'extérieur du Canada?*

Les variables se rapportant à un point dans le temps
Elles révèlent une situation courante au moment de l'entrevue. Par exemple : *Combien y a-t-il de pièces dans l'endroit où vous logez?*

Les variables se rapportant à un évènement
Elles réfèrent à des évènements qui ont lieu entre une date de début et une date de fin. Ces évènements sont d'une durée variable pouvant s'étendre au-delà des vagues de collecte et non spécifiée à l'avance dans le questionnaire. On retrouve ce type de variable surtout dans les entités constituées de listes d'évènements tels que Liste des études (ST), Liste d'emplois (JB), Liste des endroits habités (WL), Liste des cours de langue seconde de vague 1 (LC) et Liste des cours de langue seconde de vague 3 (LD). Par exemple : *Pourquoi avez-vous cessé de suivre ce cours ou programme?*

Les variables cumulatives
Elles couvrent deux ou trois vagues. Elles sont plutôt rares dans l'ELIC. Par exemple, cette question demandée à la vague 2 : *Depuis votre arrivée au Canada, à quelle fréquence avez-vous été victime de discrimination ou d'un tel traitement injuste?*

6.5 Comment s'y retrouver : la table de concordance

La table de concordance indique non seulement s'il y a eu un changement, mais aussi la nature du ou des changements. Chaque ligne représente un item spécifique. Il devient ainsi facile de voir quels items ont été abandonnés et quels items se sont ajoutés. Chaque item possède une courte description qui lui est propre. Ces descriptions sont identiques aux étiquettes associées à chacune des variables contenues dans les cartes de syntaxe (formats pré-définis).

La colonne « Note » présente un ou plusieurs codes révélant la nature des modifications, ou des commentaires concernant la variable. L'utilisateur pourra consulter le livre de codes afin d'obtenir plus de détails sur les différences entre les vagues. Les codes suivants ont été attribués aux types de changements décrits à la section 6.4.

Code	Modification
Wm	la formulation ou le concept
In	les directives aux intervieweurs
Rc	catégories de réponse
Un	univers
St	structure du questionnaire
Pt	type de période de référence

Il est à noter que la table de concordance indique les changements de mesure d'un item d'une vague à une autre aussi bien que les changements qui ont donné lieu à la création d'un nouvel item.

6.6 Index des changements importants

Voici une liste des révisions les plus importantes qu'ont subi les questionnaires de l'ELIC au cours du temps :

1) Le type de période de référence pour les variables sur le revenu
À la première vague, les répondants devaient rapporter les montants reçus de différentes sources depuis l'arrivée au Canada (variables de vague). À la deuxième et à la troisième vague, les répondants doivent rapporter les montants reçus pour les douze mois précédant la date d'entrevue (variables annuelles). Le fait de recueillir des revenus sur une base annuelle permettra des comparaisons avec de nombreuses autres sources de données puisque la majorité des enquêtes recueille le revenu annuel.

2) La structure de l'information en ce qui concerne la formation générale et la formation linguistique
Les changements en ce qui concerne la formation régulière et la formation linguistique sont pour le moins complexes. En voici un aperçu :

- À la première vague, le module sur l'éducation recueillait des données sur tous les types de cours suivis par le répondant. Afin d'obtenir des détails spécifiques en ce qui concerne les cours de langue, nous avons décidé de séparer les cours de langue des autres types de formation dès la deuxième vague. Le module éducation des vagues 2 et 3 ne traite donc uniquement que de la formation régulière, excluant la formation linguistique. Certaines questions ont été ajoutées au module Compétences linguistiques (LS) afin d'obtenir des détails en ce qui concerne les cours de langue.

- À la vague 2, les cours de formation linguistique ne sont plus recueillis. La collecte des cours de formation linguistique a repris à la vague 3 grâce à l'ajout d'un module (Renseignements détaillés sur la formation linguistique). Au final, seulement les cours de langue ayant eu lieu dans les six premiers mois (vague 1) et ceux ayant eu lieu de la deuxième à la quatrième année après l'arrivée au Canada (vague 3) sont recueillis. Entre les deux, c'est le vide.

- Pour clarifier cette situation, nous avons créé dans les fichiers de vague 3 deux entités distinctes pour les cours de formation linguistique : Liste des cours de langue seconde de vague 1 (LC) et Liste des cours de langue seconde de vague 3 (LD). On retrouvera des variables dérivées en rapport avec la formation linguistique de vague 1 dans l'entité ED (Éducation) et celles en rapport avec la formation linguistique de vague 3 dans l'entité LS (Compétences linguistiques). Il s'agit d'une différence importante par rapport à ce qui

était fait avant. En effet, les utilisateurs les fichiers de données précédents de l'ELIC se souviendront que dans les fichiers des vagues 1 et 2, l'entité ST (Liste des études) contient tous les cours, y compris les cours de langue qui ont été recueillis à la vague 1.

- Bien que les cours de langue ne soient pas recueillis à la vague 2, l'entité LS (Compétences linguistiques) contient quelques variables pertinentes à ce sujet. Il est en effet possible de déterminer si le répondant a suivi des cours de langue pour apprendre ou perfectionner l'anglais ou le français et de connaître certains détails en ce qui concerne ces cours, sans qu'il soit toutefois possible de connaître le nombre de cours suivis et les dates de début et de fin des cours (et la durée par le fait même).

Pour plus de détails, il importe de consulter les annexes 2 et 3 de la Table de concordance.

3) La sélection aléatoire d'un enfant dans le ménage

Trois modules -- Éducation, Santé et Valeurs et attitudes -- comportent des questions au sujet des enfants des répondants longitudinaux (RL). Par exemple, certaines questions du module Santé demandent si les enfants du RL ont eu des problèmes dentaires depuis leur arrivée au Canada (voir questions HL_Q08B et HL_Q29 pour les vagues 1 et 2 respectivement). À la vague 1, on posait des questions au répondant au sujet de ses enfants en général. Aux vagues 2 et 3, afin d'obtenir de l'information plus précise, un enfant était sélectionné au hasard parmi tous les enfants âgés de 2 à 18 ans. Bien que les questions portaient sur l'enfant sélectionné à la vague 2 et 3, et sur tous les enfants à la vague 1, l'unité d'analyse demeure le RL, c'est-à-dire que les données portant sur les enfants doivent être utilisées comme caractéristiques du RL, et ce, à toutes les vagues. Cependant, à cause du changement dans la méthodologie, il peut être impossible de faire de l'inférence à la population d'intérêt des vagues 2 et 3 basé sur certaines des questions posées au sujet de l'enfant. Par exemple, l'utilisation de la question HL_Q29 pour estimer le nombre d'immigrants qui étaient parents d'un enfant qui a eu des problèmes dentaires résultera en une sous estimation, puisqu'un répondant pour lequel l'enfant sélectionné n'a pas eu un tel problème peut avoir un autre enfant avec ce problème. L'utilisateur est donc averti de faire preuve de prudence dans l'interprétation des analyses qui seront faites à partir des questions sur les enfants.

4) L'ajout de questions filtres

Plusieurs questions de la première vague tentaient de forcer une réponse du répondant. Par exemple, on leur demandait : *« Quels problème(s) ou difficulté(s) avez-vous eu(s) à trouver un emploi au Canada? »* sans leur avoir précédemment demandé s'ils avaient eu des problèmes à trouver un emploi.

À la deuxième et à la troisième vague, les questions de ce genre ont été reformulées. Pour chacune d'elles, une question filtre a été ajoutée. Suivant l'exemple précédent, la question : *« Depuis votre dernière entrevue, avez-vous eu des problèmes ou difficultés à trouver un emploi? »* a été demandée. Seulement les personnes ayant répondu avoir eu des problèmes se sont ensuite vus demander lesquels.

5) Le type de période de référence pour certaines variables du module Santé

Le type de période de référence change à la vague 3 pour un grand nombre de variables de l'entité Santé (HL). Alors qu'aux vagues 1 et 2 les questions couvraient toute la vague en terme de temps (variables de vague), à la vague 3 la plupart des questions couvrent les 12 derniers mois (variables annuelles). Ce changement devrait permettre une plus grande comparabilité avec des données provenant d'autres enquêtes.

7.0 Sélection de l'échantillon

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) vise à recueillir des données longitudinales sur les immigrants afin de mieux comprendre le processus d'adaptation des nouveaux immigrants à la société canadienne. Elle fournira des renseignements sur les facteurs qui favorisent ou entravent leur adaptation ainsi que sur les façons dont ils peuvent contribuer à la société ou à l'économie canadienne.

L'enquête consiste en trois interviews (vagues) : la première (vague 1) a eu lieu six mois après l'arrivée des immigrants au Canada; la deuxième (vague 2), deux ans après leur arrivée; et, la troisième (vague 3) a lieu quatre ans après leur arrivée.

Afin de produire des estimations fiables, on a sélectionné un échantillon représentatif d'environ 20 300 nouveaux immigrants. Ce chapitre décrit la méthode de sélection de l'échantillon de l'ELIC.

7.1 Populations de l'enquête

La ***population cible*** de l'enquête consiste en l'ensemble des immigrants qui répondent aux trois critères suivants :

- immigrants arrivés au Canada entre le 1er octobre 2000 et le 30 septembre 2001;
- immigrants âgés de 15 ans ou plus au moment de leur arrivée au Canada;
- immigrants reçus de l'extérieur du Canada qui ont présenté une demande par l'entremise d'une mission canadienne à l'étranger.

Sont exclues de l'enquête les personnes qui ont présenté une demande d'établissement en sol canadien. Il se peut que ces personnes aient été au Canada depuis longtemps au moment où elles ont officiellement obtenu le droit d'établissement et que, de ce fait, elles présentent, du point de vue de l'intégration, des caractéristiques très différentes de celles qui sont nouvellement arrivées au pays. Sont également exclus du champ de l'enquête les réfugiés faisant leur demande d'asile ici au Canada.

La population cible contient environ 169 400[1] des 250 000 personnes admises au Canada durant cette période. Le champ de l'enquête s'étend à toutes les régions métropolitaines de recensement et à toutes les agglomérations de recensement non éloignées.

La ***population d'intérêt*** comprend les immigrants de la population cible qui habitent toujours au Canada au moment d'un cycle donné. Pendant la période de six mois comprise entre l'arrivée des immigrants et la première interview et pendant les périodes entre les interviews subséquentes, certains immigrants ont quitté le Canada pour rentrer dans leur pays d'origine ou se rendre dans un autre pays. Ces personnes sont exclues de la population d'intérêt. Au premier cycle, cette population d'intérêt était évaluée à environ 164 200 immigrants et, au deuxième cycle, à 160 800 immigrants, et au troisième cycle, à 157 600 immigrants.

7.2 Base de sondage

La population cible est représentée par la base de sondage à partir de laquelle l'échantillon est sélectionné. La base de sondage de l'ELIC est la base de données administratives de Citoyenneté et Immigration Canada sur tous les immigrants ayant obtenu le droit d'établissement. Désignée sous le nom de SSOBL (Système de soutien des opérations des bureaux locaux), cette base de données fournit des renseignements sur les diverses caractéristiques de chaque

[1] Taille de la population cible selon une mise à jour de la base de sondage; au moment de la sélection de l'échantillon, on avait identifié environ 165 000 membres de la population cible (voir Tableau 7.1).

immigrant qui peuvent être utilisés dans la conception de l'enquête : nom, âge, sexe, langue maternelle, pays d'origine, connaissance de l'anglais ou du français, catégorie d'immigrants, date d'arrivée et province de destination prévue.

Deux mois après le mois de référence, Statistique Canada recevait du SSOBL des données détaillées sur chaque immigrant ayant obtenu le droit d'établissement au cours de la période de référence de l'enquête (c.-à-d., d'octobre 2000 à septembre 2001). On a pu ainsi construire la base de sondage mois après mois en y ajoutant simplement les nouveaux arrivants.

7.3 Conception de l'enquête

L'enquête fait appel à l'échantillonnage probabiliste. Le plan d'échantillonnage est un plan stratifié à deux degrés. En premier lieu, on a effectué la sélection des unités immigrantes (UI) à l'aide de la méthode d'échantillonnage avec probabilité proportionnelle à la taille (PPT). En deuxième lieu, on a sélectionné un membre au sein de chaque UI. On désigne sous le nom de répondant longitudinal (RL) le membre ainsi choisi, avec lequel on communique en vue de sa participation à l'enquête. Seul le RL fait l'objet d'un suivi tout au cours de l'enquête et aucune interview n'est menée auprès d'autres membres de l'UI ou du ménage du RL.

7.3.1 Échantillon longitudinal

L'ELIC est une enquête longitudinale, les immigrants étant interviewés à trois moments différents, soit six mois, deux ans et quatre ans après leur arrivée au Canada. Le plan d'échantillonnage a été conçu en se fondant sur l'approche dite entonnoir ou monotone; par conséquent, seuls les répondants de la première vague furent retracés pour la deuxième vague et seuls les répondants de la deuxième vague furent retracés pour la troisième vague.

On a opté pour l'approche entonnoir en raison de la nature de l'enquête et de ses objectifs analytiques. Les données recueillies portent sur les perceptions, les valeurs et les attitudes de l'immigrant à des moments particuliers et visent à évaluer son intégration durant ses premières années au Canada. Si les données n'étaient recueillies qu'une fois (c.-à-d., durant la quatrième année suivant l'arrivée au Canada), il pourrait en résulter des erreurs de mémoire et de réponse importantes. En outre, afin de faciliter une étude complète de l'adaptation de l'immigrant, il faut obtenir auprès de chaque répondant longitudinal toute la gamme des données longitudinales.

7.3.2 Stratification

La première variable de stratification utilisée fut le mois de l'arrivée au Canada. Il y avait 12 cohortes d'immigrants correspondant à chacun des mois de référence. À l'intérieur de chaque mois, deux autres variables de stratification furent utilisées : la province de destination prévue et la catégorie d'immigrants.

Les provinces furent regroupées en cinq blocs : Québec, Ontario, Alberta, Colombie-Britannique et le reste des provinces (les territoires furent exclus).

Les immigrants furent stratifiés en six catégories d'immigrants : les immigrants de la catégorie famille, les travailleurs qualifiés ou indépendants, les entrepreneurs et investisseurs, les réfugiés parrainés par le gouvernement, les autres réfugiés et les autres immigrants. Une strate étant représentée par l'intersection des niveaux précédents, il y avait 30 strates pour chaque cohorte mensuelle d'immigrants, soit 360 strates au total.

7.4 Sélection et taille de l'échantillon

L'échantillon fut constitué de deux composantes : l'échantillon de base et les échantillons supplémentaires. L'échantillon de base représente la population cible tandis que les échantillons supplémentaires sont constitués de sous-populations particulières, qui sont déterminées en analysant la répartition prévue de l'échantillon à la troisième vague ainsi que par diverses exigences des ministères fédéraux et provinciaux. Les sous-groupes suivants ont été suréchantillonnés :

1) les réfugiés parrainés par le gouvernement;
2) les réfugiés autres que ceux parrainés par le gouvernement;
3) les immigrants entrepreneurs et investisseurs;
4) les immigrants de la catégorie famille en Colombie-Britannique;
5) l'ensemble des immigrants en Alberta; et
6) les immigrants indépendants au Québec (travailleurs qualifiés et entrepreneurs et investisseurs).

La stratification a permis de contrôler la taille des échantillons supplémentaires sélectionnés qui représentent les divers sous-groupes.

Les tableaux 7.1, 7.2 et 7.3 indiquent la répartition de la population sur la base de sondage ainsi que la répartition prévue de l'échantillon de base et des échantillons supplémentaires pour la troisième vague.

En ce qui concerne l'échantillon de base, on a déterminé que 5 000 interviews complétées pour la troisième vague produiraient des estimations fiables[2] à l'échelle nationale, pour les provinces où l'afflux d'immigrants est particulièrement important (Québec, Ontario et Colombie-Britannique) et pour certaines catégories d'immigrants (catégorie famille et catégorie économique). En outre, il serait possible d'obtenir des estimations fiables pour d'autres combinaisons de variables, dans la mesure où l'on procéderait au nombre minimum d'interviews nécessaire. Compte tenu des besoins relatifs aux échantillons supplémentaires décrits ci-dessus, on prévoyait que le nombre minimum d'interviews à réaliser durant la troisième vague sera de 5 755.

La taille de l'échantillon de la première vague a été déterminée en fonction de plusieurs hypothèses relatives à l'érosion de l'échantillon à partir de la taille minimale durant la troisième vague. Sur la foi des résultats de diverses études longitudinales sur la population canadienne, on a estimé à 75 % le taux de réponse combiné (cas résolus et répondants) des deuxième et troisième vagues – 75 % des répondants de la première vague répondraient au questionnaire durant la deuxième vague et 75 % des répondants de la deuxième vague le feraient durant la troisième vague. En outre, on a eu recours à diverses sources pour estimer le taux de retour combiné, c.-à-d., après dépistage et classement des cas selon qu'ils sont dans le champ d'observation ou hors de celui-ci. Les résultats de l'étude pilote et de l'étude de couverture sur la langue[3] ont été utilisés comme sources de renseignements. Enfin, on s'est fondé sur les données du projet de Contre-vérification des dossiers (CVD)[4] de Statistique Canada pour estimer les taux prévus de dépistage et de cas résolus.

---

[2] « Estimations fiables » signifie que nous devons être capables d'estimer une proportion minimale de 10 % avec un coefficient de variation de 16,5 %. Pour satisfaire à cette exigence, les cellules doivent être constituées de 450 unités répondantes.

[3] Vu les contraintes opérationnelles, notamment le besoin de traduire le questionnaire en plusieurs langues et les coûts y afférents, on a réalisé une étude pour déterminer la couverture démographique selon la langue. On a établi que des traductions en 13 langues autres que l'anglais ou le français permettraient d'obtenir un taux national de couverture d'environ 93 % des immigrants ayant obtenu le droit d'établissement.

[4] Le projet CVD (1996) a été entrepris afin d'estimer le sous-dénombrement au Recensement de 1996. Cette étude est fondée sur une base qui comprend les immigrants qui sont arrivés au Canada entre le Recensement de 1991 et celui de 1996.

L'échantillon initial fut sélectionné sur une période de 12 mois. En répartissant l'échantillon proportionnellement au nombre d'immigrants arrivés chaque mois ainsi qu'entre les strates dans un mois donné, on aurait réduit au minimum la variance totale. Cependant, pour des raisons d'ordre opérationnel (p. ex., maintien d'un nombre constant d'interviews durant chaque mois de collecte), on a procédé à une répartition égale entre les mois d'arrivée, en dépit de la variation saisonnière du flux d'immigrants. Le tableau 7.4 présente la taille de l'échantillon final de la première vague.

Tableau 7.1 Nombre total d'immigrants âgés de 15 ans et plus, selon la province et la catégorie d'immigrants, d'octobre 2000 à septembre 2001

Provinces	Famille	Travailleurs qualifiés, catégorie économique	Entrepreneurs, catégorie économique	Réfugiés — gouvernement	Autres réfugiés	Autres	Total
Québec	4 680	12 694	2 977	1 238	887	78	22 554
Ontario	26 579	64 346	3 591	2 054	2 123	216	98 909
Alberta	3 250	5 651	444	623	307	125	10 400
Colombie-Britannique	8 532	15 048	2 489	679	317	235	27 300
Autres provinces	1 199	2 074	494	948	427	707	5 849
Canada	**44 240**	**99 813**	**9 995**	**5 542**	**4 061**	**1 361**	**165 012**

Tableau 7.2 Répartition prévue des répondants de la troisième vague – Échantillon de base

Provinces	Famille	Travailleurs qualifiés, catégorie économique	Entrepreneurs, catégorie économique	Réfugiés — gouvernement	Autres réfugiés	Autres	Total
Québec	151	312	94	46	25	5	633
Ontario	810	1 870	125	46	72	12	2 935
Alberta	104	156	21	13	6	4	304
Colombie-Britannique	287	505	108	12	10	10	932
Autres provinces	41	74	19	25	12	25	196
Canada	**1 393**	**2 917**	**367**	**142**	**125**	**56**	**5 000**

Tableau 7.3 Répartition prévue des répondants de la troisième vague – Échantillon de base et échantillons supplémentaires

Provinces	Famille	Travailleurs qualifiés, catégorie économique	Entrepreneurs, catégorie économique	Réfugiés — gouvernement	Autres réfugiés	Autres	Total
Québec	151	346	125	146	28	5	801
Ontario	810	1 870	153	146	79	12	3 070
Alberta	154	231	36	47	9	6	483
Colombie-Britannique	450	505	132	38	11	10	1 146
Autres provinces	41	74	23	79	13	25	255
Canada	**1 606**	**3 026**	**469**	**456**	**140**	**58**	**5 755**

Tableau 7.4 Répartition de l'échantillon finale de la première vague

Provinces	Famille	Travailleurs qualifiés, catégorie économique	Entrepreneurs, catégorie économique	Réfugiés — gouvernement	Autres réfugiés	Autres	Total
Québec	463	1 230	437	377	111	12	2 630
Ontario	2 653	6 920	599	630	269	23	11 094
Alberta	531	928	93	234	59	22	1 867
Colombie-Britannique	1 560	1 634	423	210	40	26	3 893
Autres provinces	121	225	81	293	46	72	838
Canada	**5 328**	**10 937**	**1 633**	**1 744**	**525**	**155**	**20 322**

8.0 Collecte des données

8.1 Interviews assistées par ordinateur

Pour la collecte des données de l'Enquête longitudinale auprès des immigrants du Canada (ELIC), on fait beaucoup appel à la technologie d'interviews assistées par ordinateur (IAO). L'emploi de cette technologie permet une collecte de données de grande qualité pour des contenus complexes sur des populations particulières. Par exemple, le système facilite la collecte de données sur les liens entre tous les membres du ménage (c.-à-d., la grille des liens). Cette mine de renseignements permettra une analyse détaillée des structures familiales, concept important pour l'analyse des données. Ce genre de collecte serait très difficile à mettre sur pied dans le cas d'une interview papier et crayon.

Le système IAO se divise en deux grands sous-systèmes :

1) Gestion de cas

Le système de gestion de cas sert à contrôler l'affectation des cas et la transmission des données pour l'enquête. Dans le cadre de l'ELIC, un cas correspond à une personne échantillonnée. Le système enregistre automatiquement des données de gestion pour chaque contact (ou tentative de contact) avec les enquêtés et fournit des rapports de gestion de l'activité de collecte.

Le système de gestion des cas permet d'acheminer les applications questionnaires et le fichier échantillon du bureau central aux bureaux régionaux et de ces derniers aux ordinateurs portatifs des intervieweurs. Les données recueillies font le chemin inverse. Par souci de confidentialité, on chiffre toutes les données en vue de leur transmission et on les déchiffre seulement lorsqu'elles se trouvent en lieu sûr dans un ordinateur séparé sans accès de l'extérieur.

2) Composantes particulières de l'enquête

Localisation des répondants
La population cible de la troisième vague de l'ELIC se compose d'immigrants qui sont au Canada depuis quatre ans. Or, pour toutes sortes de raisons, les nouveaux immigrants représentent une population très mobile durant leurs premières années au Canada. C'est pourquoi il faut procéder à des opérations de dépistage des répondants.

Afin d'aider à repérer les répondants, on a conçu un questionnaire contact pour demander l'adresse de l'immigrant au Canada (si elle est connue) ainsi que l'adresse d'une personne-ressource au pays. Le formulaire renferme également une déclaration de consentement par laquelle le répondant autorise Statistique Canada à avoir accès, à des fins de dépistage seulement, aux renseignements détenus par d'autres organismes fédéraux et provinciaux, tel qu'un ministère provincial de la santé. Ce formulaire était annexé à la documentation que les missions canadiennes à l'étranger ont fourni aux immigrants au moment de la délivrance du visa.

On n'a accès aux renseignements supplémentaires de dépistage que si le répondant éventuel a donné son consentement. Ce consentement permet à Statistique Canada d'avoir accès aux renseignements de dépistage contenus dans les dossiers de tous les ministères provinciaux de la santé, sauf celui de la Nouvelle-Écosse. On a jugé que cette source de renseignements est celle qui renferme les données les plus à jour quant à l'adresse du répondant.

Contact avec les répondants longitudinaux
À chaque vague, on a établi le premier contact avec les répondants sélectionnés en se reportant à l'adresse et au numéro de téléphone indiqués par le Bureau central dans le fichier de l'échantillon. L'intervieweur a fait confirmer que le répondant vivait à cette adresse. Après avoir établi qu'il parlait à la bonne personne, l'intervieweur a pris d'autres précautions pour vérifier que c'était bien le cas. Plus précisément, l'intervieweur s'est enquis de la date de naissance et de la date d'arrivée au Canada.

Après avoir vérifié qu'il s'agissait bien de la bonne personne, l'intervieweur a validé ou corrigé les données de contact (adresse postale et adresse de résidence, numéro de téléphone). Ensuite, il a pris rendez-vous en vue de la poursuite de l'interview sur place.

Dans les cas où il était impossible de localiser le répondant, les cas étaient transférés à l'équipe de dépistage au bureau régional afin d'effectuer une recherche plus minutieuse.

Dépistage des répondants
À chaque vague, l'équipe de dépistage du bureau régional a fait un suivi auprès d'autres sources afin de localiser le répondant. Les répertoires téléphoniques électroniques ont été la seule source d'information publique aux fins du dépistage. Pour retracer les répondants sélectionnés, on a eu recours aux sources d'information suivantes :

- dossiers administratifs de Citoyenneté et Immigration Canada;
- questionnaires contacts de l'enquête;
- adresses inscrites sur les cartes santé provinciales (dans les cas où un accord a été conclu avec la province et où on a obtenu le consentement du répondant);
- répertoires téléphoniques électroniques (Québec, Ontario et Colombie-Britannique).

Personne la mieux renseignée
Dans l'ELIC, les interviews par procuration ne sont pas permises. La seule exception est dans le module Revenu où la personne la mieux renseignée (PMR) sur le revenu familial était désignée pour répondre aux questions.

8.2 Collecte

Période de collecte
L'ELIC est une enquête longitudinale, ce qui implique que les mêmes répondants sélectionnés sont interviewés à plusieurs moments différents. Dans l'ELIC, les répondants sont interviewés à trois reprises. La première entrevue a lieu six mois après l'arrivée du répondant au Canada, car il est souhaitable d'évaluer le plus tôt possible le degré d'intégration. La deuxième entrevue est tenue deux ans après son arrivée et la dernière, quatre ans après celle-ci.

Pour bien représenter les différentes tendances d'immigration au Canada sur une période d'un an, l'échantillon se compose de 12 cohortes, c'est-à-dire de 12 échantillons mensuels indépendants sélectionnés sur une période de 12 mois consécutifs.

En théorie, un immigrant arrivé en octobre 2000 serait interviewé en avril 2001, octobre 2002 et octobre 2004. En pratique toutefois, cela peut varier. D'une part, la collecte de la deuxième vague a commencé deux mois plus tard que prévu, soit en décembre 2002, et celle de la troisième vague a commencé un mois plus tard que prévu, soit en novembre 2004. D'autre part, cela pouvait prendre jusqu'à trois mois aux vagues 1 et 2 et jusqu'à deux mois à la vague 3 pour réaliser toutes les entrevues d'un échantillon mensuel.

Date d'arrivée : octobre 2000 à septembre 2001		
Vague	**Début de la collect**	**Fin de la collect**
1	Avril 2001	Mai 2002
2	Décembre 2002	Décembre 2003
3	Novembre 2004	Novembre 2005

Méthodes de collecte
À la première vague, la plupart des interviews (68 %) ont été faites sur place, tandis que les autres interviews (32 %) ont été effectuées au téléphone pour diverses raisons (lieu de l'interview, exigences linguistiques particulières, etc.). À la deuxième vague, un peu plus de la moitié des interviews ont été faites sur place. À la troisième vague, la proportion d'entrevue sur place s'est élevée à 63 %.

Les interviews ont été menées dans l'une des 15 langues les plus fréquemment parlées par la population cible : anglais, français, chinois (mandarin, cantonais), panjabi, farsi/dari (une langue), arabe, espagnol, russe, serbo-croate, ourdou, coréen, tamoul, tagalog et gujarati. Les 15 langues sélectionnées couvraient environ 93 % de la population des nouveaux immigrants au Canada.

Durée de l'interview
Les interviews de la Vague 1 duraient environ 90 minutes en moyenne. Quinze minutes étaient consacrées aux composantes Entrée et Sortie et le reste (75 minutes) à l'enquête proprement dite. Pour les vagues suivantes, les interviews duraient environ 65 minutes en moyenne.

9.0 Traitement des données

Le principal résultat de l'Enquête longitudinale auprès des immigrants du Canada (ELIC) est un fichier maître de données « épuré ». Nous présentons dans le présent chapitre un bref résumé des étapes du traitement des données reliées à la production de ce fichier.

9.1 Vérifications préliminaires faites par l'application

Vérifications informatiques
Tel que mentionné précédemment, tous les renseignements recueillis auprès des personnes échantillonnées ont été obtenus en personne, à l'aide d'une application d'interview sur place assistée par ordinateur (IPAO), ou au moyen d'une interview téléphonique si une interview en personne n'était pas possible. On a pu ainsi inclure diverses fonctions de vérification dans le questionnaire afin de recueillir des données de grande qualité. Voici quelques exemples illustrant le genre de vérifications faites dans le cadre du processus d'interviews assistées par ordinateur (IAO) exécutées pour l'ELIC.

Vérifications de cheminement de questions
Tous les cheminements de questions étaient intégrés dans le système IAO. Par exemple, pour les questions concernant un conjoint/partenaire ou un enfant, le système IAO se réfère automatiquement aux renseignements sur les liens entre tous les membres du ménage inscrits dans la composante entrée afin de déterminer si un partenaire/conjoint ou un enfant vit avec le répondant longitudinal (RL). Le cas échéant, le système IAO poursuit en posant des questions précises à leur sujet. Dans la négative, le système IAO saute automatiquement ces questions.

Vérifications de cohérence générale
On a prévu un certain nombre de vérifications de cohérence dans le système IAO, et les intervieweurs étaient en mesure de revenir à des questions déjà posées pour rectifier les incohérences. Les intervieweurs recevaient aussi des instructions à l'écran pour traiter ou corriger des problèmes de réponse incomplète ou erronée. Par exemple, si un répondant avait indiqué dans le module sur la langue que l'anglais était la langue la plus souvent parlée à la maison, il ne pouvait répondre qu'il ne parlait pas anglais à une des questions subséquentes. Le cas échéant, le système signalait l'erreur en fenêtre instantanée et demandait à l'intervieweur de modifier l'une ou l'autre des réponses.

Vérifications d'intervalles dans les champs numériques
On a également intégré des vérifications d'intervalles dans le système IAO pour les questions exigeant d'indiquer des valeurs numériques. Si les chiffres indiqués ne s'inscrivaient pas dans l'intervalle, une fenêtre apparaissait instantanément pour indiquer l'erreur et demander à l'intervieweur de corriger la réponse erronée. Par exemple, dans le sous-module des Détails sur l'emploi, le nombre maximum d'heures travaillées par semaine a été fixé à 168 heures (nombre d'heures dans une semaine). Une fenêtre instantanée apparaissait pour indiquer le dépassement de la limite lorsqu'un répondant affirmait qu'il travaillait plus de 168 heures par semaine.

9.2 Exigences minimales en matière de réponse

Une des premières étapes du traitement des données de l'ELIC a consisté à définir les exigences à l'égard du nombre minimum de réponses requises pour considérer un enregistrement valide.

Aucune information
Parfois, on ne recueillait pas de données pour une personne échantillonnée. Il se pouvait que l'intervieweur soit incapable de dépister un immigrant sélectionné ou de prendre contact avec l'intéressé tout au long de la période de collecte. Il se pouvait aussi que l'intéressé refuse de

participer à l'enquête, soit absent pendant toute la période de collecte ou ne puisse être interviewé en raison de barrières linguistiques (une personne qui ne parlait aucune des 15 langues utilisées pour l'enquête).

Si on ne recueillait pas de renseignements auprès d'un immigrant, ce dernier était retranché du fichier de l'ELIC. On augmentait les facteurs de pondération de l'échantillon des immigrants répondants en fonction de ces immigrants retranchés.

Réponse complète ou partielle
La plupart du temps, on a obtenu une réponse complète des répondants; c'est-à-dire que tous les modules ont été complétés. Dans d'autres cas, il était possible d'effectuer une partie de l'interview, mais celle-ci n'était pas complète pour diverses raisons. Certains répondants n'avaient qu'un temps bien limité à consacrer à l'interview; parfois aussi, l'intervieweur faisait une partie de l'interview avec le répondant et prenait rendez-vous pour la terminer, mais sans pouvoir reprendre contact avec l'intéressé. Enfin, certains répondants peuvent avoir refusé de répondre à un ou plusieurs modules concernant des sujets auxquels ils sont plus sensibles.

Critères de réponse partielle
Pour les cas où l'interview n'a pas permis d'obtenir une réponse complète, il était nécessaire de définir des critères pour juger si l'enregistrement était valide (réponse partielle). On a considéré l'enregistrement comme partiel lorsque l'interview a permis d'obtenir suffisamment d'information pour permettre d'appliquer des stratégies d'imputation dans le but de compléter les questions restantes.

À la première vague, on a considéré un enregistrement comme partiel lorsque certains modules étaient incomplets, à l'exception des deux premiers : Entrée et Antécédents. À la deuxième et à la troisième vague, les critères étaient légèrement différents. Pour qu'un enregistrement soit valide, le répondant devait au minimum avoir fourni des réponses au module Entrée. Les cas de réponse partielle ont été conservés dans l'échantillon de répondants.

Composantes manquantes et imputation massive
En ce qui a trait aux immigrants ayant fourni des réponses partielles, toutes les variables des composantes manquantes ont été considérées non déclarées ou ont été imputées, sauf pour trois modules – « Valeurs et attitudes », « Citoyenneté » et « Impression sur la vie au Canada ». Les questions posées dans ces modules visaient à s'enquérir des opinions et perceptions du RL, qui variaient beaucoup trop pour qu'on puisse établir une solide stratégie d'imputation massive. Pour plus de détails sur l'imputation, voir le chapitre 11.0.

Nombre total de répondants
Au total, 7 716 répondants longitudinaux ont été jugés suffisamment complets pour être conservés dans le fichier définitif de la vague 3.

Ces immigrants avaient résidé à 12 593 endroits avant de s'installer dans le lieu de résidence actuel (renseignements recueillis grâce au sous-module des Endroits habités). Ils avaient suivi au total 6 315 cours ou séances de formation (excluant les cours de langues) et 2 721 cours de langue pour la période de temps couverte par la vague 1 seulement (recueillis grâce au sous-module Éducation – renseignements détaillés). À la vague 3, ils avaient suivis 744 cours de langue (recueillis grâce au sous-module Formation linguistique – renseignements détaillés). Ils ont déclaré 8 560 attestations d'études de toutes sortes (dans le sous-module des Attestations d'études). Finalement, le nombre total de leurs emplois ou entreprises depuis leur arrivée au Canada s'élevait à 14 221 (recueillis dans le sous-module des Détails sur les emplois).

9.3 Codage

Trois différents genres de codage ont été faits selon qu'il s'agissait de questions ouvertes, de questions de type recensement ou de textes inscrits dans les champs « Autre – Précisez ». Vu le nombre de nouvelles catégories qui ont été ajoutées aux questions à l'étape du codage, celui-ci a été fait avant l'étape de la vérification préliminaire, afin de réduire au minimum les corrections à cette étape et à celle des vérifications de cheminement.

9.3.1 Codage à des questions ouvertes

Les intervieweurs ont enregistré sur le questionnaire quelques éléments d'information sous forme de réponses à des questions ouvertes. Ainsi, dans le module sur l'Emploi, ils ont posé aux RL qui avaient travaillé depuis leur arrivée au Canada un ensemble de questions ouvertes au sujet de chaque emploi occupé :

- Quel genre d'industrie, d'entreprise ou de service est-ce/était-ce?
- Quel genre de travail faites/faisiez-vous à cet emploi?
- Dans cet emploi, quelles sont/étaient vos fonctions les plus importantes?

Dans le questionnaire de la première vague, dans le module Impressions sur la vie au Canada, les deux dernières questions posées étaient des questions ouvertes :

- Quelle est la chose la plus utile qui a facilité votre installation au Canada?
- Quelle est la chose la plus utile qui aurait facilité votre installation au Canada?

Mode d'enregistrement des réponses
L'intervieweur inscrivait littéralement les réponses données par le répondant à ces questions. Au Bureau central, les énoncés écrits ont été convertis en codes (p. ex., d'industrie ou de profession) afin d'assurer la comparabilité des données.

Mode de codage
Les questions ouvertes ont été codées au moyen de plusieurs classifications types. Les questions portant sur les professions ont été codées à l'aide de la Classification type des professions de 1991 (CTP) tandis que celles portant sur les industries l'ont été à l'aide du Système de classification des industries de l'Amérique du Nord (SCIAN 1997).

À la première vague, les variables portant sur le principal domaine d'études dans le module Éducation et le sous-module Éducation – renseignements détaillés ont été codées selon l'ensemble de codes établi du principal domaine d'études (PDÉ). Aux vagues 2 et 3, ces mêmes variables ont été codées à l'aide de la Classification des programmes d'enseignement (CPE, Canada, 2000). C'est la classification qu'utilise actuellement Statistique Canada pour le domaine des études. Pour permettre la comparabilité des données des trois vagues, on a recodé les variables de la première vague à l'aide de cette même classification.

Des ensembles de codes ont été conçus expressément pour l'ELIC afin de coder les réponses à des questions comme les deux exemples tirés du module Impressions sur la vie au Canada cités plus haut.

9.3.2 Codage des variables de type recensement

Quelques-unes des questions de l'ELIC ont aussi été posées dans le cadre du Recensement de 2001. Ces questions portaient sur le pays de naissance, le pays de citoyenneté, la langue, la religion, le groupe ethnique et l'appartenance à une minorité visible.

Mode d'enregistrement
Pour la plupart de ces questions, une liste de choix a été incluse dans le questionnaire. Dans beaucoup de cas, l'intervieweur a choisi la catégorie « Autre – Précisez » et inscrit un énoncé.

Mode de codage
Au Bureau central, chacune de ces questions a été codée au moyen de l'ensemble de codes correspondant aux éléments du dictionnaire de données du Recensement de 2001. Les catégories résultant du codage sont donc tout à fait comparables avec les données du recensement.

9.3.3 Codage des réponses de la catégorie « Autre – Précisez »

Les réponses à plusieurs questions du questionnaire de l'ELIC s'inscrivaient dans une catégorie « Autre – Précisez », qui permettait à l'intervieweur d'inscrire un texte lorsque la réponse fournie ne figurait pas dans la liste des choix.

Mode de codage
À la suite d'un examen consciencieux, les éléments inscrits dans les champs « Autre – Précisez » ont été codés selon trois scénarios possibles:

- ils ont parfois reçu le code d'une catégorie existante (lorsque le concept était similaire);
- ils sont parfois restés dans la catégorie « autre »;
- on a parfois ajouté des catégories à celles qui existaient initialement dans le questionnaire, soit lorsque les réponses pour une catégorie représentaient environ 5% ou plus de l'ensemble des réponses.

9.4 Vérification au Bureau central

Vérifications préliminaires
Avant de procéder aux vérifications préliminaires, on a créé des bases de données pour la section principale du questionnaire, pour les renseignements recueillis sur le ménage du RL et pour chacun des sous-modules.

Une étape importante du processus de vérification préliminaire a consisté à décomposer les questions « Inscrivez toutes les réponses qui s'appliquent » et à transformer leurs valeurs en réponses Oui (1) ou Non (2). Les valeurs de non-réponse du système IAO ont aussi été recodées en fonction de codes standard de non-réponse pour les refus, ne sait pas et non déclaré.

Conversion des codes de non-réponse en codes standards

Ne sait pas
Au cours d'une interview assistée par ordinateur, il se peut que le répondant ne connaisse pas la réponse à une question particulière. Le système IAO comporte une touche de fonction sur laquelle l'intervieweur appuie dans une telle situation.

Dans les fichiers de l'ELIC, le code utilisé pour indiquer que le répondant ne connaissait pas la réponse à une question particulière est « 7 ». Dans le cas d'une variable à deux chiffres, le code est « 97 », pour une variable à trois chiffres « 997 », etc.

Refus
Le répondant peut choisir de refuser de répondre à une question particulière. Le système IAO comporte une touche de fonction sur laquelle l'intervieweur appuie pour indiquer un refus. L'information est enregistrée pour la question et transmise au Bureau central.

Dans les fichiers de l'ELIC, une question refusée comporte un code « 8 ». Dans le cas d'une variable à deux chiffres, le code est « 98 », pour une variable à trois chiffres « 998 », etc.

Non déclaré
Lors du traitement qui se fait au Bureau central, on code parfois comme « non déclaré » la réponse à une question. On indique par ce code que la question n'a pas été posée au répondant. On attribue de tels codes pour trois raisons principales :

1) Lors de l'interview assistée par ordinateur, l'intervieweur pouvait entrer un code « refus » ou « ne sait pas », comme nous l'avons expliqué plus haut. Le système IAO était souvent programmé en pareil cas pour sauter cette section particulière du questionnaire. En cas de refus, on supposait que les questions posées étaient délicates et qu'il était probable que le répondant ne veuille pas répondre à d'autres questions à ce sujet. Dans les cas « ne sait pas », on supposait que le répondant n'était pas suffisamment informé pour répondre à d'autres questions et on ignorait si les questions subséquentes s'appliquaient à lui. Lors du traitement des données de l'ELIC, on a décidé que toutes les questions subséquentes se verraient attribuer un code « non déclaré ».

2) Dans certains cas, des sections ou des modules entiers du questionnaire n'avaient pas été commencés ou avaient été commencés puis interrompus prématurément. Par exemple, il a pu se produire une interruption où le répondant a dit ne pas vouloir continuer. Si on avait obtenu suffisamment de renseignements pour considérer le module comme rempli, on attribuait un code d'enchaînement valide aux questions restantes. Si le répondant n'avait pas répondu à un module complet, on effectuait une imputation massive – sauf pour les modules Citoyenneté, Valeurs et attitudes et Impressions sur la vie au Canada où on a attribué aux questions non répondues le code « non déclaré ».

3) La troisième situation où on a eu recours au code « non déclaré » est lors des vérifications de cohérence. Si on décelait une erreur de cohérence entre des groupes de variables, on attribuait le code « non déclaré » à une ou plusieurs des variables en question.

Dans le cas des variables dérivées, si on avait attribué le code « non déclaré » à une ou plusieurs des variables sources, on attribuait également le code « non déclaré » à la variable dérivée.

On a attribué un code « 9 » aux cas « non déclaré ». Dans le cas d'une variable à deux chiffres, le code est « 99 », pour une variable à trois chiffres « 999 », etc.

Vérifications de cheminement et attribution de codes d'enchaînement valide
La dernière étape du processus de vérification préliminaire a consisté à traiter les cheminements de questions dans chacun des fichiers et à attribuer des codes standard « d'enchaînement valide » (6, 96 et 996).

Par exemple, on a attribué un code « d'enchaînement valide » à toutes les variables relatives au

« conjoint » dans tous les cas où un conjoint ou un partenaire d'union libre ne vivait pas au sein du ménage du RL.

9.5 Vérification de la cohérence

Vérification de la cohérence
Les vérifications de cohérence consistent à vérifier les liens entre deux variables ou plus. Un exemple de problème de cohérence qui doit être corrigé durant le traitement des données est lorsque le revenu personnel du RL est plus élevé que le revenu total pour toute la famille, dont il ne devrait pourtant constituer qu'une partie. On corrige le problème en utilisant le plus d'informations possible provenant d'autres variables. Si possible, on change la réponse incorrecte pour la valeur qui semble correcte. Mais dans les autres cas, on applique un statut « non déclaré » à la valeur incohérente. En conséquence, il n'y a plus d'incohérences entre le revenu personnel du RL et le revenu familial dans le fichier final.

Vérifications de liens
La vérification des liens est une autre forme de vérification de cohérence. Pour diverses raisons, les données sur les liens recueillies dans la composante Entrée sont parfois erronées. L'étape de la vérification des liens permet de produire un fichier épuré et d'assurer la cohérence des liens entre les membres du ménage.

Par exemple, certains répondants dont le conjoint avait des enfants ont déclaré ne pas avoir de lien de parenté avec eux. En fait, selon les définitions du recensement, ces personnes auraient dû se considérer comme des beaux-parents, concept qui n'est pas bien connu de certains nouveaux immigrants au Canada. De même, certains parents de famille d'accueil déclarent ne pas avoir de lien de parenté avec l'enfant en foyer nourricier, alors qu'ils devraient indiquer qu'ils sont des parents de famille d'accueil.

9.6 Variables dérivées

Utilité des variables dérivées
Les variables dérivées facilitent le travail des analystes en fournissant de l'information condensée dont l'extraction demande un certain travail de programmation. Par exemple, une variable dérivée peut être le résultat de la combinaison de réponses provenant de plusieurs questions. Les variables qui fournissent le compte des évènements (comme les emplois par exemple) que l'on retrouve dans les entités-listes pour un enregistrement constituent d'autres exemples de variables dérivées.

Par ailleurs, certaines variables dérivées ont été créées afin de permettre la comparabilité des données d'une vague à l'autre. En effet, dans certains cas, on pouvait combiner les réponses de plusieurs questions pour créer une variable comparable (mesurant un même item) à une autre existant à la vague précédente.

Noms des variables dérivées
Toutes les variables dérivées dans les fichiers de données de l'ELIC reçoivent un « d », un « g » ou un « l » en quatrième position de leur nom.

Certaines variables dérivées dans le fichier original de la première vague ont dû être renommées dans les fichiers de données de vague 2 et vague 3. À la première vague, la numérotation de ces variables commençait à 001. Cela causait le problème que deux variables pouvaient avoir le même identificateur d'item, par exemple : HS1Q001 et HS1D001. À la vague 2, la variable HS1D001 a été renommée HS1D117 pour éviter les confusions possibles.

10.0 Non-réponse

Les taux de réponse à une enquête constituent une mesure de l'efficacité de l'échantillonnage de la population et du processus de collecte et sont également un bon indicateur de la qualité des estimations produites. Comme dans les autres enquêtes, l'Enquête longitudinale auprès des immigrants du Canada (ELIC) présente un certain niveau de non-réponse. Le présent chapitre comprend des précisions sommaires qui permettent de faire la distinction entre les deux types de non-réponse, soit la non-réponse totale et la non-réponse partielle.

Non-réponse totale :
Aucune donnée n'a été recueillie sur l'unité échantillonnée. C'est le cas de l'information incomplète décrite à la section 9.2. Dans le cas d'une non-réponse totale, on a eu recours à des méthodes d'ajustement de la pondération pour compenser. Cette question est examinée plus en détail au chapitre 12.0.

Non-réponse partielle :
Au moins un des modules, mais pas la totalité, était complet. Les critères définissant un module complet sont énoncées à la section 9.2. On a corrigé la non-réponse partielle par imputation.

10.1 Définition du statut de réponse

Les définitions suivantes sont requises à la compréhension du contenu des tableaux qui suivent.

À la vague 3, un **immigrant hors du champ de l'enquête** est un immigrant inclus dans le fichier de l'échantillon de la vague 3 mais qui, après certaines vérifications, ne satisfaisait pas aux critères définissant la population d'intérêt. Les immigrants décédés, ceux qui vivaient en établissement et ceux qui avaient quitté le Canada sont autant d'exemples d'immigrants hors du champ de l'enquête.

À la vague 3, un **immigrant répondant** est le répondant longitudinal (RL) sélectionné qui avait participé à la vague 1 et à la vague 2 et qui est soit un répondant partiel, soit un répondant complet (voir la section 9.2). À l'issue de la vague 3, on a identifié 7 716 enregistrements utilisables en tant qu'unités déclarantes.

Les cas dits **non résolus** ou **non dépistés** sont ceux identifiés à l'étape de la collecte de la vague 3 pour lesquels il n'y a eu aucun contact avec l'immigrant sélectionné. Aucun renseignement n'a été recueilli permettant de le repérer.

On entend par **non-répondants** les cas identifiés à l'étape de la collecte de la vague 3 pour lesquels on a réussi à repérer l'immigrant sélectionné et à confirmer sa présence au Canada mais, pour une raison donnée, il n'a pu répondre à l'interview.

Bien que les cas non-résolus et de non-réponse entraînent les uns comme les autres la production d'enregistrements inutilisables, la principale différence est que dans un cas de non-réponse, on a obtenu confirmation que l'immigrant sélectionné appartenait à la population d'intérêt de la vague 3.

Tableau 10.1 Résultats de la troisième vague de la collecte, selon le mois et l'année de référence

Mois et années	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
octobre 2000	596	81	15	55	747
novembre 2000	655	78	12	55	800
décembre 2000	631	59	14	54	758
janvier 2001	618	67	19	43	747
février 2001	677	68	16	64	825
mars 2001	661	66	5	51	783
avril 2001	639	70	5	57	771
mai 2001	680	71	13	59	823
juin 2001	673	71	13	41	798
juillet 2001	658	64	16	56	794
août 2001	649	57	15	40	761
septembre 2001	579	63	20	53	715
Total	**7 716**	**815**	**163**	**628**	**9 322**

Le mois de référence et l'année de référence sont des termes utilisés pour désigner le mois et l'année d'arrivée.

Tableau 10.2 Résultats de la troisième vague de la collecte, selon la catégorie d'immigrants

Catégories d'immigrants	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
Économique	4 509	434	119	350	5 412
Famille	1 993	288	30	186	2 497
Réfugiés	1 133	85	12	87	1 317
Autre	81	8	2	5	96
Total	**7 716**	**815**	**163**	**628**	**9 322**

Tableau 10.3 Résultats de la troisième vague de la collecte, selon le groupe d'âge

Groupes d'âge	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
15 à 24	1 347	162	30	144	1 683
25 à 34	2 883	293	61	273	3 510
35 à 44	2 150	170	41	123	2 484
45 à 64	1 141	156	27	77	1 401
65 et plus	195	34	4	11	244
Total	**7 716**	**815**	**163**	**628**	**9 322**

Tableau 10.4 Résultats de la troisième vague de la collecte, selon le sexe

Sexes	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
Hommes	3 819	400	77	318	4 614
Femmes	3 897	415	86	310	4 708
Total	**7 716**	**815**	**163**	**628**	**9 322**

Tableau 10.5 Résultats de la troisième vague de la collecte, selon la province de destination prévue

Provinces	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
Terre-Neuve-et-Labrador	19	1	0	3	23
Île-du-Prince-Édouard	6	2	1	0	9
Nouvelle-Écosse	43	3	2	1	49
Nouveau-Brunswick	36	6	1	2	45
Québec	1 139	93	26	83	1 341
Ontario	3 833	452	65	376	4 726
Manitoba	168	21	1	15	205
Saskatchewan	64	8	0	7	79
Alberta	925	86	13	34	1 058
Colombie-Britannique	1 483	143	54	107	1 787
Canada	**7 716**	**815**	**163**	**628**	**9 322**

Tableau 10.6 Résultats de la troisième vague de la collecte, selon le lieu de naissance

Lieux de naissance	Répondants	Non-répondants	Population hors du champ de l'enquête	Cas non résolus	Total
Afrique	786	74	17	66	943
Amérique	551	43	9	58	661
Asie	4 854	565	108	427	5 954
Europe	1 464	130	29	72	1 695
Océanie	61	3	0	5	69
Total	**7 716**	**815**	**163**	**628**	**9 322**

Tableau 10.7 Profile de réponse sur les trois vagues de l'ELIC

	Vague 1	Vague 2	Vague 3
Cas résolus	**14 571**	**10 892**	**8 694**
Répondants	12 040	9 322	7 716
Non-répondants	2 120	1 370	815
Hors du champ	411	200	163
Cas non résolus	**5 751**	**1 148**	**628**
Prédits dans la population d'intérêt	5 577	1 122	618
Prédits hors du champ	174	26	10
Total	**20 322**	**12 040**	**9 322**

La *Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie* (http://www.statcan.ca/francais/about/policy/infousers_f.htm) de Statistique Canada exige que les informations sur la non-réponse soient fournies afin d'informer les utilisateurs de la qualité des données. Le tableau 10.8 présente les taux pour la collecte et les taux (particuliers à chaque vague et longitudinaux) pour l'estimation. Ces taux sont définis ci-dessous.

Taux de réponse pour la collecte des données

$$\frac{\textit{Unités répondantes}}{\textit{Unités dans le champ de l'enquête} + \textit{Unités non résolues}}$$

Taux de réponse pour l'estimation

$$\frac{\textit{Unités répondantes}}{\textit{Unités dans le champ de l'enquête} + \textit{Nombre estimé d'unités dans le champ de l'enquête}}$$

Taux de réponses longitudinal (pour l'estimation)

$$\frac{\textit{Unités répondantes}}{\textit{Taille de l'échantillon initial} - \textit{Unités hors du champ de l'enquête} - \textit{Nombre estimé d'unités hors du champ de l'enquête}}$$

(Unités hors du champ de l'enquête et Nombre estimé d'unités hors du champ de l'enquête réfèrent à la vague courante et vague(s) précédente(s).

Tableau 10.8 Taux de réponse sur les trois vagues de l'ELIC

	Vague 1	Vague 2	Vague 3
Taux de collecte (%)	60,5	78,7	84,2
Taux pour l'estimation (%)	61,0	78,9	84,3
Taux longitudinal (pour l'estimation) (%)	61,0	47,8	39,9

11.0 Imputation

Essentiellement, l'imputation est le processus qui consiste à remplacer les valeurs manquantes ou incohérentes par des valeurs plausibles. Il s'agit de construire des valeurs qui produiront des estimateurs approximativement sans biais. Il existe de nombreuses méthodes bien connues auxquelles on peut avoir recours pour imputer des valeurs pour une variable ou un enregistrement donné. Effectuée correctement, l'imputation améliore la qualité des données en réduisant le biais dû à la non-réponse. Dans le cas de l'Enquête longitudinale auprès des immigrants du Canada (ELIC), l'imputation visait à produire un ensemble complet de données pour des variables ou des enregistrements et à réduire au minimum le nombre de champs « non déclaré » dans le fichier de microdonnées.

Les deux sections suivantes incluent, respectivement, une description de l'imputation par la méthode du plus proche voisin, utilisée pour les modules incomplets, et une description des techniques utilisées pour l'imputation d'items dans le module Revenu.

11.1 Imputation massive

11.1.1 Imputation longitudinale

L'imputation massive à la vague 3 était d'ordre longitudinal en ce sens qu'elle s'est faite simultanément pour les données des trois vagues.

La première étape a consisté à identifier les modules qui devraient faire l'objet d'une imputation longitudinale. À cette fin, on a produit des codes d'achèvement longitudinaux. Comme il a été expliqué à la section 9.2, on a défini des champs clés pour la vague 3 suivant les mêmes principes qui s'étaient appliqués aux vagues 1 et 2. On a défini un code d'achèvement longitudinal en se fondant sur les codes d'achèvement des trois vagues. On a considéré qu'un répondant longitudinal (RL) de la vague 3 était un répondant longitudinal complet uniquement s'il avait été un répondant complet aux trois vagues. Dans le cas contraire, on a considéré que le RL était un répondant partiel longitudinal. Cette règle a eu pour conséquence qu'un module a été classifié comme longitudinalement incomplet s'il avait été incomplet à l'une des trois vagues. Par conséquent, dans les cas où un module avait été complet pour un ou deux vagues mais pas à l'autre, les données valides associées à ce module particulier ont été écrasées. Heureusement, le nombre de RL pour lesquels cela a présenté un problème était peu élevé (641 sur 7 716).

Le tableau 11.1 présente les différents schémas d'achèvement de modules longitudinaux, pour tous les enregistrements des répondants. Dans ce tableau, un « 1 » signifie que le module est complet, c'est-à-dire que tous les *champs clés* du module (vague 1, vague 2 et vague 3) contiennent des valeurs valides, tandis qu'un « 2 » signifie que le module est incomplet (l'information était incomplète pour une, deux ou trois vagues).

Tableau 11.1 Répartition des modules longitudinaux remplis

EN	IS	LS	HS	ED	EM	HL	IN	Nombre de cas	Pourcentage
1	2	2	2	2	2	2	2	2	0,03 %
2	1	1	1	2	2	2	2	1	0,01 %
2	1	1	1	1	2	2	2	3	0,04 %
2	1	1	1	1	1	2	2	2	0,03 %
2	1	1	1	1	1	1	2	3	0,04 %
2	1	1	1	1	1	1	1	14	0,18 %
1	2	2	2	2	2	2	2	36	0,47 %
1	2	2	1	1	1	1	1	1	0,01 %
1	2	1	2	2	2	2	2	2	0,03 %
1	2	1	2	1	1	2	2	1	0,01 %
1	2	1	1	1	1	2	2	2	0,03 %
1	2	1	1	1	1	2	1	1	0,01 %
1	2	1	1	1	1	1	2	8	0,10 %
1	2	1	1	1	1	1	1	41	0,53 %
1	1	2	2	2	2	2	2	2	0,03 %
1	1	2	1	2	1	1	1	1	0,01 %
1	1	2	1	1	2	2	2	1	0,01 %
1	1	2	1	1	1	1	1	14	0,18 %
1	1	1	2	2	2	2	2	9	0,12 %
1	1	1	2	2	1	2	2	1	0,01 %
1	1	1	2	2	1	1	2	1	0,01 %
1	1	1	2	1	1	1	2	6	0,08 %
1	1	1	2	1	1	1	1	4	0,05 %
1	1	1	1	2	2	2	2	9	0,12 %
1	1	1	1	2	1	1	2	5	0,06 %
1	1	1	1	2	1	1	1	51	0,66 %
1	1	1	1	1	2	2	2	42	0,54 %
1	1	1	1	1	2	1	2	3	0,04 %
1	1	1	1	1	1	2	2	42	0.54 %
1	1	1	1	1	1	2	1	33	0.43 %
1	1	1	1	1	1	1	2	314	4.07 %
1	1	1	1	1	1	1	1	7 061	91,51 %

EN : Entrée; SI : Interactions sociales; LS : Compétences linguistiques; HS : Logement; ED : Éducation; EM : Emploi; HL : Santé; IN : Revenu.

Selon le tableau 11.1, le module sur le revenu a le taux de non-réponse le plus élevé, soit de 6,4 %. Dans le cas du module sur le revenu, on a utilisé une autre méthode de traitement. Cette méthode est décrite à la section 11.2.

11.1.2 Stratégie d'imputation longitudinale

Pour la réponse partielle longitudinale à la vague 3, on a procédé à une imputation massive pour les modules incomplets en appliquant la technique du donneur par le plus proche voisin. La méthode de l'imputation par donneur n'altère généralement pas la distribution des données, ce qui est un inconvénient associé à beaucoup d'autres techniques d'imputation. Elle visait à remplacer l'information manquante pour un répondant partiel longitudinal avec les valeurs fournies par un répondant complet longitudinal qui était « semblable » au premier répondant. Cela fonctionnait de la façon suivante : en se fondant sur une distance statistique calculée au moyen de certaines informations sociodémographiques, on identifiait un donneur (répondant complet longitudinal) considéré comme étant le plus près du receveur (répondant partiel longitudinal) et on a utilisé les valeurs associées au donneur pour remplacer les valeurs manquantes aux deux vagues pour le receveur. Cela s'est fait module par module. Il convient de souligner que les variables sociodémographiques utilisées dans la sélection des donneurs incluaient celles qui déterminaient l'enchaînement des questions, à savoir la présence du conjoint et des enfants du RL, et aussi la présence d'enfants d'âge scolaire du RL.

Dans le cas des répondants partiels longitudinaux pour lesquels plus d'un module était incomplet, on a utilisé le même enregistrement donneur pour tous les modules incomplets. Il convient de souligner que l'on a utilisé comme donneurs possibles uniquement les enregistrements complets vérifiés. Pour assurer l'uniformité des variables, on a imputé à l'enregistrement receveur l'ensemble complet des variables d'un module donné de l'enregistrement donneur. À la fin de ce processus, tous les modules de tous les enregistrements étaient complets. Une variable indicatrice signalant si le module était imputé a été créée.

11.1.3 Imputation relative à des événements

Un autre aspect de l'imputation massive à la vague 3 avait trait à la rectification de différentes variables de date tirées des listes d'événements (logement, éducation et formation, antécédents professionnels). Pour s'assurer d'avoir une cohérence au niveau des dates, on a simplement utilisé comme dates imputées les dates du donneur. Les dates d'interview avec le RL donneur dans les deux vagues, la date d'arrivée et le nombre de jours écoulés entre la date d'arrivée et les dates d'interview sont contenues sur les enregistrements receveurs et ont été utilisées pour la dérivation des variables connexes. Les données imputées donnent du répondant receveur le portrait qu'il aurait eu s'il était arrivé et avait été interviewé aux mêmes dates que ses répondants donneurs. Cette méthode est la même que celle utilisée à la vague 2.

Les enregistrements receveurs contiennent également des variables supplémentaires relatives à la destination prévue du donneur à l'arrivée de même que des renseignements géographiques relatifs aux déménagements au Canada et de l'information sur le métier ou la profession que le donneur a exercé ou prévoyait exercer à son arrivée au Canada. Il est possible pour un RL ayant fait l'objet d'une imputation d'avoir un déménagement à l'extérieur de la province alors que dans les faits il n'y a eu qu'un déménagement local ou pas de déménagement, ou encore d'avoir un changement complet de profession avant et après l'imputation. Par conséquent, l'information géographique relative au donneur est utile pour déterminer les habitudes de résidence et de déménagement des receveurs au Canada à partir des données sur le donneur. De même, il est important de connaître la profession du donneur à l'arrivée pour déterminer si le receveur a conservé la même profession ou changé de profession avec le temps.

Il conviendra de se rappeler que pour les receveurs, les données imputées correspondent aux périodes, aux lieux et aux caractéristiques se rapportant aux donneurs. Il faut simplement s'abstenir de comparer les données imputées relatives aux receveurs avec leurs données réelles, particulièrement dans le cas des données de liste d'événements.

11.1.4 Exemples d'imputation massive

Pour mieux comprendre l'imputation massive, nous vous présentons ici deux exemples qui décrivent bien les étapes nécessaires à sa réalisation.

D'abord, comme il a été clairement décrit à la section 11.1.1, nous devons identifier les modules à imputer. Dans le tableau qui suit, chaque module a son code d'achèvement (1 : complet; 2 : partiel) pour chacune des vagues et la dernière ligne représente le résultat pour l'ensemble des trois vagues.

Tableau 11.2

Vague	EN	IS	LS	HS	ED	EM	HL	IN
1	1	1	1	1	2	1	1	1
2	1	1	1	1	1	1	1	2
3	1	1	1	1	1	1	1	2
Total	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**2**

La prochaine étape consiste à calculer un pointage pour chaque donneur potentiel basé sur les différentes caractéristiques du receveur (sexe, âge, conjoint, enfant, réponses aux questions clés, emploi, etc.). Si le donneur potentiel a la même caractéristique que le receveur, il obtient un certain nombre de points pour cette caractéristique et ainsi de suite pour toutes les caractéristiques sélectionnées. À la fin de ce processus, nous faisons la somme de tous les points et le donneur ayant obtenu le pointage le plus haut est choisi; en cas d'égalité nous choisissons le donneur aléatoirement parmi ceux qui ont obtenu le maximum de points.

Dans ce cas particulier, nous remarquons que le module « éducation » (ED) est incomplet pour la vague 1 et que le module « revenu » (IN) est incomplet pour les vagues 2 et 3. L'enregistrement receveur va donc conserver ses données pour tous les modules à l'exception des modules « éducation » et « revenu » qui seront imputés pour les trois vagues par le même donneur, celui ayant eu le plus haut pointage.

Comme second exemple, nous allons impliquer le module « emploi » qui contient des listes d'événements et des dates, ainsi que le module « logement » qui contient une liste d'événements et des adresses.

Tableau 11.3

Vague	EN	IS	LS	HS	ED	EM	HL	IN
1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	1	2
3	1	1	1	2	1	1	1	1
Total	**1**	**1**	**1**	**2**	**1**	**2**	**1**	**2**

Le module « logement », le module « emploi » et le module « revenu » sont donc les modules à imputer par un donneur. Il y a une liste d'événements sur le déménagement joint au module « logement » et deux listes d'événements (liste des emplois et détails sur l'emploi) joints au module « emploi ». Lorsqu'ils sont imputés, ces modules et listes d'événements sont remplis avec les données du donneur (dates et adresses aussi). Il est fort probable qu'il y ait des incohérences avec les dates et les adresses qui sont fournies dans le module « entré » (non imputé) et les modules imputés. Ce dernier point est discuté plus en détail dans la section 11.1.5.

Cet enregistrement sera imputé, pour les trois modules mentionnés, par le donneur ayant obtenu le maximum de points, avec les dates et adresses du donneur.

11.1.5 Mise en garde pour l'emploi de données imputées

Certaines analyses utiliseront des informations (souvent les dates ou la géographie) provenant de deux ou plusieurs entités (modules), où une entité est imputée et l'autre ne l'est pas. Par exemple, l'entité « ménage » (HH), qui contient des informations sur l'adresse du répondant au moment de l'entrevue et qui n'est jamais imputée, et l'entité « liste des endroits habités » (WL), qui contient les informations sur toutes adresses antérieures pourraient être utilisées afin d'examiner les transitions entre les résidences. Si l'entité WL est imputée et on s'intéresse à un certain aspect de la transition (disons un changement de régions métropolitaines de recensement (RMR) ou d'agglomération de recensement (AR)) entre l'adresse au moment de l'entrevue et une adresse antérieure, l'emploi des données imputées de WL avec la vraie adresse du répondant de HH pourrait créer une transition artificielle ou illogique. C'est-à-dire, en comparant l'item HH_037 (RMR/AR de l'adresse courante) avec item WL_030 (RMR/AR de l'adresse antérieure) la transition pourrait indiquer un changement d'une grande région métropolitaine de recensement (RMR) à une région rurale (non-RMR/AR). Par conséquent, si les données imputées sont utilisées dans l'analyse, la procédure appropriée serait de comparer le RMR/AR du donneur au moment de l'entrevue (HH_044) avec WL_030.

Pour cette raison, les dates et informations géographiques du donneur sont fournies pour certaines entités non-imputées.

11.1.6 Impact de l'imputation massive

Nous avons fait une étude pour déterminer si l'imputation massive pouvait avoir un impact sur les estimations. Comme le nombre d'enregistrements imputés par module est très faible et que nous avons utilisé l'imputation par le plus proche voisin qui tend à conserver les distributions, l'impact ne devrait pas être significatif. Les résultats obtenus suite à notre étude ont bien confirmé notre hypothèse et, nous pouvons donc vraiment conclure que l'impact de l'imputation massive sur les estimations est négligeable.

Tableau 11.4

Module	Nombre d'enregistrements imputés	Pourcentage (%)
Réseau Social	94	1,2
Compétences linguistiques	57	0,7
Logement	65	0,8
Éducation	121	1,6
Emploi	111	1,4
Santé	189	2,4
Revenu	495	6,4

11.2 Imputation par champ des variables sur le revenu

L'interview de l'immigrant dans le cadre de l'ELIC comporte plusieurs questions relatives au revenu. On recueille des renseignements sur le revenu familial du répondant longitudinal selon les sources à l'intérieur du Canada et à l'extérieur du Canada. On recueille aussi des renseignements sur le revenu personnel du répondant longitudinal selon toutes les sources (à l'intérieur et à l'extérieur du Canada) et sur ses montants d'épargnes et de prêts.

Le revenu est un sujet délicat. Certains répondants refusent de donner des réponses aux questions détaillées se rapportant aux diverses sources du revenu. Il arrive que certains répondants fournissent tout de même une estimation du revenu familial total ou une estimation de leur revenu personnel, parfois au moyen d'intervalles de revenu. De plus, parmi ceux qui répondent aux questions, il arrive que les montants indiqués aux sections touchant le revenu soient incompatibles avec les réponses données à la section relative à l'emploi (par exemple, un répondant qui, selon les réponses données à la section relative à l'emploi, a travaillé au cours des 12 derniers mois mais qui ne déclare pas de salaire ou de revenu net d'un emploi autonome dans la section sur le revenu). On procède donc à une imputation du revenu pour combler les valeurs manquantes attribuables à la non-réponse partielle (section 11.2.2) et, dans une moindre mesure, pour corriger les données incohérentes lorsque cela est possible (section 11.2.1).

11.2.1 Détection et imputation des valeurs aberrantes

Avant d'effectuer l'imputation par champ des valeurs manquantes, les variables quantitatives de revenu passent d'abord par un processus de détection de valeurs aberrantes. Ce processus permet, entre autres, de définir le bassin de donneurs qui sera utilisé pour l'imputation des valeurs manquantes. Pour chacune des variables quantitatives, la distribution empirique pondérée est produite et représentée graphiquement afin de comparer les données obtenues entre elles et d'identifier les valeurs extrêmes. À noter que les données de revenu sont généralement asymétriques, l'asymétrie étant caractérisée par un plus grand étalement vers les valeurs élevées de la variable et par le fait que certaines données peuvent prendre des valeurs négatives (par ex. : la perte de revenus dans le cas d'un travail autonome). Les valeurs identifiées comme extrêmes sont inspectées manuellement. Deux résultats sont possibles suite à l'inspection :

1) Il s'agit d'une valeur aberrante : dans ce cas, la médiane ou une valeur plus plausible que la médiane[1] est imputée;
2) Il s'agit d'une valeur extrême mais acceptable étant donné d'autres informations : dans ce cas, la valeur n'est pas changée mais est identifiée pour être exclue du bassin de donneurs pour l'imputation.

11.2.2 Imputation par champ des valeurs manquantes

Les valeurs manquantes du module sur le revenu sont ensuite imputées par la méthode du plus proche voisin. Cette méthode consiste à retracer un répondant ayant fourni une réponse à la section sur le revenu (un donneur) et dont les caractéristiques sont semblables à celles de la personne ou la famille n'ayant pas fourni de renseignement complet sur le revenu (un receveur). Une fois qu'on a identifié le voisin le plus proche, le montant déclaré par le donneur est imputée au receveur. Comme les règles pour trouver un donneur sont différentes selon la source à imputer, l'imputation est effectuée par champ (c.-à-d. indépendamment pour chacune des sources). En d'autres mots, dans un cas où il y aurait plus d'une source de revenus à imputer, il pourrait y avoir plus d'un donneur.

Le fichier de données diffusé pour la vague 3 est un fichier longitudinal, c.-à-d. qu'il contient les données des vagues 1, 2 et 3 des répondants de la vague 3. L'imputation massive a donc été effectuée longitudinalement afin d'assurer la cohérence entre les données de la vague 1, de la vague 2 et celles de la vague 3 (voir la section 11.1). Ceci a eu pour effet de changer certaines données des vagues 1 et 2, dont celles du module sur le revenu. Le processus d'imputation par champ des variables de revenu doit donc être effectué pour les données de la vague 1, pour celles de la vague 2 et pour celles de la vague 3. Les trois processus d'imputation sont effectués de façon indépendante. Pour l'imputation des variables des vagues 1 et 2, le bassin de donneurs est cependant réduit aux immigrants qui étaient aussi répondants à la vague 3. Bien que les répondants des vagues 1 et 2 qui n'ont pas répondu à la vague 3 pourraient techniquement servir de donneur pour les données des vagues 1 et 2, il se pourrait que ces individus aient des caractéristiques différentes des immigrants qui ont répondu aux trois vagues. Ils sont donc exclus du bassin de donneurs pour éviter d'introduire un biais potentiel dans les données.

Dans le cadre de l'ELIC, seuls les montants de revenu familial provenant de 11 sources à l'intérieur du Canada sont imputés, en plus du revenu personnel du répondant longitudinal. Parmi les variables qui représentent des sources de revenu à l'intérieur du Canada, six sont en lien avec le marché du travail et cinq sont des paiements de transfert, c'est-à-dire des revenus provenant d'un gouvernement au Canada. La liste des variables pour lesquelles on a procédé à l'imputation figure au tableau 11.5. Le tableau montre le taux d'imputation global pour chacune des variables, pour la vague 1, la vague 2 et la vague 3 respectivement. Il convient de signaler que, même si l'imputation améliore généralement la qualité des données dans l'ensemble, les données artificielles créées sont utilisées aux fins d'estimation et peuvent mener à une sous-estimation substantielle de la variance, surtout si le taux d'imputation est grand. Des indicateurs d'imputation sont intégrés au fichier de l'ELIC pour identifier les variables ayant fait l'objet d'une imputation dans un enregistrement. Les utilisateurs peuvent donc mesurer l'ampleur de l'imputation pour une variable en particulier. Pour tous les indicateurs d'imputation du fichier de données de l'ELIC, un « I » apparaît au quatrième caractère du nom de la variable. Ainsi, IN2I004 représente l'indicateur d'imputation pour le revenu familial de tous les emplois (IN2Q003).

1. Une valeur plus plausible que la médiane est imputée entre autres pour les cas où une erreur de saisie s'est produite – ex. : 200 000 a été entré au lieu de 20 000, alors 20 000 sera la valeur imputée.

Tableau 11.5 Taux d'imputation du revenu et des gains

Description de la variable	Vague	Nom de la variable	Nom de l'indicateur d'imputation de la variable	Nombre de cas excluant les enchaînements valides	Nombre de valeurs imputées	Taux d'imputation
Revenu de tous les emplois	Vague 1	IN1Q003	IN1I004	5096	583	11.44 %
	Vague 2	IN2D003x	IN2I004	6324	595	9.41 %
	Vague 3	IN3D003x	IN3I004	6612	538	8.14 %
Revenu d'un travail autonome	Vague 1	IN1Q005	IN1I006	285	87	30.53 %
	Vague 2	IN2D005x	IN2I006	949	219	23.08 %
	Vague 3	IN3D005x	IN3I006	1298	205	15.79 %
Pension d'une entreprise canadienne ou d'une société canadienne	Vague 1	IN1Q027	IN1I028	24	9	37,50%
	Vague 2	IN2D027x	IN2I028	29	1	3.45 %
	Vague 3	IN3D027x	IN3I028	31	1	3.23 %
Parrain privé	Vague 1	IN1Q030	IN1I031	31	4	12.90 %
	Vague 2	IN2D030x	IN2I031	52	5	9.62 %
	Vague 3	IN3D030x	IN3I031	67	6	8.96 %
Placements	Vague 1	IN1Q033	IN1I034	195	39	20.00 %
	Vague 2	IN2D033x	IN2I034	307	22	7.17 %
	Vague 3	IN3D033x	IN3I034	360	22	6.11 %
Autres sources	Vague 1	IN1Q036	IN1I037	369	18	4.88 %
	Vague 2	IN2D036x	IN2I037	374	9	2.41 %
	Vague 3	IN3D036x	IN3I037	376	13	3.46 %
Assistance sociale	Vague 1	IN1Q008	IN1I009	1043	27	2.59 %
	Vague 2	IN2D008x	IN2I009	1063	26	2.45 %
	Vague 3	IN3D008x	IN3I009	643	28	4.35 %
Assurance emploi	Vague 1	IN1Q011	IN1I012	143	26	18.18 %
	Vague 2	IN2D011x	IN2I012	1164	63	5.41 %
	Vague 3	IN3D011x	IN3I012	1082	48	4.44 %
Prestations fiscales ou crédits pour enfants	Vague 1	IN1Q014	IN1I015	2562	132	5,15%
	Vague 2	IN2D014x	IN2I015	3946	241	6.11%
	Vague 3	IN3D014x	IN3I015	3804	203	5.34 %
Régime de pension canadien ou Régime des rentes du Québec	Vague 1	IN1Q017	IN1I018	64	18	28.13 %
	Vague 2	IN2D017x	IN2I018	118	11	9.32%
	Vague 3	IN3D017x	IN3I018	85	9	10.59%
Autres sources du gouvernement	Vague 1	IN1Q023	IN1I024	510	25	4.90 %
	Vague 2	IN2D023x	IN2I024	837	22	2.63 %
	Vague 3	IN3D023x	IN3I024	1124	22	1.96 %
Revenu personnel du répondant longitudinal reçu de toutes les sources	Vague 1	IN1D067	IN1I068	7716	109	1.41 %
	Vague 2	IN2D067x	IN2I068	7716	242	3.14 %
	Vague 3	IN3D067x	IN3I068	7716	274	3.55 %

12.0 Traitement de la non-réponse totale et de la pondération

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) est une enquête probabiliste. Comme dans le cas de toute enquête probabiliste, l'échantillon est sélectionné de manière à représenter le plus fidèlement possible une population de référence, la population d'immigrants, à une date précise dans le cadre de cette enquête. Pour ce faire, chaque unité dans l'échantillon doit donc représenter un certain nombre d'unités dans la population. L'échantillon complet de la vague 3 est un sous-ensemble de l'échantillon de la vague 2, qui est lui-même un sous-ensemble de l'échantillon de la vague 1. Il se compose uniquement des immigrants répondants à la vague 1 et à la vague 2. Bien que certains liens entre les trois vagues sont faits dans le présent chapitre, la pondération de la vague 3 en est le sujet principal. Pour de plus amples détails sur la pondération de la vague 1 et de la vague 2, veuillez consulter respectivement le chapitre 10.0 du guide de l'utilisateur de la vague 1 et le chapitre 12.0 du guide l'utilisateur de la vague 2.

12.1 Représentativité des poids

Pour la plupart des enquêtes, la somme des poids finaux représente les chiffres estimatifs de la population cible, qui sont habituellement égaux à la population d'intérêt. Dans le cas de l'ELIC, toutefois, étant donné la mobilité de la population et les objectifs de l'enquête (voir le chapitre 3.0 du guide de l'utilisateur de la vague 1), la population d'intérêt représente en réalité une partie de la population cible, c'est-à-dire les immigrants qui demeuraient encore au pays au moment de l'interview. De plus, la population d'intérêt de la vague 3 diffère de la population d'intérêt des deux autres vagues puisqu'elle n'est qu'un sous-ensemble de ces derniers.

Rappelons tout d'abord que la base de sondage couvre la population cible, c.-à-d. les immigrants qui satisfont à tous les critères suivants :

- arrivés au Canada entre le 1er octobre 2000 et le 30 septembre 2001;
- âgés de 15 ans ou plus au moment de l'arrivée au pays;
- arrivés de l'étranger; doivent avoir présenté leur demande par l'entremise d'une mission canadienne à l'étranger.

Toutefois, certains de ces immigrants ont résidé au Canada pendant un certain temps avant de retourner dans leur pays d'origine ou d'émigrer vers un autre pays. Ces immigrants n'ont pas les mêmes caractéristiques d'adaptation que ceux qui sont des résidents permanents du Canada. Il est biaisé d'inclure dans le même ajustement de la pondération les immigrants qui ont quitté le Canada et ceux qui continuent d'y résider. La population cible inclut ces deux sous-groupes de base.

La **population d'intérêt** (PI) **de la vague 3** est constituée des immigrants de l'ELIC qui sont toujours au Canada quatre ans après leur arrivée (à titre comparatif, la population d'intérêt de la vague 1 était constituée des immigrants de l'ELIC qui étaient toujours au Canada six mois après leur arrivée et celle de la vague 2 était constituée des immigrants de l'ELIC qui étaient toujours au Canada deux après leur arrivée). Le poids final de la vague 3 donne des estimations de la population d'intérêt de la vague 3. La **population hors du champ d'intérêt** (OOI) est constituée d'immigrants qui n'habitent plus au Canada, c.-à-d. qui ont quitté le Canada après avoir été admis comme immigrants.

12.2 Aperçu des ajustements de poids

À l'étape de la collecte, les immigrants sélectionnés ont été classés dans l'une de quatre catégories, soit répondant, non-répondant, non inclus dans la population d'intérêt et cas non résolu. Dans le cas des immigrants classés dans les trois premières catégories, on a procédé à un premier contact avec l'immigrant ou avec une personne qui a pu confirmer le statut de celui-ci. Ces cas sont considérés comme résolus puisque le statut de l'immigrant est connu. La dernière

catégorie est celle des cas non résolus, pour lesquels on n'a pas établi de contact et qui sont donc restés non résolus. On ne possédait pas de renseignements permettant de déterminer si ces personnes se trouvaient toujours au Canada. Les ajustements de poids ont tenu compte de ces résultats.

On peut répartir l'échantillon d'abord entre les cas résolus et les cas non résolus :

Échantillon : $S = S_U + S_R$

où S_U = unités échantillonnées dont le cas n'est pas résolu

S_R = unités échantillonnées dont le cas est résolu

En outre, dans la partie concernant les unités résolues, $S_R = S_{RR} + S_{RN} + S_{RO}$

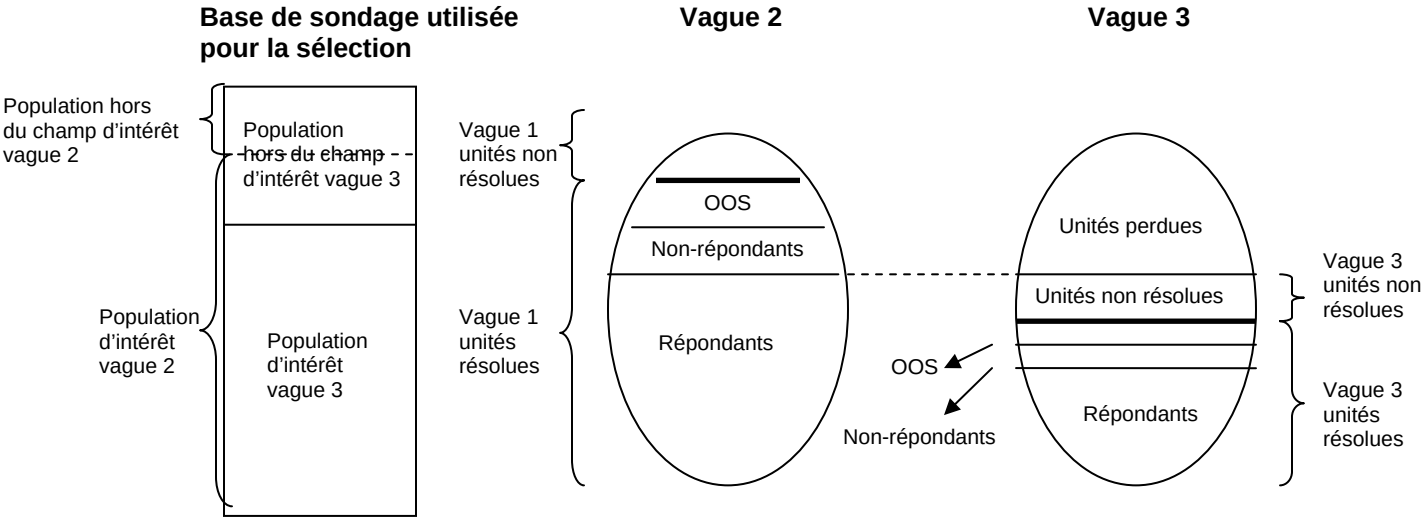
où S_{RR} = unités échantillonnées résolues qui sont des répondants

S_{RN} = unités échantillonnées résolues qui sont des non-répondants

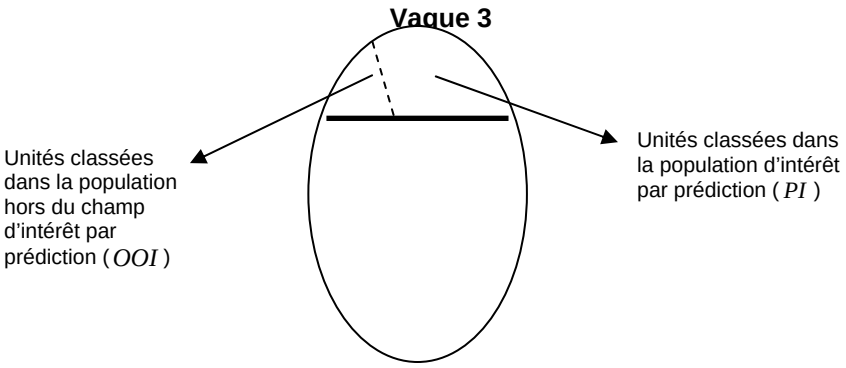
S_{RO} = unités échantillonnées résolues qui ne font pas partie de la population d'intérêt (hors du champ d'intérêt ou OOI).

Les personnes qui sont hors du champ de l'enquête sont représentées par OOS.

Le schéma suivant donne un aperçu de ces concepts dans la mesure où ils se rapportent à la pondération et démontre le passage de la base de sondage à l'échantillon de la vague 2 et ensuite à l'échantillon de la vague 3.



Conceptuellement, il est raisonnable de supposer que l'ensemble d'unités non résolues (S_U) se compose des unités faisant partie de la population d'intérêt (PI) et de la population hors du champ d'intérêt (OOI). Toutefois, on ne disposait pas de renseignements à cette étape du processus. Par conséquent, la première étape du processus de pondération consistait à prévoir, dans le cas des unités non résolues, si elles auraient été classées ou non dans la population d'intérêt. Au moyen de modèles, en utilisant les renseignements disponibles dans la base de sondage, les informations recueillies aux vagues 1 et 2 et les informations sur les unités résolues de la vague 3, on a procédé au classement par prédiction des unités non résolues dans PI ou OOI, tel qu'indiqué au schéma suivant.



Après cette première étape, nous avons pour chaque unité sélectionnée un statut (prévu ou confirmé) indiquant qu'elle faisait ou ne faisait pas partie de la population d'intérêt. Il convient de signaler que, dans le cas des unités résolues, la population d'intérêt est constituée de répondants et de non-répondants. Ainsi, nous utiliserons dans les sections qui suivent la notation suivante :

Dans le cas des unités non résolues (S_U) :

$j \in S_U$ = unités échantillonnées non résolues où $S_U = \hat{S}_{U\_PI} + \hat{S}_{U\_OOI}$

$j \in \hat{S}_{U\_PI}$ = unités échantillonnées non résolues classées PI par prédiction

$j \in \hat{S}_{U\_OOI}$ = unités échantillonnées non résolues classées OOI par prédiction

Dans le cas des unités résolues (S_R) :

$i \in S_R$ = unités échantillonnées résolues où $S_R = S_{RR} + S_{RN} + S_{RO}$

$i \in S_{RN}$ = unités non-répondantes résolues

$i \in S_{RR}$ = unités répondantes résolues

$i \in S_{RO}$ = unités OOI résolues

12.3 Pondération longitudinale des immigrants répondants

La stratégie de pondération de l'ELIC repose sur une série d'ajustements appliqués en cascade. Le poids longitudinal final est obtenu par une série d'ajustements du poids initial. Le poids final est défini à partir de quatre poids intermédiaires, soit le poids initial, le poids d'ajustement pour la non-réponse, le poids d'ajustement pour la non-résolution et le poids post-stratification. Le tableau 12.1 montre la relation entre les diverses catégories de résultats liées à l'ajustement.

Tableau 12.1 Processus de classification du statut des répondants vague 3

<table>
<tr><th>Sous-échantillon</th><th>Dépistage</th><th>Statut</th><th colspan="2">Réponse</th></tr>
<tr><td rowspan="6">Unités répondantes à la vague 2</td><td rowspan="5">Unités résolues</td><td rowspan="5">PI : Unités faisant partie du champ d'enquête</td><td colspan="2">Unités répondantes</td></tr>
<tr><td rowspan="4">Unités non-répondantes</td><td>Refus</td></tr>
<tr><td>Problèmes de langue</td></tr>
<tr><td>RL absent</td></tr>
<tr><td>Autre non-réponse</td></tr>
<tr><td colspan="3">OOI (a quitté le Canada, décédé, etc.)</td></tr>
<tr><td colspan="4">Unités non résolues</td></tr>
</table>

À noter que dans le fichier de microdonnées, seules les unités répondantes résolues, ($i \in S_{RR}$), ont un poids final puisque ce sont les seules pour lesquelles les enregistrements sont complets. La population hors du champ d'intérêt ($i \in S_{RO}$) a aussi un poids final, mais les données sur cette population ne sont pas disponibles dans le fichier de microdonnées puisque tous les enregistrements ne sont pas complets. Seules des totalisations pour cette sous-population utilisant les poids finaux sont disponibles.

Dans les sections qui suivent, nous décrivons le poids initial, (section 12.3.1), les deux poids d'ajustement, c.-à-d. l'ajustement pour la non-réponse et l'ajustement pour les unités non résolues (section 12.3.2), puis nous expliquons la post-stratification à la section 12.3.3.

12.3.1 Poids initial

Au moment de la sélection, un poids de sondage initial est attribué à la personne sélectionnée. Ce poids est simplement l'inverse de la probabilité de sélection des immigrants et cette dernière est fonction de la méthode de sélection. Comme une méthode d'échantillonnage à deux degrés a été employée pour l'ELIC, le poids de sondage attribué à chaque personne sélectionnée est égal à l'inverse de la probabilité de sélection de l'unité immigrante dans laquelle la personne se trouve, multiplié par le nombre de personnes admissibles dans cette unité immigrante.

Pour la pondération de la vague 1, le poids initial était le poids de sondage décrit ci-haut. Ce poids a ensuite été ajusté pour tenir compte de la non-réponse et de la non-résolution. Un ajustement de post-stratification a finalement été appliqué afin d'être en accord avec des chiffres de population mis à jour. Pour de plus amples détails sur le poids de sondage et les différents ajustements de la vague 1, veuillez consulter le chapitre 10.0 du guide de l'utilisateur de la vague 1.

Pour la pondération de la vague 2, le poids initial est le poids avant post-stratification de la vague 1, c'est-à-dire le poids de sondage ajusté pour la non-réponse et la non-résolution de la vague 1. Pour plus de détails consulter le chapitre 12.0 du guide de l'utilisateur de la vague 2.

Pour la pondération de la vague 3, le poids initial est le poids avant post-stratification de la vague 2 et est formulé comme suit :

$$Poid\ initial = (poids\ de\ sondage)*(ajustement\ pour\ la\ non-réponse_{Vague\ 1})*(ajustement\ pour\ la\ non-résolution_{Vague\ 1})*$$
$$(ajustement\ pour\ la\ non-réponse_{Vague\ 2})*(ajustement\ pour\ la\ non-résolution_{Vague\ 2})*$$

Algébriquement, le poids initial pour la pondération de la vague 3 est :

$$w_{initial} = w_D * \left[\frac{\sum_{G_1^{(1)}}\sum_{i\in S_{RR}} w_D + \sum_{G_1^{(1)}}\sum_{i\in S_{RN}} w_D}{\sum_{G_1^{(1)}}\sum_{i\in S_{RR}} w_D}\right] * \left[\frac{\sum_{G_2^{(1)}}\sum_{j\in \hat{S}_{U\_PI}} w_D + \sum_{G_2^{(1)}}\sum_{i\in S_{R\_PI}} w_1}{\sum_{G_2^{(1)}}\sum_{i\in S_{R\_PI}} w_1}\right] *$$
$$\left[\frac{\sum_{G_1^{(2)}}\sum_{i\in S_{RR}} w_{D^*} + \sum_{G_1^{(2)}}\sum_{i\in S_{RN}} w_{D^*}}{\sum_{G_1^{(2)}}\sum_{i\in S_{RR}} w_{D^*}}\right] * \left[\frac{\sum_{G_2^{(2)}}\sum_{j\in \hat{S}_{U\_PI}} w_{D^*} + \sum_{G_2^{(2)}}\sum_{i\in S_{R\_PI}} w_2}{\sum_{G_2^{(2)}}\sum_{i\in S_{R\_PI}} w_2}\right]$$

où $w_1 = w_D * \left[\frac{\sum_{G_1^{(1)}}\sum_{i\in S_{RR}} w_D + \sum_{G_1^{(1)}}\sum_{i\in S_{RN}} w_D}{\sum_{G_1^{(1)}}\sum_{i\in S_{RR}} w_D}\right]$ et,

$$w_2 = w_{D^*} * \left[\frac{\sum_{G_1^{(2)}}\sum_{i\in S_{RR}} w_{D^*} + \sum_{G_1^{(2)}}\sum_{i\in S_{RN}} w_{D^*}}{\sum_{G_1^{(2)}}\sum_{i\in S_{RR}} w_{D^*}}\right]$$

où $w_{initial}$ = poids initial pour la vague 3

$G_1^{(1)}$ = classe d'ajustement pour la non-réponse de la vague 1

$G_1^{(2)}$ = classe d'ajustement pour la non-réponse de la vague 2

$G_2^{(1)}$ = classe d'ajustement pour la non-résolution de la vague 1

$G_2^{(2)}$ = classe d'ajustement pour la non-résolution de la vague 2

w_D = poids de sondage (pour plus de détail, voir la section 10.3 du guide de l'utilisateur de la vague 1)

w_{D^*} = poids de sondage ajusté pour la non-réponse et la non-résolution de la vague 1

12.3.2 Ajustement des poids pour la non-réponse et les cas non résolus

Dans le cas des unités répondantes résolues de la vague 3 ($i \in S_{RR}$), l'ajustement de la pondération est formulé comme suit [avant ajustement pour la post-stratification] :

$$Poids\ \text{int}ermédiaire = (poids\ \ initial)*(ajustement\ pour\ la\ \ non-réponse)*(ajustement\ pour\ la\ \ non-résolution)$$

ou

$$Poids\ \text{int}\ ermédiaire\ =\ poids\ initial\ *\frac{somme\ \ pondérée\ \ des\ unités\ \ résolues\ \ (répondants\ \ et\ non-répondants\ )}{somme\ \ pondérée\ \ des\ répondants}*\frac{somme\ \ pondérée\ \ des\ unités\ \ résolues\ \ et\ résolues\ \ PI\ \ par\ \ prédiction}{somme\ \ pondérée\ \ des\ unités\ \ résolues}$$

ou algébriquement

$$w_{\text{int}\_PI} = w_{initial} * \left[\frac{\sum_{G_1^{(3)}}\sum_{i\in S_{RR}} w_{initial} + \sum_{G_1^{(3)}}\sum_{i\in S_{RN}} w_{initial}}{\sum_{G_1^{(3)}}\sum_{i\in S_{RR}} w_{initial}}\right] * \left[\frac{\sum_{G_2^{(3)}}\sum_{j\in \hat{S}_{U\_PI}} w_{initial} + \sum_{G_2^{(3)}}\sum_{S_{R\_PI}} w_3}{\sum_{G_2^{(3)}}\sum_{i\in S_{R\_PI}} w_3}\right]$$

$$où\ w_3 = w_{initial} * \left[\frac{\sum_{G_1^{(3)}}\sum_{i\in S_{RR}} w_{initial} + \sum_{G_1^{(3)}}\sum_{i\in S_{RN}} w_{initial}}{\sum_{G_1^{(3)}}\sum_{i\in S_{RR}} w_{initial}}\right]$$

et $G_1^{(3)}$ = classe d'ajustement pour la non-réponse de la vague 3

$G_2^{(3)}$ = classe d'ajustement pour la non-résolution de la vague 3

W_{int_PI} = poids intermédiaire à la vague 3de la population d'intérêt PI

$W_{initial}$ = poids initial à la vague 3

Note : La section 12.3.4 traite plus en détails du concept des classes d'ajustement pour la non-réponse et la non-résolution.

Pour la population hors du champ d'intérêt résolue de la vague 3 ($i \in S_{RO}$), il n'y a qu'un ajustement, c.-à-d. un ajustement pour compenser pour la population classée hors du champ d'intérêt par prédiction ($j \in \hat{S}_{U\_OOI}$) dans l'ajustement pour la non-résolution.

$$w_{\text{int}\_OOI} = w_{initial} * \left[\frac{\sum_{G_2^{(3)}}\sum_{j\in \hat{S}_{U\_OOI}} w_{initial} + \sum_{G_2^{(3)}}\sum_{i\in S_{RO}} w_{initial}}{\sum_{G_2^{(3)}}\sum_{i\in S_{RO}} w_{initial}}\right]$$

12.3.3 Post-stratification

La post-stratification a pour but d'assurer la cohérence entre les estimations produites à partir de l'enquête et les estimations démographiques produites par une source externe indépendante. Or, comme les poids finaux de la vague 3 de l'ELIC donnent des estimations de la population d'intérêt de la vague 3, et non de la population cible (voir la section 12.1 ***Représentativité des poids***) et qu'il n'existe pas de source administrative externe indépendante à ce sujet (comme à la vague 2), les totaux de post-stratification doivent être estimés. On peut estimer les totaux de post-stratification pour la vague 3 de la façon suivante :

$$\hat{N}_k^{(3)} = \hat{N}_k^{(2)} - \sum_{i \in k \cap OOI} W_{fi}^{(2)}$$
$$= \sum_{i \in k \cap PI} W_{fi}^{(2)}$$

où $\hat{N}_k^{(3)}$ = la taille estimée de la population d'immigrants PI dans la post-strate k (le total de post-stratification de la post-strate k pour la vague 3)

$\hat{N}_k^{(2)}$ = la taille estimée de la population d'immigrants dans la post-strate k (le total de post-stratification de la post-strate k pour la vague 2)

$W_{fi}^{(2)}$ = le poids final de la vague 2 de l'immigrant i

Pour l'échantillon de la vague 3, la population d'intérêt est l'ensemble des immigrants de l'ELIC qui sont toujours au Canada quatre ans après leur arrivée. Par conséquent, l'ajustement de post-stratification pour cet échantillon assure la cohérence entre la somme des poids et l'estimation démographique associée à cette période pour chaque combinaison d'âge, sexe, lieu de naissance (agrégé selon la région du monde) et catégorie d'immigrants. Les catégories détaillées sont présentées dans les tableaux 12.2 à 12.5.

Tableau 12.2 Groupes d'âge

15 à 24 ans
25 à 34 ans
35 à 44 ans
45 ans et plus

Tableau 12.3 Sexe

Hommes
Femmes

Tableau 12.4 Lieux de naissances

Régions	Régions du monde (WA)
Afrique centrale	1 – Afrique
Afrique de l'Est	
Afrique du Nord	
Afrique du Sud	
Afrique de l'Ouest	
Amérique centrale	2 – Amérique
Amérique du Nord	
Amérique du Sud	
Caraïbes et Bermudes	
Asie de l'Est	3 – Asie
Asie du Sud-Est	
Asie du Sud	
Asie centrale de l'Ouest et Moyen-Orient	
Europe de l'Est	4 – Europe
Europe du Nord	
Europe du Sud	
Europe de l'Ouest	
Océanie	5 – Océanie

Tableau 12.5 Catégories d'immigrants

Catégorie de la famille
Catégorie économique – Travailleurs qualifiés (Demandeur principal)
Catégorie économique – Travailleurs qualifiés (Conjoint(e) et personnes à charge)
Catégorie économique – Gens d'affaires indépendants et autres immigrants indépendants
Réfugiés parrainés par le gouvernement
Réfugiés autres

Les variables font l'objet de totalisations croisées sauf dans les cas suivants :

Afrique

- Pour les hommes et femmes dans la catégorie de la famille, les groupes d'âge 25 à 34 et 35 à 44 sont regroupés.
- Pour les hommes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés.

- Pour les femmes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés, ainsi que les groupes 35 à 44 et 45 et plus.
- Pour les hommes et les femmes dans la catégorie économique – travailleurs qualifiés (conjoint(e) et personnes à charge) les groupes d'âge 35 à 44 et 45 et plus sont regroupés.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, on ne fait pas la distinction par sexe ni par groupe d'âge.
- Pour les réfugiés parrainés par le gouvernement les groupes d'âge 35 à 44 et 45 et plus sont regroupés pour les hommes et les femmes.
- Pour les réfugiés autres on ne fait pas la distinction par sexe ni par groupe d'âge.

Amérique

- Pour les hommes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés.
- Pour les femmes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés, ainsi que les groupes 35 à 44 et 45 et plus.
- Pour les hommes et les femmes dans la catégorie économique – travailleurs qualifiés (conjoint(e) et personnes à charge) les groupes d'âge 35 à 44 et 45 et plus sont regroupés.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, on ne fait pas la distinction par sexe ni par groupe d'âge.
- Pour les réfugiés parrainés par le gouvernement les groupes d'âge 35 à 44 et 45 et plus sont regroupés pour les hommes.
- Pour les réfugiés parrainés par le gouvernement les groupes d'âge 15 à 24 et 25 à 34 sont regroupés pour les femmes, ainsi que les groupes 35 à 44 et 45 et plus.
- Pour les réfugiés autres on ne fait pas la distinction par sexe ni par groupe d'âge.

Asie

- Pour les hommes et les femmes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, on ne fait pas la distinction par sexe.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, les groupes d'âge 25 à 34 et 35 à 44 sont regroupés.
- Pour les réfugiés parrainés par le gouvernement les groupes d'âge 35 à 44 et 45 et plus sont regroupés pour les hommes et les femmes.
- Pour les réfugiés autres on ne fait pas la distinction par sexe ni par groupe d'âge.

Europe

- Pour les femmes dans la catégorie économique – travailleurs qualifiés (demandeur principal), les groupes d'âge 15 à 24 et 25 à 34 sont regroupés, ainsi que les groupes 35 à 44 et 45 et plus.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, on ne fait pas la distinction par sexe.
- Pour la catégorie économique – gens d'affaires indépendants et autres immigrants indépendants, les groupes d'âge 25 à 34 et 35 à 44 sont regroupés.

- Pour les réfugiés parrainés par le gouvernement les groupes d'âge 35 à 44 et 45 et plus sont regroupés pour les hommes et les femmes.
- Pour les réfugiés autres on ne fait pas la distinction par sexe ni par groupe d'âge.

Océanie

- Pour la catégorie de la famille on ne fait pas la distinction par sexe ni par groupe d'âge.
- Toutes autres catégories d'immigrants sont regroupées, sans faire la distinction par sexe ou groupe d'âge.

L'ajustement prend la forme suivante :

$$\text{Poids final} = \text{Poids intermédiaire} * \frac{\text{Taille estimée de la population d'immigrants PI}}{\text{Estimation des chiffres de la population utilisant les poids intermédiaires}}$$

ou algébriquement pour $i \in S_{RR}$,

$$W_f = \sum_{i \in S_{RR}} w_{\text{int\_}PI} * \frac{\hat{N}_k^{(3)}}{\sum_k \sum_{i \in S_{RR}} w_{\text{int\_}PI} + \sum_k \sum_{i \in S_{RO}} w_{\text{int\_}OOI}}$$

12.3.4 Classes d'ajustement : Groupes homogènes

Les classes d'ajustement de la pondération et les groupes de post-stratification reposent sur la même hypothèse. Ils doivent être des groupes homogènes ayant un rapport avec la correction apportée : les classes d'ajustement de la non-réponse se fondent sur l'homogénéité des réponses d'une même classe, ce qui signifie que la probabilité de réponse est la même; les classes d'ajustement des cas non résolus regroupent des cas homogènes ou des cas présentant la même propension à être résolus et à faire partie du champ de l'enquête.

Pour l'ELIC, on a établi les catégories d'ajustement pour la non-réponse et la non-résolution par régression logistique permettant de prédire, respectivement, la probabilité de réponse et la probabilité de résolution. Dans le cas de ce dernier modèle, les variables explicatives utilisées pour prédire l'appartenance à la population d'intérêt étaient incluses dans le modèle par défaut.

Les prédicateurs ou variables explicatives du modèle de prédiction de réponse étaient :

- Groupe d'âge comme défini à la vague 2
- Pays de citoyenneté provenant de la base de sondage (Système de soutien des opérations des bureaux locaux (SSOBL), c'est-à-dire la base de données administratives de Citoyenneté et Immigration Canada)
- Classe d'immigrant du répondant longitudinal (RL) (SSOBL)
- Code de bureau Citoyenneté et immigration Canada à l'arrivée (SSOBL)
- Variable de vague 1 indiquant si le RL a décidé de venir au Canada pour démarrer une entreprise
- Variable de vague 1 indiquant si le RL a obtenu une carte santé provinciale
- Variable de vague 1 indiquant si le RL a choisi de vivre dans la ville où il habite actuellement en raison de la (les) langue(s) parlée(s) là
- Variable de vague 2 indiquant qui a répondu aux questions sur le revenu

- Variable de vague 2 indiquant si le RL a reçu un revenu provenant de sources de l'extérieur du Canada
- Variable de vague 2 indiquant le degré de satisfaction du RL en ce qui concerne la vie au Canada
- Variable de vague 2 indiquant si le RL viendrait au Canada s'il avait à choisir de nouveau
- Variable de vague 2 indiquant le degré de satisfaction du RL en ce qui concerne leur bien-être matériel au Canada
- Variable de vague 1 indiquant si le RL est membre d'un groupe ou d'une organisation au Canada
- Code de priorité de l'interview vague 2
- Taux de coopération du RL à la vague 1

Les variables explicatives du modèle de prédiction de la propension à être **résolu** étaient :

- Nombre de membres dans l'unité immigrante du RL noté à la vague 1
- Province de destination (SSOBL)
- Groupe d'âge comme défini à la vague 2
- pays de dernière résidence permanente du RL (SSOBL)
- indicateur des homonymes
- Code de priorité de l'interview vague 1
- Variable de vague 2 indiquant dans quelle mesure il est important que les gens qui donnent des soins de santé soient du même groupe ethnique ou culturel que le RL
- Variable de vague 1 indiquant si le RL prévoit de s'établir au Canada
- Variable de vague 1 indiquant la chose la plus utile qui a facilité l'installation du RL au Canada
- Taux de coopération du RL à la vague 2
- État matrimonial du RL à la vague 1

Dans ce modèle, les **statuts prédits d'appartenance à la population d'intérêt,** étaient inclus par défaut. Les variables explicatives étaient :

- Province de destination (SSOBL)
- Code de priorité de l'interview vague 1
- Nombre de membres dans l'unité immigrante du RL noté à la vague 1
- Classe d'immigrant du RL (SSOBL)
- Variable indiquant l'interview de la vague 2 s'est déroulé dans une des langues officielles (anglais ou français) ou non
- Variable de vague 1 indiquant si le RL prévoit de s'établir au Canada
- Variable de vague 2 indiquant dans quelle mesure il est important que les gens qui donnent des soins de santé parlent la même langue que le RL
- Variable de vague 1 indiquant la chose la plus utile qui aurait facilité l'installation du RL au Canada

Les catégories ont été construites en utilisant des probabilités similaires obtenues des modèles respectifs. On a établi le nombre de catégories pour chaque ajustement selon un algorithme de convergence de manière à garantir des estimations non biaisées.

13.0 Qualité des données et couverture

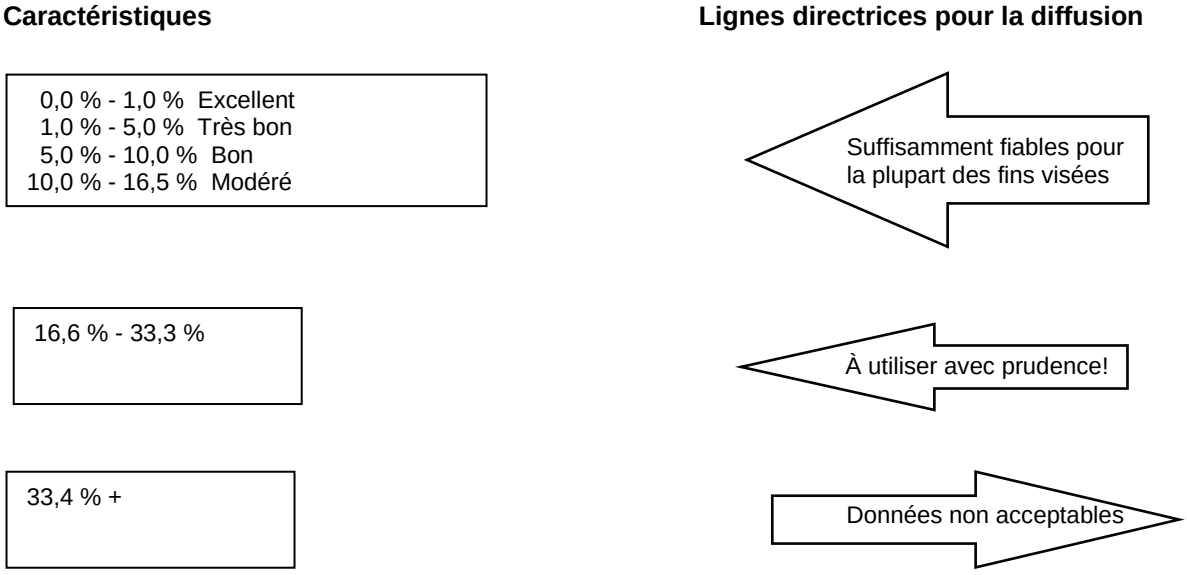
Ce chapitre permet à l'utilisateur de prendre connaissance des différents éléments qui influent sur la qualité des données de l'enquête. Les erreurs sont divisées en deux grandes catégories : les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage. L'erreur d'échantillonnage est l'écart entre une estimation obtenue à partir d'un échantillon et celle que donnerait un recensement pour lequel on a utilisé les mêmes méthodes de collecte des données. Tous les autres types d'erreurs, comme l'erreur de couverture de la base de sondage, de réponse, de traitement et de non-réponse, sont des erreurs non dues à l'échantillonnage. Bon nombre de ces erreurs sont difficiles à identifier et à quantifier. Voir à ce sujet la section 13.2.

13.1 Erreurs d'échantillonnage

Les estimations dérivées de cette enquête sont fondées sur un échantillon d'immigrants et non sur un dénombrement complet (recensement) effectué dans des conditions similaires. La différence est appelée erreur d'échantillonnage. Les *normes et lignes directrices concernant la documentation sur la qualité des données et la méthodologie*[1] énoncent qu'il faut donner aux utilisateurs externes une indication de l'ampleur de l'erreur d'échantillonnage. Il est fortement recommandé que les utilisateurs qui analysent les données ou produisent des estimations à partir des fichiers de données de l'Enquête longitudinale auprès des immigrants du Canada (ELIC) fournissent également à leur public des indicateurs de la qualité des données.

La mesure de l'importance éventuelle des erreurs d'échantillonnage est fondée sur l'erreur-type des estimations, estimées des résultats de l'enquête. Cependant, en raison de la diversité des estimations que l'on peut tirer d'une enquête, l'erreur-type d'une estimation est habituellement exprimée en fonction de l'estimation à laquelle elle se rapporte. La mesure résultante, appelée coefficient de variation (CV) d'une estimation, s'obtient en exprimant l'erreur-type de l'estimation en pourcentage de l'estimation elle-même. Plus le CV est petit, plus la variabilité d'échantillonnage est faible, de sorte que les CV plus faibles sont plus souhaitables. Le CV est fonction de la taille de l'échantillon sur lequel l'estimation est fondée, la taille de la population et la répartition de l'échantillon, c.-à-d. la fraction de sondage, des unités du domaine estimé. Le schéma suivant présente les caractéristiques de certains coefficients de variation et les lignes directrices de Statistique Canada quant à la diffusion.

[1] Statistique Canada. *Normes et lignes directrices concernant la documentation sur la qualité des données et la méthodologie*, 2002, www.statcan.ca/francais/about/policy/infousers_f.htm.



13.2 Erreurs non due à l'échantillonnage

Les sources d'erreurs non dues à l'échantillonnage sont nombreuses et ces erreurs peuvent survenir à pratiquement n'importe quelle étape d'une enquête. Il est possible que les intervieweurs comprennent mal les instructions, que les répondants fassent des erreurs en répondant aux questions, que des réponses soient mal inscrites sur le questionnaire et que des erreurs soient introduites au moment du traitement et de la totalisation des données. Pour l'ELIC, des mesures d'assurance de la qualité ont été mises en œuvre à chaque étape et cycle de collecte et de traitement des données pour surveiller la qualité des données. Ces mesures comprenaient la formation des intervieweurs portant sur les procédures d'enquête et le questionnaire, l'observation des interviews pour déceler les problèmes de conception du questionnaire ou de mauvaise interprétation des instructions, la surveillance du codage final et les vérifications de la qualité du codage et de la vérification des données afin de confirmer la logique de traitement. Les procédures de traitement des données sont présentées au chapitre 9.0. D'autres types d'erreurs non dues à l'échantillonnage sont plus facilement quantifiables, particulièrement la non-réponse et la qualité de la couverture de la population; ces sujets sont abordées dans les deux sections qui suivent.

13.3 Non-réponse et cas non résolus

La non-réponse et les cas non résolus, s'ils ne sont pas corrigés de la façon appropriée, sont des types d'erreur qui peuvent entraîner des biais dans les estimations de l'enquête. Pour l'ELIC, ces deux catégories de réponses ont réduit significativement le nombre d'enregistrements utilisables. Des estimations biaisées peuvent être produites si les caractéristiques des non-répondants diffèrent considérablement de celles des répondants. Comme à la vague 1, on a effectué des études afin de comprendre le mécanisme de la non-réponse. Les résultats ont montré que les unités non-répondantes et les unités non résolues affichaient des tendances différentes et des taux différents ont été obtenus pour des caractéristiques différentes des immigrants.

Après de nombreuses études des différents taux et des différentes caractéristiques, il était raisonnable de supposer des tendances et propensité de réponse et de résolution non aléatoires.Les tendances observées pour les unités répondantes et non-répondantes ainsi que pour les unités résolues et non résolues étaient différents. Il faut habituellement corriger les tendances non aléatoires en utilisant comme classes d'ajustement, les caractéristiques/variables

qui démontrent des tendances différentes. Par exemple, si le sexe est une variable explicative dans le modèle de prédiction de la réponse (c.-à-d. des taux de réponse différents pour les hommes et pour les femmes), alors il faut utiliser le sexe pour la correction.

Pour ces raisons, on a calculé les ajustements de poids par étapes distinctes pour les unités répondantes et puis, pour les unités résolues, tel que décrit à la section 12.3. Des modèles de prédiction de réponse et de résolution ont été utilisés pour définir les classes d'ajustement appropriées pour corriger pour les différents taux de réponse et taux de résolution. On souligne l'importance d'utiliser les poids finaux aux fins de toute totalisation ou analyse à partir des données de l'ELIC. Tout résultat provenant d'estimation non-pondérée est un résultat biaisé.

13.4 Couverture

La couverture indique dans quelle mesure la base de sondage couvre la population cible ou, dans le cas de l'ELIC, la population d'intérêt. Il y a surdénombrement lorsque, par exemple, la base de sondage comprend des unités qui n'auraient pas dû être incluses, par exemple décès, enregistrements en double ou date de naissance incorrecte saisie dans le fichier. Il y a sous-dénombrement lorsque, par exemple, la base de sondage ne comprend pas certaines unités qui auraient dû être incluses. Pour la première vague, il y avait un léger surdénombrement, qui a été corrigé par l'application d'une technique de poststratification à un fichier plus à jour. Faute d'une source plus fiable, le même fichier a été utilisé pour la deuxième vague et encore pour la troisième vague (voir la section 12.3.3). Par conséquent, la taille de la population d'intérêt de la troisième vague est en soi une estimation fondée sur les données de la première vague et les résultats de la collecte de la deuxième et la troisième vague.

14.0 Lignes directrices pour la totalisation, l'analyse et la diffusion de données

Ce chapitre de la documentation renferme un aperçu des lignes directrices que doivent respecter les utilisateurs qui totalisent, analysent, publient ou autrement diffusent des données calculées à partir des fichiers de microdonnées de l'enquête. Ces lignes directrices devraient permettre aux utilisateurs de microdonnées de produire les mêmes chiffres que ceux produits par Statistique Canada, tout en étant en mesure d'obtenir des chiffres actuellement inédits de façon conforme à ces lignes directrices établies.

14.1 Lignes directrices pour l'arrondissement d'estimations

En premier lieu, il faut établir une distinction entre l'arrondissement destiné à protéger la confidentialité du répondant et l'arrondissement effectué aux fins de la précision implicite. L'arrondissement est souvent utilisé comme contrôle de la divulgation pour éviter que des résultats publiés soient reliés à des répondants individuels dans un fichier de microdonnées à grande diffusion (FMGD). Comme aucun FMGD n'est, ni ne sera, produit pour l'Enquête longitudinale auprès des immigrants du Canada (ELIC), il n'y a aucune raison de craindre que des liens puissent être établis entre les résultats. Ceci étant dit, l'ELIC publie des données géographiques détaillées et, étant donné le caractère reconnaissable des répondants de l'ELIC, **les calculs pondérés fondés sur la géographie infraprovinciale doivent être arrondis à la cinquantaine près**. Les utilisateurs de fichiers de microdonnées de l'ELIC doivent se conformer aux lignes directrices suivantes concernant l'arrondissement de ces estimations :

a) Les estimations qui figurent dans le corps d'un tableau statistique doivent être arrondies à la cinquantaine près selon la méthode d'arrondissement classique.

b) Les totaux partiels marginaux et les totaux marginaux des tableaux statistiques doivent être calculés d'après leurs éléments correspondants non arrondis, puis arrondis à leur tour à la cinquantaine près selon la méthode d'arrondissement classique. Il est également acceptable, du point de vue de la confidentialité, de calculer les totaux marginaux en utilisant les comptes arrondis.

c) Les moyennes, les proportions, les taux et les pourcentages doivent être calculés à partir d'éléments arrondis (par exemple les numérateurs ou les dénominateurs).

d) Les sommes et les différences d'agrégat (ou de rapports) doivent être calculées à partir de leurs éléments correspondants arrondis, puis arrondies à leur tour à la cinquantaine près selon la méthode d'arrondissement classique.

L'arrondissement est également utilisé pour ne pas donner l'impression que les estimations sont plus précises qu'elles ne le sont en réalité. Dans une grande partie de la recherche de l'ELIC publiée et produite par Statistique Canada, l'arrondissement tel qu'il est décrit ci-dessus est utilisé à la centaine près. Pour que les résultats publiés soient comparables, nous conseillons vivement aux utilisateurs de s'en tenir à cette pratique. Ceci étant dit, dans les cas où les estimations devant être publiées ou diffusées d'une autre façon sont différentes des estimations correspondantes publiées par Statistique Canada, nous conseillons vivement aux utilisateurs d'indiquer la raison de ces divergences dans le ou les documents à publier ou à diffuser.

14.2 Lignes directrices pour la pondération de l'échantillon en vue de la totalisation

Le plan d'échantillonnage utilisé pour l'ELIC était auto-pondéré. Lorsqu'ils produisent des estimations simples, y compris des tableaux statistiques ordinaires, les utilisateurs doivent appliquer le poids final. Si l'on n'utilise pas des poids finaux, on ne peut considérer les

estimations calculées à partir des fichiers de microdonnées représentatives de la population visée par l'enquête et ces estimations ne correspondront pas à celles produites par Statistique Canada. De fait, le poids attribué à chaque immigrant reflète le nombre d'immigrants représentés par un répondant donné.

Les utilisateurs devraient également noter que certains progiciels pourraient peut-être ne pas permettre la production d'estimations correspondant exactement à celles qu'offre Statistique Canada, en raison du mode de traitement du champ de poids par ces progiciels (p.ex. troncation ou arrondissement de poids qui ne sont pas des nombres entiers).

Le fichier vague 3 de l'ELIC a été établi de façon à ce que le répondant longitudinal constitue l'unité d'analyse. Le poids qui paraît sur chaque enregistrement (WT3L) est un poids correspondant à un immigrant (le répondant longitudinal). Les analyses utilisant les enfants, le conjoint, la famille ou le ménage du répondant comme unité d'analyse ne peuvent être effectuées au moyen des données de l'ELIC. Toutes les questions de recherche doivent être formulées en terme du répondant longitudinal.

14.3 Définitions de types d'estimations : catégoriques et quantitatives

Estimations catégoriques

Les estimations catégoriques sont des estimations du nombre ou du pourcentage de membres de la population visée par l'enquête possédant certaines caractéristiques ou faisant partie d'une catégorie définie. Le nombre ou la proportion d'immigrants qui prévoient acheter une maison ou un appartement au cours des prochaines années constituent des exemples de telles estimations. On peut aussi appeler une estimation du nombre de personnes possédant une certaine caractéristique une estimation d'un agrégat.

Exemples de questions catégoriques :

Q : Est-ce que vous ou vous et votre famille planifiez d'acheter une maison ou un appartement au cours des prochaines années?
R : Oui / Non / N'est pas certain(e)

Q : Quel est votre niveau de satisfaction à l'égard de votre logement?
R : Très satisfait(e) / Satisfait(e) / Insatisfait(e) / Très insatisfait(e)

Estimations quantitatives

Les estimations quantitatives sont des estimations de totaux ou de moyennes, de médianes ou d'autres mesures d'une tendance centrale de quantités reposant sur certains ou sur la totalité des membres de la population visée par l'enquête. Elles comprennent aussi expressément des estimations de la forme $\hat{X}/\hat{Y}$ où $\hat{X}$ est une estimation de la quantité totale de membres de la population visée par l'enquête et $\hat{Y}$ une estimation du nombre de personnes de la population visée par l'enquête ayant contribué à en arriver à cette quantité totale.

Un exemple d'estimation quantitative est le montant mensuel moyen payé en coûts de location ou de logement. Le numérateur est une estimation du montant total payé par mois pour les immigrants qui habitent en logement et le dénominateur est le nombre d'immigrants qui habitent en logement.

Exemples de questions quantitatives :

Q : Combien payez-vous par mois pour votre logement? (Inclure «le loyer, les taxes, le chauffage, l'alimentation en eau, l'électricité, le stationnement, les frais de copropriété/l'hypothèque, etc., mais exclure les frais de téléphone et de câble.)
R : |_|_|_|_| $/mois

Q : Dans cet emploi, quel est/était votre salaire ou traitement avant impôts et autres déductions?
R : |_|_|_|_|_|_| $

14.3.1 Totalisation d'estimations catégoriques

On peut obtenir des estimations du nombre d'immigrants possédant une certaine caractéristique à partir des fichiers de microdonnées en additionnant les poids finals de tous les enregistrements présentant la ou les caractéristiques qui nous intéressent. Ces estimations peuvent être transversales ou longitudinales. On obtient des proportions et des rapports de la forme $\hat{X}/\hat{Y}$ en :

a) additionnant les poids finals des enregistrements présentant la caractéristique qui nous intéresse pour le numérateur $\left(\hat{X}\right)$,
b) additionnant les poids finals des enregistrements présentant la caractéristique qui nous intéresse pour le dénominateur $\left(\hat{Y}\right)$, puis en
c) divisant l'estimation a) par celle de b) $\left(\hat{X}/\hat{Y}\right)$.

14.3.2 Totalisation d'estimations quantitatives

On peut obtenir des estimations de quantités à partir des fichiers de microdonnées en multipliant la valeur de la variable qui nous intéresse par le poids final établi pour chaque enregistrement, puis en additionnant cette quantité pour tous les enregistrements qui nous intéressent. Par exemple, pour obtenir une estimation du montant total payé mensuellement en coûts de logement, multipliez le montant mensuel en coûts de logement de l'immigrant par le poids final établi pour l'enregistrement, puis additionnez cette valeur pour tous les enregistrements indiquant un immigrant qui habite en logement.

Pour obtenir une moyenne pondérée de la forme $\hat{X}/\hat{Y}$, le numérateur $\left(\hat{X}\right)$ est calculé comme une estimation quantitative et le dénominateur $\left(\hat{Y}\right)$ est calculé comme une estimation catégorique. Pour estimer, par exemple, le montant mensuel moyen payé pour le logement par les immigrants habitant en logement :

a) estimez le montant mensuel total payé en coûts de logement $\left(\hat{X}\right)$ tel qu'il est décrit ci-dessus,
b) estimez le nombre d'immigrants qui habitent en logement $\left(\hat{Y}\right)$ en additionnant les poids finals de tous les enregistrements correspondant à cette catégorie, puis
c) divisez l'estimation a) par l'estimation b) $\left(\hat{X}/\hat{Y}\right)$.

14.4 Lignes directrices pour l'analyse statistique

L'ELIC repose sur un plan d'échantillonnage complexe comportant une stratification, plusieurs degrés de sélection et des probabilités inégales de sélection des répondants. L'utilisation des données tirées d'enquêtes aussi complexes présente des problèmes pour les analystes, parce que le plan d'enquête et les probabilités de sélection influencent les procédures d'estimation et de calcul de la variance qui doivent être utilisées. Il faut utiliser des poids de l'enquête pour que les estimations et les analyses des données de l'enquête soient exemptes de biais.

Bien que de nombreuses procédures d'analyse que l'on trouve à l'intérieur de progiciels statistiques permettent l'utilisation de poids, la signification ou la définition du poids inclus dans ces procédures diffère de ce qui convient à la base de sondage d'une enquête-échantillon, de sorte que les estimations produites au moyen de ces progiciels sont correctes dans bien des cas, les variances qui sont estimées sont mauvaises. Les variances approximatives d'estimations simples comme les totaux, les proportions et les rapports (pour les variables qualitatives et pour les domaines communs) peuvent être calculées à partir du Module d'extraction de coefficients de variation (MECV) vague 3 de l'ELIC qui est fourni comme outil complémentaire. Le MECV est examiné à la section 15.3.

Pour d'autres techniques d'analyse (par exemple, la régression linéaire, la régression logistique et l'analyse de variance), il existe une méthode qui peut rendre les variances calculées par l'application des progiciels standards plus significatives en intégrant les probabilités inégales de sélection. L'application de cette méthode entraîne une remise à l'échelle de poids de façon à ce que le poids moyen soit de 1. Les logiciels d'analyse couramment utilisés (SAS et SPSS par exemple) comportent souvent des options dans bon nombre de procédures qui permettent de modifier l'échelle de pondération. Cependant, les variances calculées de cette façon ne tiennent pas compte des gains ou des pertes d'efficacité causés par la stratification et l'effet des grappes du plan d'échantillonnage. Les méthodes et logiciels qui permettent d'établir une estimation appropriée des variances sont examinés au chapitre 15.0.

14.5 Lignes directrices pour la diffusion de coefficients de variation

Avant de diffuser et/ou de publier toute estimation établie à partir de l'ELIC, les utilisateurs devraient premièrement déterminer le niveau de qualité de l'estimation. Les niveaux de qualité sont *acceptable*, *médiocre* et *inacceptable*. Comme il en a été question au chapitre 13.0, des erreurs d'échantillonnage et des erreurs non dues à l'échantillonnage influent sur la qualité des données. Cependant, aux fins du présent document, le niveau de qualité d'une estimation est déterminé seulement en fonction de l'erreur d'échantillonnage illustrée par le coefficient de variation, tel qu'il est indiqué au tableau ci-dessous.

On devrait premièrement déterminer le nombre d'immigrants retenus pour le calcul de l'estimation. Si ce nombre est inférieur à 10, l'estimation pondérée ne peut pas être diffusée. Pour les estimations pondérées fondées sur les tailles d'échantillons composés de 10 immigrants ou plus, les utilisateurs devraient déterminer le coefficient de variation de l'estimation et suivre les lignes directrices relatives au niveau de qualité qui figurent ci-dessous. Celles-ci devraient être appliquées aux estimations pondérées.

Lignes directrices relatives au niveau de qualité de l'estimation

Niveau de qualité de l'estimation	Lignes directrices
1) Acceptable	Les estimations proviennent d'une taille d'échantillon de 10 ou plus, et présentent de faibles coefficients de variation, de l'ordre de 0,0 à 16,5 %. Aucune mise en garde n'est requise.
2) Médiocre	Les estimations proviennent d'une taille d'échantillon de 10 ou plus, et présentent des coefficients de variation élevés, de l'ordre de 16,6 à 33,3 %. Ces estimations devraient être signalées par la lettre M (ou un quelconque identificateur similaire). Elles devraient être accompagnées d'une mise en garde avertissant les utilisateurs subséquents des niveaux élevés d'erreur associés aux estimations.
3) Inacceptable	Les estimations proviennent d'une taille d'échantillon de 10 ou plus, et présentent des coefficients de variation très élevés, supérieurs à 33,3 %. Statistique Canada recommande de ne pas diffuser d'estimations de qualité inacceptable. Si un utilisateur choisit cependant de le faire, ces estimations devraient alors être signalées à l'aide de la lettre I (ou d'un quelconque identificateur similaire) et devraient être accompagnées de la mise en garde suivante : « Nous informons l'utilisateur que ces estimations (désignées avec la lettre I) ne respectent pas les normes de qualité de Statistique Canada. Les conclusions qui reposeront sur ces données ne seront pas fiables et seront très probablement invalides. »

15.0 Calcul de la variance

L'Enquête longitudinale auprès des immigrants du Canada (ELIC) est une enquête probabiliste, c'est-à-dire, qu'un échantillon a été sélectionné pour représenter la population cible. Une certaine variabilité est associée à la sélection aléatoire d'un échantillon. Cette variabilité est identifiée comme étant l'erreur d'échantillonnage tel qu'il a été décrit à la section 13.1. À cette erreur, s'ajoutent des corrections pour tenir compte des unités de non-réponses et non résolues qui font partie de l'estimation de la variabilité. Ce chapitre explique l'importance de calculer la variance et présente divers outils pour la calculer.

15.1 Importance du calcul de la variance

La variabilité ou variance d'une estimation est une bonne indication de la qualité d'une estimation. Une estimation avec une variance trop élevée est considérée comme non fiable. Afin de quantifier ce qui est une variance trop élevée, une mesure relative de la variabilité est utilisée, à savoir le coefficient de variation (CV). Le coefficient de variation est défini comme étant le ratio de la racine carrée de la variance sur l'estimation. La racine carrée de la variance est aussi connue sous le nom d'écart-type. L'utilisation du coefficient de variation plutôt que la variance permet à l'analyste de comparer sur une même échelle des estimations de magnitudes diverses. Ainsi, il est possible de quantifier la qualité de toute estimation avec le CV.

De plus, pour différents tests statistiques tels que des hypothèses sur la différence entre deux estimations, le calcul de la variance ou du CV est requis afin de déterminer si la différence est statistiquement différente ou non. Le calcul de la variance ou du CV est donc primordial.

Méthode d'obtention de la variance d'une estimation
La complexité du plan de sondage, les ajustements de poids et la post-stratification font qu'il est presque impossible d'établir une formule exacte pour le calcul de la variance dans le cas de l'ELIC. Un excellent moyen d'obtenir une approximation de la variance réelle consiste à recourir aux méthodes de rééchantillonnage, plus particulièrement la méthode bootstrap. Cette méthode, qui se fonde sur une technique de réplication de l'échantillon, produit de bonnes approximations de la valeur réelle de la variance. Un fichier contenant 1 000 poids bootstrap est disponible. Le calcul de la variance au moyen de 1 000 poids bootstrap comprend le calcul de l'estimation en appliquant chacun de ces 1 000 poids, puis le calcul de la variance de ces 1 000 estimations.

Les paragraphes qui suivent présentent des logiciels et outils capables de produire des estimations bootstrap de variance. L'utilisation d'un ou de plusieurs de ces outils dépend du type d'analyse et du degré de précision requis.

Le Bootvar est un programme de macros permettant le calcul d'estimations de variances à partir de la méthode du bootstrap. Le programme est considéré générique , c'est-à-dire qu'il est possible de l'utiliser avec les données de tout enquête de Statistique Canada diffusant des poids bootstrap. Le Bootvar est disponible en format SAS et SPSS.

Il existe en outre des logiciels commerciaux (SUDAAN, Stata et WesVar) qui peuvent produire des estimations de variance en utilisant les poids bootstrap. L'avantage de ces logiciels est que, en plus de produire des estimations de la variance selon la méthode bootstrap pour une gamme plus large de statistiques, ils permettent d'apporter des corrections fondées sur le plan à d'autres statistiques utiles. L'utilisation des poids bootstrap avec ces logiciels est examinée dans *Comment utiliser les poids bootstrap avec WesVar et SUDAAN*[1].

[1] http://www.statcan.ca/francais/freepub/12-002-XIF/2004002/pdf/phillips_f.pdf

Le Module Excel d'extraction des CV (MECV) est un outil convivial développé en utilisant les poids bootstrap afin de produire des CV approximatifs pour les totaux et les proportions pour un grand nombre de domaines. Le MECV est surtout utile à l'étape exploratoire de l'analyse. L'application et la documentation sont incluses avec les données de l'ELIC. Vous trouverez plus de détails dans la prochaine section.

15.2 Module Excel d'extraction des coefficients de variation

Cette application a été développée au moyen de macros Excel et rendue conviviale grâce à une interface facile à utiliser qui permet à l'utilisateur d'extraire l'information souhaitée de deux façons. L'une consiste à décrire le domaine d'intérêt au moyen des neuf variables disponibles et l'autre, à préciser la taille d'un domaine. L'information affichée comprend l'estimation de la proportion, le nombre de répondants dans le domaine précisé, la population estimée de ce domaine, des statistiques élémentaires et le coefficient de variation pour la proportion sélectionnée. Ici, nous définissons un domaine comme étant les totalisations croisées des variables décrites dans le tableau à la section 15.2.1.

Plus de 44 000 domaines sont couverts par l'ensemble de chiffriers, qui donnent un CV approximatif pour huit proportions distinctes dans chaque domaine, soit, en tout, plus de 352 000 CV. Des simulations ont été exécutées pour calculer les variances, les coefficients de variation, ainsi que les intervalles de confiance à 95 % pour diverses proportions, c.-à-d. 1 %, 5 %, 10 %, 15 %, 20 %, 30 %, 40 % et 50 %. Ces proportions sont fondées sur la répartition de la population. Pour une répétition donnée, la proportion observée dans l'échantillon aléatoire peut différer de celle observée pour la population cible. Par conséquent, on a utilisé la moyenne de 100 répétitions pour tenir compte de cette variabilité.

15.2.1 Normes de qualité de Statistique Canada

Les utilisateurs des données devraient noter que, pour éviter la divulgation de renseignements confidentiels, lors de l'utilisation d'une variable dichotomique, il faut que la taille de l'échantillon et le CV correspondent simultanément aux normes permettant la publication des données. Les utilisateurs devraient toujours s'assurer de la qualité des estimations, particulièrement pour les faibles proportions calculées sur de petits domaines. Pour faciliter le repérage des CV de valeur élevée, un système de codes de couleurs est utilisé dans l'application Excel pour l'affichage d'un CV. Pour les indicateurs décrits ci-après, les couleurs utilisées sont le rouge pour un CV supérieur à 33,3 % et le jaune pour ceux compris entre 16,6 % et 33,3 %. Des renseignements supplémentaires figurent dans le Guide de l'utilisateur du MECV. Suit la liste des variables disponibles dans le MECV.

Champ	Description
Classe d'immigrant	
Groupe d'âge	
Région de résidence	
Région mondiale de naissance	
Sexe	
État matrimonial	
Statut d'emploi	
Niveau le plus élevé de scolarité	
Connaissance des langues officielles	
Proportion cible	Proportion théorique utilisée pour simuler une variable. Peut prendre la valeur de 1 %, 5 %, 10 %, 15 %, 20 %, 30 %, 40 % ou 50 %.
Y hat	Moyenne de 100 proportions calculées. Ce chiffre devrait s'approcher de la proportion cible.
N	Taille moyenne de l'échantillon du domaine spécifié calculée d'après 100 répétitions.
Bs_var	Moyenne de 100 valeurs de la variance pour le domaine spécifié.
Bs_sd	Moyenne de 100 valeurs de l'erreur-type pour le domaine spécifié.
Cil95	Moyenne de 100 valeurs de la borne inférieure de l'intervalle de confiance à 95 %.
Ciu95	Moyenne de 100 valeurs de la borne supérieure de l'intervalle de confiance à 95 %.

Les normes de qualité qui suivent devraient être utilisées à titre de référence :

1) Une estimation est dite **acceptable** si la taille de l'échantillon est de 10 ou plus et un faible coefficient de variation de l'ordre de 0,0 % à 16,5 %.

2) Une estimation est dite **médiocre** si la taille de l'échantillon est de 10 ou plus et un coefficient de variation élevé de l'ordre de 16,6 % à 33,3 %. Cette estimation devrait être accompagnée d'une mise en garde avertissant les utilisateurs du niveau élevé d'erreur qui y est relié.

3) Une estimation est dite **inacceptable** si la taille de l'échantillon est de 10 ou plus et un coefficient de variation très élevé de plus de 33,3 %. Statistique Canada recommande de ne pas diffuser les estimations de qualité inacceptables (voir la section 14.5).

Pour de plus amples renseignements, consulter la publication intitulée *Lignes directrices sur le niveau de qualité de Statistique Canada*, (N° 12-539-XIF au catalogue).

15.3 Comment calculer le coefficient de variation pour des estimations catégoriques

Règle 1 : Estimation du nombre d'immigrants possédant une caractéristique (agrégats)

Le coefficient de variation dépend uniquement de la taille de l'estimation proprement dite. On peut dire que le CV de l'estimation s'approche (quoiqu'il soit légèrement supérieur) de celle de la proportion qu'il représente. Donc, pour obtenir une approximation du CV d'une estimation, les utilisateurs devraient utiliser le MECV en spécifiant la taille du domaine et en calculant la proportion appropriée. Par exemple, supposons qu'on estime que *Y* hat = *30 000* individus possèdent une caractéristique particulière. Si nous les comparons aux *100 000* personnes comprises dans le domaine d'intérêt, alors le CV de *Y* hat devrait s'approcher d'une proportion c'est-à-dire *30 000 / 100 000 = 30,0 %*. Pour obtenir un CV exact, il faut se servir des programmes qui utilisent les poids bootstrap. Les programmes bootstrap sont disponibles pour les utilisateurs de SAS et STATA.

Règle 2 : Estimations d'une proportion ou d'un pourcentage d'immigrants possédant une caractéristique

Les CV calculés dans le MECV le sont pour des proportions. Donc, ils peuvent être utilisés tels qu'ils sont donnés dans le chiffrier.

Règle 3 : Estimations d'une différence entre agrégats, pourcentages ou rapports

Pour obtenir le CV d'une différence, les programmes bootstrap conviennent le mieux car il n'existe aucun moyen facile de le calculer pour chaque CV. Les programmes permettent de calculer les CV pour les différences entre totaux et entre rapports.

Règle 4 : Estimations de rapports
Si le dénominateur d'un rapport est considéré comme une « taille de domaine », on peut utiliser le MECV de la même façon qu'il l'est expliqué à la règle 2. Sinon, les programmes bootstrap peuvent être utilisés à condition de définir correctement le numérateur et le dénominateur.

15.4 Comment utiliser les coefficients de variation pour calculer des limites de confiance

Bien que les coefficients de variation soient beaucoup utilisés, l'intervalle de confiance d'une estimation est une mesure plus intuitivement significative de l'erreur d'échantillonnage. Un intervalle de confiance constitue une déclaration du niveau de confiance selon laquelle la valeur vraie pour la population se situe à l'intérieur d'une gamme précisée de valeurs. Par exemple, un intervalle de confiance de 95 % peut être décrit comme suit :

> Si l'échantillonnage de la population se répète à l'infini, chacun des échantillons donnant un nouvel intervalle de confiance pour une estimation, alors, dans 95 % des cas, l'intervalle contiendra la valeur vraie de la population.
>
> Une fois déterminée l'erreur-type d'une estimation, on peut calculer des intervalles de confiance pour les estimations en partant de l'hypothèse qu'en procédant à un échantillonnage répété de la population, les diverses estimations obtenues pour une caractéristique de la population sont réparties selon une distribution normale autour de la valeur vraie de la population. Selon cette hypothèse, il y a environ 68 chances sur 100 que l'écart entre une estimation de l'échantillon et la valeur vraie pour la population soit inférieur à une erreur-type, environ 95 chances sur 100 que l'écart soit inférieur à deux l'erreurs-types et environ 99 chances sur 100 que l'écart soit inférieur à trois l'erreurs-

types. On appelle ces différents degrés de confiance des niveaux de confiance.

Les intervalles de confiance pour une estimation, $\hat{X}$, sont généralement exprimés sous forme de deux nombres, l'un étant inférieur à l'estimation et l'autre supérieur à celle-ci, sous la forme $\left(\hat{X} - k,\ \hat{X} + k\right)$, où on détermine k selon le niveau de confiance désiré et l'erreur d'échantillonnage de l'estimation.

Les intervalles de confiance à 95 % d'une estimation peuvent être tirés directement des tables de CV figurant dans le chiffrier. L'utilisateur qui souhaite déterminer d'autres intervalles de confiance doit utiliser la formule qui suit pour convertir le coefficient de variation de l'estimation en un intervalle de confiance $\left(IC_{\hat{x}}\right)$:

$$IC_{\hat{x}} = \left(\hat{X} - t\hat{X}\alpha_{\hat{x}}, \hat{X} + t\hat{X}\alpha_{\hat{x}}\right)$$

Où $\alpha_{\hat{x}}$ est le coefficient de variation déterminé pour $\hat{X}$, et

t = 1 si l'on désire un intervalle de confiance de 68 %;
t = 1,6 si l'on désire un intervalle de confiance de 90 %;
t = 2,6 si l'on désire un intervalle de confiance de 99 %.

Note d'avertissement concernant les intervalles de confiance
Les lignes directrices concernant la diffusion des données s'appliquant également aux intervalles de confiance. Par exemple, si l'estimation est « médiocre », alors l'intervalle de confiance est médiocre et devrait être accompagné d'une mise en garde avertissant les utilisateurs des hauts niveaux d'erreur qui y sont reliés.

Exemple d'utilisation du coefficient de variation pour obtenir les limites de confiance
L'intervalle de confiance à 90 % de la proportion estimée de femmes titulaires d'un diplôme universitaire se calcule comme suit.

$\hat{X}$ = 47,4 % (ou exprimé sous forme de proportion 0,474)

t = 1,6

$\alpha_{\hat{x}}$ = 1,21 % (0,0121 exprimé sous forme de proportion) est le coefficient de variation de cette estimation tel que calculé au moyen des poids bootstrap.

$IC_{\hat{x}}$ = {0,474 - (1,6) (0,474) (0,0121), 0,474 + (1,6) (0,474) (0,0121)}

$IC_{\hat{x}}$ = {0,474 - 0,009, 0,474 + 0,009}

$IC_{\hat{x}}$ = {0,465, 0,483}

Donc, on peut dire avec un degré de confiance de 90 % que de 46,5 % à 48,3 % de femmes possèdent un diplôme universitaire.

15.5 Test d'hypothèse (test t)

On peut aussi utiliser les erreurs-types pour effectuer des tests d'hypothèses, une procédure qui permet de faire la distinction entre les paramètres d'une population à l'aide d'estimations d'un échantillon. Ces estimations peuvent être des nombres, des moyennes, des pourcentages, des rapports, etc. Les tests peuvent être effectués à divers niveaux de signification, à un niveau de signification est la probabilité de conclure que les caractéristiques sont différentes quand, en fait, elles sont identiques.

Supposons que $\hat{X}_1$ et $\hat{X}_2$ sont des estimations d'un échantillon pour deux caractéristiques d'intérêt. L'erreur-type de la différence $\hat{X}_1 - \hat{X}_2$ peut être obtenue en se servant des programmes qui utilisent les poids bootstrap. Représentons l'erreur-type de la différence par $\sigma_{\hat{d}}$.

Si $t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$ se situe entre -2 et 2, aucune conclusion à propos de la différence entre les caractéristiques n'est alors justifiée au niveau de signification de 5 %. Si, cependant, ce rapport est inférieur à -2 ou supérieur à +2, la différence observée est significative au niveau de 0,05. C'est-à-dire que la différence entre les estimations est significative.

15.6 Coefficients de variation d'estimations quantitatives

Il faudrait produire des tables spéciales afin de déterminer l'erreur d'échantillonnage d'estimations quantitatives, ce qui n'a pas été fait car la plupart des variables pour l'ELIC sont principalement de nature catégoriques.

En général cependant, le coefficient de variation d'un total quantitatif sera supérieur au coefficient de variation de l'estimation de la catégorie correspondante (c'est-à-dire l'estimation du nombre de personnes retenues dans l'estimation quantitative). Si l'estimation de la catégorie correspondante ne peut être diffusée, l'estimation quantitative ne pourra pas l'être non plus. Par exemple, le coefficient de variation du nombre total d'heures de cours des femmes qui fréquentent l'université serait plus élevé que celui de la proportion correspondante de femmes fréquentant l'université. Par conséquent, si le coefficient de variation de la proportion n'est pas diffusable, celui de l'estimation quantitative correspondante ne le sera pas non plus.

Pseudo-réplication
On peut calculer, au besoin, les coefficients de variation d'estimations de ce genre pour une estimation particulière au moyen d'une technique appelée pseudo-réplication, qui consiste à diviser les enregistrements des fichiers de microdonnées en sous-groupes (ou sous-échantillons) et à calculer la variabilité de l'estimation d'un sous-échantillon à l'autre. Les utilisateurs désireux de calculer le coefficient de variation d'estimations quantitatives peuvent demander conseil à Statistique Canada en ce qui concerne la manière de répartir les enregistrements en sous-échantillons appropriés et les formules à utiliser pour ces calculs.

15.7 Seuils approximatifs pour la diffusion des estimations

Les tableaux ci-dessous fournissent les seuils de diffusion approximatifs pour deux domaines particuliers. Ces prévisions démographiques fournissent une indication approximative des tailles de domaine acceptables, médiocres et inacceptables. L'utilisateur doit considérer ces seuils comme des lignes directrices approximatives seulement et reste tenu de calculer les CV précis avant de diffuser les résultats. L'utilisation du MECV est fortement recommandée pour obtenir une meilleure précision.

Seuils approximatifs de diffusion selon la catégorie d'immigrants

Catégories d'immigrants	CV acceptable 0,0 à 16,5 %	CV médiocre 16,6 à 33,3 %	CV inacceptable > 33,3 %
Famille	820 et plus	205 à < 820	moins de 205
Économique	560 et plus	165 à < 560	moins de 165
Réfugiés	270 et plus	80 à < 270	moins de 80
Total	**530 et plus**	**145 à < 530**	**moins de 145**

Seuils approximatifs de diffusion selon la région géographique

Provinces	CV acceptable 0,0 à 16,5 %	CV médiocre 16,6 à 33,3 %	CV inacceptable > 33,3 %
Québec	565 et plus	155 à < 550	moins de 155
Ontario	625 et plus	155 à < 625	moins de 155
Alberta	385 et plus	95 à < 385	moins de 95
Colombie-Britannique	485 et plus	175 à < 485	moins de 175
Autre	385 et plus	200 à < 385	moins de 200
Canada	**530 et plus**	**145 à < 530**	**moins de 145**