

## WEIGHTING CHALLENGES FOR THE LONGITUDINAL SURVEY OF IMMIGRANTS TO CANADA (LSIC)

Jean-François Dubois and Michelle Simard<sup>1</sup>

### ABSTRACT

As in other surveys, the Longitudinal Survey of Immigrants to Canada (LSIC) is facing nonresponse. In order to get adequate population estimates, the survey weights have to be corrected by using a nonresponse adjustment. A less biased estimator is obtained if this adjustment is calculated within given classes. This is especially true if the response pattern differs from one class to the next. In most surveys, the adjustment is calculated and then applied to the survey weights to get the final weights. For LSIC, calculating the adjustment will not be as straightforward as there is a significantly higher unresolved rate than most surveys. Unresolved units are units that could not be traced nor contacted during the collection period. We are aware that some groups of immigrants land in Canada but then go to the USA or go back to their original countries after a certain period for various reasons. This leads to think that there are two main reasons why there is unresolved units: 1) the immigrant is no longer in Canada so even the best sources of tracing wouldn't permit to find him; 2) the immigrant is really in Canada but operational constraints prevent us from finding him. The paper presents some new approaches to adjust the responding units based on various models that predict an estimated rate of inscope immigrants for unresolved units. Some methods to create the classes of adjustment will be presented. Different strategies will be discussed and evaluated.

KEY WORDS: Calibration adjustment; Nonresponse adjustment; Response mechanism; Reweighting.

### RÉSUMÉ

Comme dans bien des enquêtes, l'Enquête longitudinale auprès des immigrants du Canada (ELIC) fait face à la non-réponse. Une des façons d'obtenir des estimations adéquates consiste à calculer un facteur d'ajustement de non-réponse. Afin d'obtenir un estimateur moins biaisé, l'ajustement est calculé à l'intérieur de classes données. Ceci est particulièrement pertinent si le mécanisme de réponse diffère d'une classe à l'autre. Dans la plupart des enquêtes, l'ajustement est calculé puis appliqué aux poids de sondage afin d'obtenir les poids finaux. Pour l'ELIC, le calcul de l'ajustement ne sera pas aussi simple puisque que nous observons un taux plus important d'unités non résolues que la plupart des enquêtes. Les unités non résolues sont des unités n'ayant pu être dépistées ou contactées au cours de la période de collecte. Nous sommes conscients que certains groupes d'immigrants arrivent au Canada mais finissent par aller aux É.-U. ou dans leur pays d'origine pour diverses raisons. Ceci porte à penser qu'il existe deux raisons principales expliquant la présence d'unités non-réponses: 1) l'immigrant n'est plus au Canada et donc même les meilleures sources de dépistage ne permettraient pas de le retracer; 2) l'immigrant est effectivement au Canada mais ne peut être retrouvé à cause de contraintes opérationnelles. L'article présente de nouvelles approches pour ajuster les unités répondantes basées sur divers modèles permettant de prédire un taux estimé d'immigrants parmi les unités non résolues. Diverses méthodes permettant de créer les classes d'ajustement seront présentées. Différentes stratégies seront discutées et évaluées.

MOTS CLÉS : Ajustement par calage, ajustement pour la non-réponse, mécanisme de réponse, repondération.

### 1. INTRODUCTION

The Longitudinal Survey of Immigrants to Canada (LSIC) is designed to study how recent immigrants adjust to life in Canada. Recent arrivals will be interviewed three times during their first four years in Canada – six months, two years and four years after their arrival – allowing for the creation of a dynamic picture of their experiences.

With the exception of a pilot test for this survey (conducted in 1996-97 by Statistics Canada for Citizenship and Immigration Canada), this is the first national survey conducted with the recent immigrant population since 1970. Data collection for the first wave of interviews began in April 2001 and was spread over one year.

---

<sup>1</sup> Jean-François Dubois and Michelle Simard, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, [dubojfr@statcan.ca](mailto:dubojfr@statcan.ca), [simamic@statcan.ca](mailto:simamic@statcan.ca).

Recent immigrants represent a highly mobile population and are difficult to trace and subsequently interview. Hence finding the appropriate methodology for weighting is a challenge. Before presenting different options for the weighting strategy we will describe the background of the survey, some results of the first wave of data collection and the challenges of tracing newly arrived immigrants.

## **2. SURVEY BACKGROUND**

### **2.1 Objective and content of the survey**

There exists a growing need for information on recent immigrants to Canada, specifically the integration process, the factors that affect integration and the services used by immigrants to facilitate the process. While full integration may take several generations to achieve, the LSIC is designed to examine the process during the critical first 4 years of settlement, during which newcomers establish economic, social and cultural ties to Canadian society.

Respondents are asked questions on various aspects of their life, ranging from their reasons for choosing to relocate to Canada to problems encountered in finding housing, employment and education. The questionnaire includes a self-assessment of the respondent's abilities in English and French (oral and reading skills). It also collects general information on their children's education and health. Questions pertaining to the ability to access services have been incorporated throughout the questionnaire.

### **2.2 Target population**

The target population for the survey consists of immigrants who meet the following criteria :

- a) landed immigrants arriving from outside Canada;
- b) aged 15 years or older at the time of landing;
- c) date of arrival between October 2000 and September 2001.

Immigrants who apply and land from within Canada are excluded from the survey. These people may have been in Canada for a considerable length of time before officially "landing" and would therefore likely demonstrate quite different integration characteristics to those recently arrived in Canada.

### **2.3 Sampling frame**

The sampling frame is Citizenship and Immigration Canada's Field Operation Support System (FOSS) – an administrative database of all landed immigrants to Canada. The database includes various characteristics of each immigrant that can be used for survey design

purposes, such as the name, age, sex, mother tongue, country of origin, knowledge of English and/or French, class of immigrant, mission of visa issuance, date of landing and intended province of destination in Canada. Each immigrant is assigned a unique visa / landed immigrant number as is each immigrating unit (IU) or group of immigration that land together.

Detailed information from FOSS on each immigrant landing during the survey reference period, i.e., October 2000 to September 2001, is provided to Statistics Canada two months after the reference month. This allows for the sampling frame to be built month after month by simply adding new monthly landings.

### **2.4 Tracing**

Tracing and contacting new immigrants is challenging, as recent immigrants are a highly mobile population. Limited information is available on the sampling frame – immigrants are only required to provide their intended province of destination, which may not necessarily be where the immigrant eventually settles. Traditional administrative sources used by other surveys for tracing such as the T1 and T4 forms of the Canada Customs and Revenue Agency, Child Tax Benefits, Unemployment Insurance or Address Registers are not available for the first wave as new immigrants are not yet included in these files.

The best source of address information are the provincial Ministry of Health Address files as immigrants can apply for a health care card within three months of their arrival. Access to these files is only granted with consent from the immigrant. Consent forms were thus provided to all landed immigrants at the time of Landing Visa issuance overseas and were collected by Immigration Officers at the Canadian Ports of Entry. While the consent from immigrants is high (79 %), the rate of return of the questionnaires has been low and, as a result, health card addresses were obtained for only 40 % of the sample.

Other administrative files such as the Citizenship and Immigration Canada (CIC) address databases and phone files are also used in tracing. Overall, address information could be provided for approximately 75 % of the sample, with an average of 6 addresses provided for each potential respondents.

It is known that some immigrants will land in Canada and then go to USA while others return to their country of origin. Furthermore, current tracing and collection procedures are not adequately developed for Statistics Canada to locate those immigrants who left Canada. This

provides a strong reason for Statistics Canada to be unable to trace a good portion of the sample.

### 3. FIRST WAVE COLLECTION RESULTS

From wave 1 collection results, we were unable to contact 31% of the 20,344 units in the sample. These cases were classified as unresolved. For the resolved cases, the response rate was around 90%. The out-of-scope units represent less than 3% percent of the resolved units. The problems seem to reside more in the difficulty to trace the units than to get an answer once a contact is made.

We have reasons to believe that there is much more out-of-scope units in the unresolved units than in the resolved units. Multiple analysis were made to try to reinforce this assertion such as the one presented in table 1 which presents the sample distribution, as column percentages, by class of immigrant for the total sample and each response status. The response status distributions can be compared to the sample distribution. A simple rule of thumb would be to say that if any response status is to be assumed random, its distribution should be more or less similar to the sample distribution (all categories of status), which is obviously not the case here. The patterns of the unresolved and nonresponding units are different and unresolved units have very similar patterns to those of the resolved out-of-scope units.

From these findings, considering the unresolved units as nonrespondents and applying a simple nonresponse adjustment seems inadequate. The challenges of using a proper weighting scheme will be presented in the next section.

### 4. WEIGHTING CHALLENGES

This section overviews the survey weighting steps, the rationale for the nonresponse weighting adjustments and and, finally, options to perform the nonresponse adjustments.

#### 4.1 Survey Weighting Steps

Like other longitudinal surveys, LSIC objective is to produce a set of final weights for each longitudinal respondent to be able to yield appropriate unbiased estimates of the population of interest. Different weighting adjustments are usually produced to reflect specific characteristics of the design, the propensity to respond and the population of interest. Usually, the final weight consists of three components and is calculated as :  $w_f = a_{nr}a_{ps}w_i$  where  $w_i$  is the sample design weight,  $a_{nr}$  is the nonresponse adjustment to be applied to the design weight and  $a_{ps}$  is the calibration adjustment to be applied to the resulting weight after nonresponse adjustment, if applicable.

#### 4.2 Nonresponse Adjustments in the Case of LSIC

In section 3 we have seen that wave 1 survey results suggest that the nonresponse mechanism is different from the unresolved mechanism and that there is a strong relationship between cases being resolved/unresolved and being in-scope/out-of-scope. A proper weighting adjustment appears crucial.

It is proposed that the nonresponse adjustment will be broken down into two components. First, there will be an adjustment to redistribute the weights of the in-scope nonresponding units to the weights of the in-scope responding units. This will be the purpose of section 4.3. Second, the weights of the resolved units will be adjusted to incorporate the characteristics of the unresolved units. Various options to adjust for the unresolved units will be presented in section 4.4.

#### 4.3 Adjusting for the Nonresponding Units

In order to adjust the weights for the (in-scope) nonresponding units, the weights of these units will be redistributed to inflate the weights of the responding units. This weight redistribution will be performed within classes of adjustment. These classes are introduced as there are indications that the response mechanism is not uniform but rather depends on some variables  $X_1, \dots, X_p$  (such as class

**Table 1 – Distribution (in percentage) of the type of immigrants by response status**

Type of immigrant	All categories of status	Respondents (in-scope)	Nonrespondents (in-scope)	Out-of-scope	Unresolved
Economic	62	58	50	69	72
Family	26	28	41	25	20
Refugee	11	13	8	6	8
Other	1	1	1	0	1
All Types	100	100	100	100	100

of immigrant, province, age group) available for all sampled units.

Several methods can be used to create the adjustment classes, but most methods are based on the creation of Response Homogeneity Groups (RHGs). For example, decision tree algorithms, such as CHAID in the software Knowledge Seeker (Kass, 1980; Angoss Software, 1995) and logistic regression models have been extensively used to create RHGs, while another method recently proposed by Eltinge and Yansaneh (1997) creates classes with similar estimated probability of responding which are predicted values from a logistic regression model. The three methods listed above are being tested and evaluated for LSIC and comparisons of each, in terms of bias, variance and mean square error are planned.

Once the RHGs are formed, we obtain an adjustment weight, say  $a_{nr-g}$ , for each RHG, denoted as  $g$ . Let  $s_g$  be the in-scope sample falling in RHG  $g$ ,  $s_{gr}$  be the responding sample falling in RHG  $g$  and  $w_{li}$  be the initial sampling weight, then we have :

$$a_{nr-g} = \frac{\sum_{i \in s_g} w_{li}}{\sum_{i \in s_{gr}} w_{li}} .$$

Note that if we then sum the product of the design weight by the resulting nonresponding adjustment weight over all the  $s_{gr}$ , we obtain the weighted estimate of the total number of in-scope units coming from the resolved portion of the sample.

#### 4.4 Adjusting for the Unresolved Units

To complete the adjustment process, we need to adjust the weights to take into account the nonresponse component that comes from the unresolved portion of the sample. We believe that there is a certain number of both in-scope units and out-of-scope units coming from the unresolved portion of the sample.

There are six methods proposed to calculate this second adjustment. They are grouped into four types. The first is to use up-to-date information on immigrants (section 4.4.1). The second type is based on models predicting values of in-scope/out-of-scope status (section 4.4.2) while the third type is based on models estimating the propensity of being resolved (section 4.4.3). The fourth type consists of modelling the propensity of being resolved but this time along with modelling either implicitly or explicitly the in-scope/out-of-scope status (section 4.4.4).

#### 4.4.1 Adjusting Based on Up to Date Information

With this type of adjustment, we obtain the in-scope/out-of-scope status, for each immigrant in the survey population from external sources or, for each immigrant in a sub-sample of the unresolved, from a follow-up. We could then use this information to adjust the weights of the in-scope and out-of-scope units in the resolved portion of the sample within some Homogeneity Groups (HG).

##### *Method A : Administrative Data*

External administrative information available on some micro-data files could be used. These files would provide an indication of the in-scope/out-of-scope status for all units of the survey population. For instance, and depending on the file, finding or matching an immigrant could be perceived as a confirmation that he/she is still residing in Canada.

Other than the fact that we can't attain full coverage from one or multiple available files, this approach is not sufficient to resolve all the cases.

##### *Method B : Follow-up*

In this approach, either all or a sub-sample of the unresolved units in the survey sample are followed-up. Re-sending cases in the field for an additional collection period would allow the use of additional sources of tracing information that were not available during the collection period. Furthermore, since the interviewers would have no extensive interview to conduct, all efforts could be put on tracing and resolving the unresolved cases, i.e., the interviewer would only check the in-scope/out-of-scope status.

Again, this approach won't resolve all the cases and budgetary constraints will limit the operations.

#### 4.4.2 Adjusting Based on Predicted In-Scope/Out-of-Scope Status

##### *Method C : Predicting in-scope/out-of-scope status*

This type of adjustment involves the use of auxiliary information predicting the in-scope/out-of-scope status in the formation of HGs for resolved units. Once the HGs are defined, each unresolved unit will then be associated with an HG. Weight adjustments will be calculated using the same approach as for adjusting for the nonresponding units (section 4.3). In this method, the resolved units will take the role of the responding units and the entire sample plays the role of the in-scope sample, i.e., the adjustment within

HG could be shown as the estimated total number of units in HG/estimated number of resolved units in HG.

The adjusted weights will allow for the estimation of the number of in-scope units and the number of out-of-scope units from the sample. The three methods proposed to create RHGs, as described in section 4.3, could be used here in a similar fashion to derive HGs, i.e., the propensity of responding would be replaced by the survey item : in-scope/out-of-scope status.

Results from wave 1 data show that variables such as the knowledge of English or French and the level of education are good predictors of the in-scope/out-of-scope status.

#### 4.4.3 Adjusting Based on Propensity of Being Resolved

*Method D : Predicting propensity of being resolved*

This adjustment involves the use of auxiliary and sampling information capable of predicting the propensity of being resolved in the formation of HGs of resolved. However, with this method we would assume that the response mechanism is not a function of the item in-scope/out-of-scope status, i.e., the mechanism is ignorable.

Results from wave 1 show that tracing information such as the quality and the number of sources of tracing are good predictors of resolved/unresolved status.

#### 4.4.4 Adjusting Based on Estimated Propensity of Being Resolved and Predicted In-Scope/Out-of-Scope Status

This type of adjustment involves the use of auxiliary and sampling information to account for the in-scope/out-of-scope status and the source and quality of tracing data in the creation of HGs of sampled units with respect to estimated probabilities of being resolved. The basic difference from the previous method is that we are accounting for the in-scope and out-of-scope status in the approach. Once the HGs are defined, weight adjustments will be calculated using the same approach as in section 4.3. Two main approaches can be used to produce HGs based on the estimated probabilities of being resolved. One approach is to implicitly account for the in-scope/out-of-scope status (method E) and the other one is to explicitly account for that status (method F).

*Method E : Predicting propensity of being resolved and implicitly accounting for in-scope/out-of-scope status*

Under this approach HGs would be created with similar propensities of being resolved based on the source of tracing information as well as variables known to explain

the in-scope/out-of-scope status. Variables explaining the in-scope/out-of-scope status could be identified using either a decision tree algorithm or a logistic regression model. For instance, these variables could be the by-product of the adjustment processes mentioned in 4.4.2. Once the explanatory variables are identified, tracing information would be added to the model to estimate the probabilities of being resolved. We could use either Kass' CHAID method or Eltinge & Yansaneh method to form HGs with respect to the probability of being resolved. To that end, we would force the classes explaining the in-scope/out-of-scope status to be part of the classes forming the HGs.

*Method F : Predicting propensity of being resolved and explicitly accounting for in-scope/out-of-scope status*

Under this approach HGs would be formed based on auxiliary and sampling information as well as tracing information. However, the application of this method would not be as straightforward as it is a variant of an approach developed to estimate or impute in the presence of a non-ignorable response mechanism. Little (1982) provided a formal definition of the concept of non-ignorable nonresponse. A particular case of non-ignorable response mechanism is : "when a survey response mechanism depends on a variable of interest measured within the same survey and observed for only a part of the sample". In most cases, the variable of interest related to the response mechanism is one of the variables of interest of the survey and is a continuous variable. In our particular application, the variable of interest is dichotomous, i.e., being in-scope or out-of-scope, and the response mechanism depending on this variable is the resolution one. Little proposes to use the maximum likelihood method to jointly estimate the parameters of the models predicting the propensity of responding and the variable of interest Y in this particular situation. Greenlees, Reece and Zieschang (1982) also use the same approach to estimate model parameters but with the goal to impute the variable of interest Y. Even if resulting probabilities of responding from the models are calculated for these approaches, they are not used explicitly.

The method F consists in specifying two models. One model would explain the probability of being resolved while the second model would explain the probability of being in-scope. The parameters of these two models would be jointly estimated. The estimation of the parameters would be done using either the maximum likelihood or the robust method proposed by Beaumont (2000). This would then allow for the calculation and the use of the estimated probabilities of being resolved required to form HGs with the Eltinge & Yansaneh method.

## 5. SUMMARY AND RESULTS FROM LSIC

The Longitudinal Survey of Immigrants to Canada is facing a high rate of unresolved units. We have reasons to believe that there's a lot of unresolved units that are really out-of-scope units. Survey results suggest that the nonresponse mechanism is different from the unresolved mechanism and that there is a strong relationship between cases being resolved or unresolved and being in-scope or out-of-scope. This indicates that the calculation of a proper weight adjustment is crucial. The most challenging issue is then to adjust not only for the nonresponding units but also to adjust the weights of the resolved units to take into account the characteristics of the unresolved units.

We propose six methods designed to calculate this adjustment. The first two are based on up to date information but are insufficient to give complete in-scope/out-of-scope status. The other four methods are based on models. The first three are appropriate for ignorable response mechanism and are based on various assumptions as to what the response mechanism depends on while the last method is suitable for non-ignorable response mechanism.

Up to now we performed the joint estimation of parameters under the assumptions of both ignorable and non-ignorable response mechanisms. The likelihood ratio test seems to indicate that the mechanism is ignorable. We have yet to decide which method we will use but at this point we are leaning towards the option of implicitly

accounting for the in-scope/out-of-scope status (method E).

## REFERENCES

- Angoss Software (1995), *Knowledge Seeker IV for Windows – User's Guide*. Angoss Software International Limited.
- Beaumont, J.-F. (2000), An estimation method for non-ignorable nonresponse. *Survey Methodology*, December 2000, Vol. 26, No. 2, pp. 131-136.
- Eltinge, J.L. and Yansaneh, I.S. (1997), Diagnostics for formation of nonresponse adjustment cells with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, June 1997, Vol. 23, No. 1, pp. 33-40.
- Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.
- Kass, G. V. (1980), An explanatory technique for investigating large quantities of categorical data. *Applied Statistics*, Vol. 29, No.2, pp.119-127.
- Little, R.J.A. (1982), Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, Vol. 77, pp. 237-250.