Imputation Strategies for the Longitudinal Survey of Immigrants to Canada Wave 2 Imputation and Evaluation of the Wave 1 Imputation

Asma Alavi and Owen Phillips
Statistics Canada, 2500 A– Main Bldg., Tunney's Pasture, 120 Parkdale Ave.
Ottawa, Ontario, Canada, K1A 0T6
asma.alavi@statcan.ca

Abstract

The Longitudinal Survey of Immigrants to Canada (LSIC) is designed to understand the process of adaptation to Canadian Society by recent immigrants, and to recognize the factors that aid or impede the immigrants' efforts in doing so.

As with most surveys, the LSIC faces non-response either full or partial. In the LSIC, partial non-response is dealt with through imputation. The idea is to impute data that are consistent and that would not distort the distribution of the variables being imputed. Being a longitudinal survey, the question of consistency across waves is of even greater importance for the LSIC.

In this paper, imputation strategies for Waves 1 and 2 of the LSIC are documented. The similarities and differences in the imputation techniques between the two waves are illustrated. A descriptive study to evaluate the quality of the Wave 1 imputation, based on the sampled units that were imputed in Wave 1 but had complete response in Wave 2, is also outlined.

Keywords: Longitudinal Surveys; Nearest-neighbour Imputation; Model-assisted Imputation, Logistic Regression, Monotonic Design

1. LSIC Background

The LSIC is a comprehensive longitudinal survey conducted jointly by Statistics Canada and Citizenship and Immigration Canada under the Policy Research Initiative. The target population is all immigrants aged 15 years and older, landing from abroad between October 2000 and September 2001. In the LSIC the sampled units, called the Longitudinal Respondents (LRs), were interviewed at three points in time. The first interview was conducted six months (Wave 1) after their arrival in Canada; the second interview was two years (Wave 2) after their arrival, with the last interview conducted four years (Wave 3) after their arrival. The sampling frame for the LSIC is an administrative database called Field Operation Support System (FOSS). The FOSS is a database of all landed immigrants to Canada which is maintained by Citizenship and Immigration Canada. Of the nearly 250,000 immigrants landed between October 2000 and September 2001, almost 68% belonged to the LSIC target population. The other 32% immigrants were either children or were landed from inside Canada.

The LSIC has a stratified two-stage sampling design. The stratification for the LSIC sampling was based on the intended province of residence, the class of immigration and the month of landing. Table 1 presents the sample allocation for the LSIC based on province and class of immigration.

The data collection vehicle for the LSIC is a detailed questionnaire that is comprised of 10 different modules. A module is a set of questions related to a specific topic such as health or income. Both Computer Assisted Personal Interviews (CAPI) and Computer Assisted Telephone Interviews (CATI) were employed to gather the data. The emphasis was on in-person interviews whenever possible. There were 15 interview languages including English and French. With the exception of the module on income, in which the person most knowledgeable about the subject was asked to respond, no interview was conducted by proxy.

The LSIC has a monotonic or "funnel-shaped" design which implies that only the respondents at a given wave would be traced and interviewed at a later wave. The result of this constraint is the reduction over time in the initial sample size of 20,322, as can be seen from Table 2. This approach was adopted as to get a complete profile for a given LR over time.

Table 1: LSIC Sample Allocation

Province	Class of Immigration						
	Family	Economic-Skilled	Economic-business	Government-refugee	Other Refugee	Other	Total
Quebec	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
British Columbia	1,560	1,634	423	210	40	26	3,893
Other Provinces	121	225	81	293	46	72	838
Canada	5,328	10,937	1,633	1,744	525	155	20,322

Table 2: Wave 1 and Wave 2 Sample Size and Number of Respondents

	Wave 1	Wave 2
Sample Size	20,322	12,040
Respondents	12,040	9,322

2. Imputation for the LSIC - Wave 1

After the collection, the record completion codes are important in defining the response status of the LRs. The record completion codes in turn depend upon the module completion status. In the LSIC, module completion status is defined by the use of *keyfields*. The *keyfields* are a set of certain questions from selected modules of the questionnaire. These questions determine the flow of the questionnaire and are asked to everyone in the sample. If all the *keyfields* in a given module have been answered satisfactorily then the module is deemed complete, and incomplete otherwise. A "don't know" answer, a refusal to answer, or missing information constitute unsatisfactory answers for the *keyfields* questions. If a given LR has all selected modules complete then the record is considered as a complete response. Else if at least one of the modules is complete then the record is deemed a partial response. A record with all modules incomplete would be dropped from the sample file and the remaining records would be re-weighted.

The partial responses are identified and imputed module-wise in the LSIC. This approach is employed for its desirable characteristics. For example, relatively few records are identified as partial respondents when compared to methods based on individual questions. Additionally, imputation of large chunks of data is done quickly thus cutting down the processing time. The donor data replaces the recipient data in the incomplete modules, while keeping the recipient data in complete modules.

The modules for which imputation was done were Social Interaction (SI), Language Skills (LS), Housing (HS), Education (ED), Employment (EM), Health (HL), and Income (IN). No imputation was done for the three modules - Values and Attitudes (VA), Citizenship (CI) and Perceptions of Settlement (PS). The questions in these three modules asked about the LR's opinions and perceptions, which varied too much to be suitable for imputation. Hence all the facts and not the perceptions modules were imputed. Table 3 represents the module completion as a percentage of Wave 2 respondents. It is evident that the modules have very high completion rates individually. The lowest completion rate is for the Income module, which is not surprising with income being a sensitive subject.

There were various rosters associated with some modules. A roster is a data file with as many records, for a given LR, as the number of events for a certain attribute of interest. For example, in the employment history roster each LR would have as many records as the number of jobs reported by him/her. In the LSIC there were the education and training roster associated with ED module, employment history roster with the EM module, and living arrangement history roster with the HS module. The rule of imputation for a roster is: if the related module is imputed then the corresponding roster is automatically imputed.

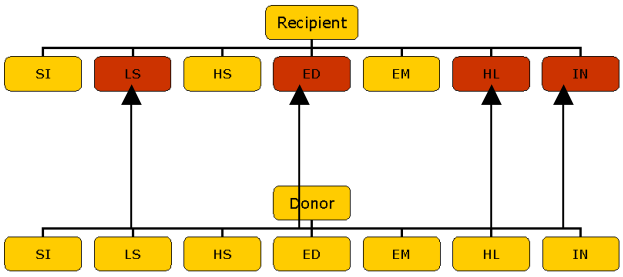
Table 3: Completion Rates based on Wave 2 Respondents (in percentage)

Module	Wave 1	Wave 2	Longitudinal
SI	99.80	99.58	99.38
LS	99.83	99.72	99.55
HS	99.76	99.69	99.46
ED	99.74	99.41	99.17
HL	99.60	98.74	98.37
EM	99.70	99.25	98.96
IN	96.65	97.92	94.74
Overall	96.38	97.18	93.80

Another aspect of imputation in Wave 1 was the adjustment of different date variables from the rosters. The dates were first imputed using the donor record and then adjusted so that the dates were consistent with the recipient's landing and interview dates. The adjustment was done based on the interview dates.

Figure 1 illustrates schematically the imputation mechanism in Wave 1. The modules in red are the incomplete modules of a partial respondent. The donor only replaces the data in the incomplete modules preserving the recipient's original data in complete modules.

Figure 1: Imputation - Wave 1



There were 12,040 respondents in Wave 1: 497 were partial respondents. Imputation was done by employing the nearest-neighbour imputation method. This imputation method generally would not alter the distribution of the data, which is a drawback of many other imputation techniques. A score function was developed based on certain variables, available for both complete and partial respondents, along with the estimated probability of response status from a logistic regression model. A complete respondent with the

highest score, among all the potential donors, was chosen as the donor. In case of multiple donors, a donor was selected at random.

3. Imputation for the LSIC - Wave 2

For Wave 2, the imputation strategy of Wave 1 could have been repeated. But, by doing so, longitudinal inconsistencies could have been introduced. These inconsistencies would have arisen for a couple of reasons: either a given LR could be complete in one wave and partial in the other; or, for a partial LR in both waves, a different donor would most likely be chosen by independent imputation. These inconsistencies were of particular concern when imputing roster data as they were used in derivation of other variables.

In order to overcome those limitations and to save potential processing time, a longitudinal imputation technique was established. The imputation at Wave 2 was longitudinal in the sense that it was done simultaneously for data collected at both waves. As a result, data imputed for partial respondents on the Wave 1 file might differ from Wave 1 imputed data for the same partial respondent on the Wave 2 file.

The first task in the longitudinal imputation was to identify which modules had to be imputed longitudinally. For this purpose longitudinal completion codes were generated. Based on Wave 1 and Wave 2 completion codes, a longitudinal response code was established. A Wave 2 LR was deemed as a longitudinal complete respondent if and only if the LR was a complete respondent in both waves. Otherwise the LR was considered as a longitudinal partial respondent. A consequence of this rule was the classification of a module as longitudinally incomplete if the module was incomplete in either wave. Thus in instances where a module was complete in one wave but not in the other, legitimate data for the particular module were overwritten for one wave. Fortunately there was a small number (552 out of 9,322) of LRs for whom that was an issue. Table 4 shows breakdown of Wave 2 responding LRs according to their completion status in both waves.

Table 4: Response Status of the LRs after Wave 2

Wave 1	Wave 2	Total
Complete response	Complete response	8,744
Complete response	Partial response	241
Partial response	Complete response	313
Partial response	Partial response	24
Total		9,322

For Wave 2, imputation for the incomplete modules was carried out using the longitudinal nearest-neighbour donor technique. The first step in that technique was to develop a logistic regression model of the longitudinal response status variable (1 if longitudinal complete response, 0 otherwise) of the LRs on various auxiliary variables. The auxiliary variables were available for both complete and partial LRs and are listed in Table A of the Appendix.

The logistic regression model resulted in a maximum-rescaled coefficient of determination (see Nagelkerke, 1991) $R^2 = 0.2817$ and a *P-value* of 0.8507 from the Hosmer-Lemeshow goodness-of-fit test. The coefficient of determination is not very high implying that the fitted model does not explain the variation in the data well and hence does not have high predictive ability. But it is important to keep in mind that the dichotomous response status variable has very few zeros thus resulting in unstable modelling. On the other hand, the high *P-value* of the goodness-of-fit test suggests a good fit implying the chosen model agrees with the data. As a result of this modelling we obtained the estimated probability of response (longitudinal complete/longitudinal partial) for all the responding LRs. This probability was then used in finding a suitable donor.

Next, based on a score function, using selected socio-demographic information and the estimated response probabilities, a donor (longitudinal complete respondent) determined to be the closest to the recipient (longitudinal partial respondent) was identified. There were 8,552 complete and edited records used as donors for 578 partial LRs. The maximum possible score, which could be achieved by potential donors, was 57. The variables that were considered for the score function are given in Table B of the Appendix. It is worth noting that the socio-demographic variables used in donor selection included the variables that determined the questionnaire skip-pattern: the presence of LR's spouse and children, and also the presence of LR's school-age children. This information, while not deemed related to the propensity of response, assured consistency between modules on recipient records. For example, a recipient with no children would not receive information for an incomplete module from a donor with children.

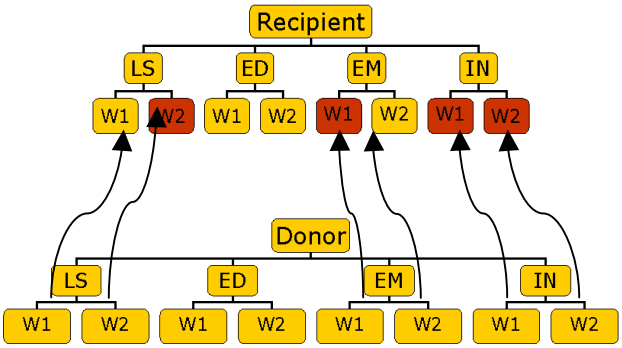
For a longitudinal partial respondent for whom more than one module was incomplete, the same donor record was used for all the incomplete modules for both waves. As mentioned earlier, to keep consistency between variables, the complete set of variables for a given module of the donor was imputed into the recipient record. Figure 2 illustrates the longitudinal imputation mechanism. At the end of this process, all records had fully completed modules. A flag indicating whether a module was imputed was created.

The adjustment of imputed dates in various rosters in Wave 1 resulted in dates that belonged neither to the donor nor to the recipient. There were also some other drawbacks of adjusting the imputed dates. Since the adjustment was done with respect to the interview date only, there was potential of some imputed dates to be earlier than the landing date of the recipient. For instance, consider a donor who landed four months earlier than the

recipient and who started work within a month of landing. In this situation the recipient would have EM roster dates earlier than the landing date. Also, adjusting the dates in that fashion altered the distribution of certain derived variables e.g. time elapsed from landing to the first job. Also by adjusting imputed dates there was a potential for the recipient LR to have seasonal employment in the wrong season, e.g. snow removal in July!

After considering the deficiencies of adjusting imputed dates in Wave 1, it was not repeated for Wave 2 longitudinal imputation. The imputed dates of the recipient were simply the donor dates. Additional information in recipient's record were provided including the donor's interview dates for both waves, landing date and the number of days between landing date and interview dates. Providing that extra information on the microdata files would help analysts to do various comparisons. It should be kept in mind that for recipients the imputed data corresponds to the time frame of the donors. One simply should not compare the imputed data of the recipients with their actual data. The imputed data paints a picture for the recipient as though they had arrived and were interviewed at the same time as their donors.

Figure 2: Longitudinal Imputation – Wave 2



4. Evaluation of Wave 1 Imputation Strategy

After Wave 2, the data were available for the two waves of the LSIC which provided opportunity to evaluate the Wave 1 imputation strategy. An evaluation study was carried out by comparing the data for the LRs that had complete responses in Wave 2 but were imputed in Wave 1. As can be seen from Table 4, there were 313 such LRs.

For the purpose of comparisons, in each module a list of questions was made that were common in both waves. Since for a given module this list could consist of a large number of questions, the attention was focused on the *keyfields* in both waves and then on the choice of the common *keyfields*. By doing that an enormous number of questions was reduced to a more manageable one: 45 questions in seven modules. All of these chosen questions had answers that were categorical in nature.

In order to evaluate the quality of the imputed data, a comparison of imputed and non-imputed data was needed. Since an inherent property of the questions being asked in the LSIC was the possibility of change in answers over time, the choice of comparison was not trivial. A simple comparison of the answers in two waves would not have provided a useful insight. Hence it was decided to do a chi-square test of homogeneity between two mutually-exclusive groups of LRs. The first group consisted of the LRs that had complete response in both waves, while the second group consisted of the LRs that were complete in Wave 2 but were imputed in Wave 1. Each of these two groups was then divided into two classes. The first class was of the LRs having the same answer in both waves for the chosen question, and the other with LRs having different answers in two waves.

For example, a selected question in EM module was about volunteer work done by the LR since arrival in Canada. The chi-square test-statistic was 0.6309 at 1 degree of freedom and the *P-value* of the test was 0.4270. The result indicated that the pattern of changes in the LR answers when imputed in Wave 1 was similar to that of the LRs that were complete in both waves. It was re-assuring that the imputation techniques used in Wave 1 did not change the distribution of changes over time.

Table 5: Distribution of Chosen Questions and Chi-Square Significance

Module	Number of Questions Chosen	Chi-Square Test was not Significant
SI	4	2
LS	10	3
HS	4	4
ED	3	2
EM	11	8
HL	8	3
IN	5	1
Total	45	23

From Table 5, it can be deduced that about 50% of the time, the chi-square test of homogeneity was not significant implying that the imputed and non-imputed groups of LRs were not different with respect to the change in answers between two waves. The qualitative analysis based on the selected common questions was preliminary and crude at best. The results did not provide a sweeping conclusion for Wave 1 imputation techniques to be perfect. It is also to be kept in mind that the size of the group with LRs imputed in Wave 1 but complete in Wave 2, was rather small with only 313 LRs. This small sample size implied even smaller size within modules. So in 15 cases the chi-square test of homogeneity was not the best choice as the expected size for some cells in the 2×2 table was less than 5. When those cases were

discarded, the percentage of the non significant tests remained at 50%.

5. Conclusion

In this paper, longitudinal imputation strategies for the Wave 2 of the LSIC were documented in detail. A brief summary of the LSIC methodology and Wave 1 imputation techniques were presented. A small study to evaluate the Wave 1 imputed data as compared to the real data was also discussed.

In the context of longitudinal nature of the LSIC, the term longitudinal imputation was used and explained. The advantages and shortcomings of imputation strategies at both waves of the LSIC were noted.

Acknowledgments

The authors would like to thank Jean-François Beaumont, Dominic Grenier, and Amélie Lévesque, of Statistics Canada, for their valuable comments and suggestions.

References

Hosmer, D. W., Lemeshow, S. (1980), "A Goodness-of-Fit Test for the Multiple Logistic regression Model", Communications in Statistics, A10, 1043-1069.

Hosmer, D. W., Lemeshow, S. (1989), "Applied Logistics Regression", Wiley Series in Probability and Mathematics.

Kalton, G., and D. Kasprzyk. (1982), "Imputing for Missing Survey Responses", 1982 Proceedings of the Section for Survey Research Methods, American Statistical Association, 22-33

Lafortune, Y. (2002), "Diagnostic Analysis of Logistic Regression Models With survey Data", SSMD-2002 006E, Methodology Branch Working Paper, Statistics Canada.

Nagelkerke, N.J.D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691 -692.

Statistics Canada (2005), "Longitudinal Survey of Immigrants to Canada – A Portrait of Early Settlement Experiences", Catalogue no. 89-614-XIE.

Statistics Canada (2003), "Microdata User Guide – Longitudinal Survey of Immigrants to Canada Wave 1".

Statistics Canada (2003), "Survey Methods and Practices", Catalogue no. 12-587-XPE.

Appendix

Table A: Variables used in the Logistic Regression

Description
interview language - Wave 2
province of residence of LR - Wave 2
method of interview-Wave 1
marital status of LR - Wave 2
class of immigration of LR - FOSS
whether paid or not for a job - Wave 2
Count of LR's children - Wave 2
size of LR's economic family - Wave 2
LR's job count - Wave 2
method of interview - Wave 2
interaction of province and interviewing method - Wave 2

Table B: Variables used in the Score Function

Description
gender of the LR
LR age grouping - Wave 2
language spoken most often at home - Wave 2
presence of LR's spouse in the household indicator - Wave 2
marital status of LR - Wave 2
estimated response probability from logistic model
number of potential earners in the LR's economic family – Wave 2
size of LR's economic family - Wave 2
job count for LR - Wave 2
whether paid or not for a job - Wave 2
selection of a child from the household indicator - Wave 2
indicator if ED questions were asked for selected child – Wave 2
class of immigration of LR - FOSS