Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Using Paradata to Manage Nonresponse in the Survey of Labour and Income Dynamics

by Wisner Jocelyn, Owen Phillips, Beatrice Baribeau and Amélie Lévesque

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

Using Paradata to Manage Nonresponse in the Survey of Labour and Income Dynamics

Wisner Jocelyn, Owen Phillips, Beatrice Baribeau and Amélie Lévesque[1]

Abstract

In recent years, the use of paradata has become increasingly important to the management of collection activities at Statistics Canada. Particular attention has been paid to social surveys conducted over the phone, like the Survey of Labour and Income Dynamics (SLID). For recent SLID data collections, the number of call attempts was capped at 40 calls. Investigations of the SLID Blaise Transaction History (BTH) files were undertaken to assess the impact of the cap on calls.The purpose of the first study was to inform decisions as to the capping of call attempts, the second study focused on the nature of nonresponse given the limit of 40 attempts.

The use of paradata as auxiliary information for studying and accounting for survey nonresponse was also examined. Nonresponse adjustment models using different paradata variables gathered at the collection stage were compared to the current models based on available auxiliary information from the Labour Force Survey.

Key Words: Management, Nonresponse, Paradata.

1. Introduction

In recent years, the use of collection process data, or paradata[2], has become increasingly important to the management of collection activities at Statistics Canada. To satisfy the need for timely and appropriate information on collection costs and quality, Statistics Canada has developed a set of processes and tools, called Active Collection Management (ACM) (Hunter and Carbonneau, 2005; Laflamme, Maydan and Miller, 2008). Particular attention has been paid to social surveys conducted over the phone, like the Survey of Labour and Income Dynamics (SLID).

The main objective of this paper is to illustrate the use of paradata in SLID and explore the possibility of increasing it, particularly in the context of nonresponse adjustment.

The paper is organized as follows: Section 2 provides an introduction to SLID and some of the concepts to be used throughout the paper, including ACM as it applies to SLID. The cap-on-calls studies and results are summarized in section 3. In section 4, the non-response adjustment models using paradata are described and preliminary results reported. Finally, concluding remarks and future work are reported in section 5.

2. Introduction to SLID

SLID is a longitudinal survey used to measure changes in the economic wellbeing of Canadians and the factors that may influence these changes. Specifically, it follows a longitudinal cohort for six years, collecting data about work, income and family circumstances annually. A panel or cohort consists of a sub-sample of approximately 17,000 respondent households from the Labour Force Survey (LFS). A new panel is introduced every three years.

[1]Wisner Jocelyn, Owen Phillips, Beatrice Baribeau and Amélie Lévesque, Household Survey Methods Division, Statistics Canada, Ottawa (Ontario), Canada K1A 0T6
[2] For a more complete definition of paradata, please see Scheuren (2005).

All members of the LFS household are followed longitudinally. As the survey is also used to produce cross-sectional estimates of family and personal income, non-longitudinal cohabitants are interviewed. Interviews are conducted by means of a computer-assisted telephone interview (CATI). While demographic information is collected for all household members, only individuals aged 16 years and older are asked the questions on labour and income. All questions refer to activity in the calendar year (the reference year) preceding data collection. To reduce the response burden, proxy response is allowed and respondents can authorize Statistics Canada to use information from their income tax return in lieu of responding to the income questions.

Response rates are calculated annually. The cross-sectional response rate is defined at the household level where a respondent household is a household in which at least one member responded to either the labour or the income interview, based on completing a minimum portion of the questionnaire. The longitudinal response rate on the other hand is defined at the individual level and is based only on the longitudinal sample, despite the availability of longitudinal information for some cohabitants.

2.1 Active Collection Management of the SLID

The SLID team has established an Active Collection Management group. The group is made up of representatives from the divisions responsible for data collection, the CATI application development, subject matter development and data processing, and the survey methodology. In pre- and post-collection, the group investigates improvements that might be implemented in the next collection and sees that proposed changes are integrated and tested in order to minimize the impact of potential problems.

Among other things, the ACM group is responsible for identifying, implementing and evaluating the effectiveness of new tools that have been introduced to better manage data collection in the SLID CATI environment. Some examples of these initiatives are:

- *Time slices* represent a partitioning of the week based on time of day and the day of week. Contact attempts are allocated to time slices in order to optimize the probability of contacting a particular household. The time slices are defined by head office prior to collection, but may be modified during collection. *Time slice sets* are defined based on the demographic characteristics of a household. They are used to determine the allocation of call attempts to the time slices.
- *Z-groups* are used in an attempt to ensure that adequate response rates are obtained for particular domains of interest to the survey. If the response rate for a particular domain is deemed low the corresponding z-group will be activated thus flagging any household with the characteristic of interest as a contact priority.
- The *call scheduler* is part of the CATI management system. It works in the background and assigns cases based on the time slice, the time slice allocation, z-group and other criteria (e.g. scheduled interview).

The goal of these tools is to make the most effective use of call attempts. Efficient management of collection activities has become increasingly important, particularly in light of the cap-on-calls.

3. Cap-on-calls

In an effort to diminish respondent burden and reduce collection costs, a maximum number of call attempts for household surveys was implemented in 2007. This maximum was set to 25 in general and 40 for SLID in particular. In the summer of 2007, two investigations of the SLID Blaise Transaction History (BTH) files were undertaken to assess the impact of the cap-on-calls. Data for the three most recent survey years were used.

The first study (Lévesque and Poulin, 2007) was conducted under the guidance of the ACM group. Its main purpose was to assess the impact of capping the number of call attempts at 25. It found that had a cap of 25 been implemented for 2007, the cross-sectional response rate would have been just over 73%, instead of the 77% achieved with a cap of 40. Additionally, the percentage of longitudinal respondents lost over the three years might have reached 21% had a cap of 25 been implemented in 2005, and 17% had the cap of 40 been in place beginning in 2005. Given that SLID allows episodes of cycle nonresponse, where typically the high-effort households are not the same from one year to the next, these estimates present a worst-case scenario that might not reflect reality. That

being said, it is not unreasonable to believe that the additional contact effort made in one year facilitates contact in subsequent years.

High-effort households also exhibited demographic and socio-economic differences from low-effort households, introducing the potential of bias in the estimates of certain domains. Not surprisingly, the proportion of households requiring more than 25 attempts is higher in households composed uniquely of young single adults. Other domains exhibiting this same tendency are immigrant households, households in the Western Provinces of Canada, and households of size five or greater. Additionally, income differences were noted in households of size one: the median income of high-effort households was more than $7,000 higher than low-effort households.

The study also examined the demographic characteristics of high-effort households, contact rates and call distributions by time of day and call sequence all in an effort to establish how to make most effective use of call attempts. It found that:

- The probability of contact was highest in the evening, yet more attempts were made during the day. This phenomenon has been observed for other CATI surveys as well (Laflamme, 2008).
- In 2007, a large portion (68%) of the capped cases was finalized by the two-thirds mark of the collection period.
- Some nonrespondents that reached the cap had not had attempts in all of the time slices. In looking at these cases, it was noted that there was a higher incidence of busy dials than for other cases.
- Only 55% of nonrespondents had at least one contact attempt in all time slices.
- *Browser* use is higher amongst nonrespondent cases and in regions where nonresponse rates are higher. The browser allows the interview to chose his/her cases, thus circumventing the case delivery by the call scheduler.

The final report made a number of suggestions for improving the management of data collection activities for 2008. Among others, it was recommended that SLID

- Maintain a cap of 40 on the number of calls for 2008 collection and transfer cases that reach 32-35 attempts to a special group in order to spread the remaining attempts over the remainder of the collection period;
- Increase the number of calls made in the evening and on the weekend;
- Define time slice sets in such a way that cases receive an attempt in all time slices;
- While in many cases the use of the browser can be justified, establish better guidelines for its use.

The second study (Chapman, 2007a and 2007b) focused on the nature of nonresponse given the limit of 40 attempts. Collection results were evaluated for the full sample and at the panel level. The 2007 data collection saw for both panels (4 and 5) a decline in response, contact and refusal rates. The latter follows the trend observed over the life of these and other panels, however the decline in response[3] and contact rates do not respect the increases generally observed in earlier panels. While the increase in nonresponse in 2007 cannot be completely attributed to the limit on call attempts, roughly 4% of all households were nonrespondent by virtue of having reached the cap.

The next section will re-examine some of these findings and suggested changes in light of the 2008 collection results.

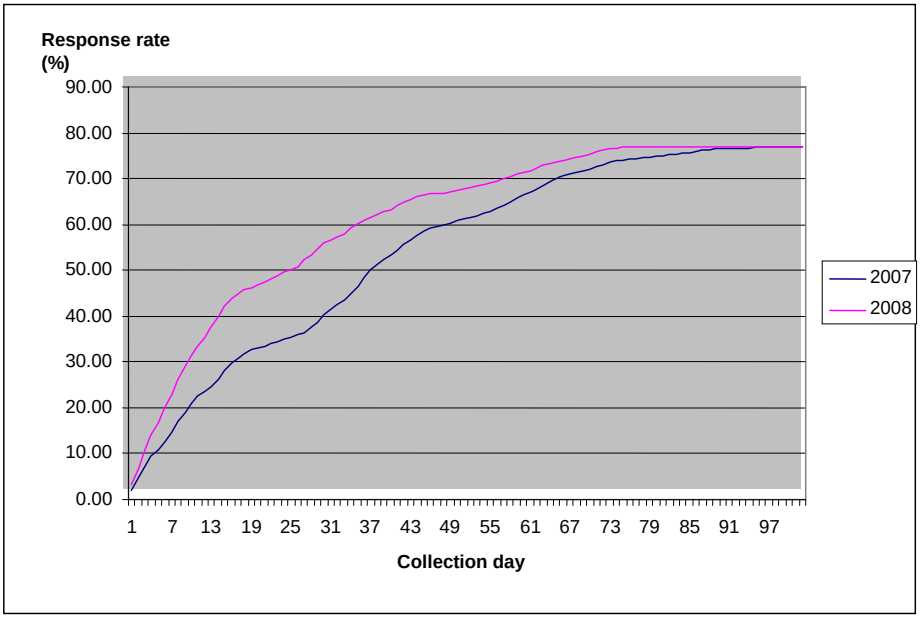
3.1 Collection results for 2008

For operational and budgetary reasons, not all of the suggested changes were introduced for 2008 collection: notably, improvements to the time slice sets to spread call attempts over the entire collection period and to ensure attempts in all time slices were not implemented. However, early collection results showed a vast improvement in response rates over 2007 (Figure 3.1-1). In general, the number of attempts made in the evenings and on the weekends increased, and browser use decreased considerably to similar levels across the regions. However, the percentage of cases requiring more than 25 call attempts increased, and, while the percentage of capped cases not having attempts in all time slices decreased (2.7% vs. 5.3% in 2007), the percentage of cases that reached the cap

[3] Some households are not returned to the field due to certain patterns of non-response in consecutive years, for example, two consecutive years of refusal. Thus the sample file is “cleaned” over the life of a panel, and collection response rates generally improve. This is referred as the *two-year rule*.

increased, and 54% of them were capped with two weeks of collection remaining. Furthermore, the average number of calls to establish a first contact increased in all regions. As collection progressed, the gap between 2008 and 2007 narrowed and at the end of collection[4] 2008 showed only a marginal increase in the final response rate. Figure 3.1-2 displayed on the next page shows the breakdown of non-response as a percentage of the entire sample, for the years 2003 to 2008. As a new panel is introduced every three years, it is useful to look at nonresponse in three-year cycles. In Figure 3.1-2, the cap reached category exists only for the years 2007 and 2008 and as a result, comparisons across years are distorted. This is because the cap reached category consists of both potential respondents and nonrespondents.

Figure 3.1-1
2008 versus 2007 response



To make things comparable and provide some insight into the nonresponse generated by the cap on calls, a cap of 40 call attempts in years 2003 to 2006 is simulated and presented in Figure 3.1-3. The greatest impact of the cap seems to be in the no contact group. In fact, for years 2003 to 2006, households that required more than 40 call attempts consisted mostly of respondent and no contact cases, and over the 4 years, the percentage of respondent cases decreased consistently but was offset by an increase in the percentage of no contact cases.

Looking at Figure 3.1-3, the percentage of refusals has decreased over a three-year cycle. This is perhaps not that surprising: there is a removal of some refusal households from data collection (see the footnote about the *two-year rule* on the previous page). Refusal was higher in 2008 than in 2005, which may indicate that the cap-on-calls has limited the opportunity to convert refusals, however this relationship is not observed for 2007 and 2004. Additionally, the general decrease in refusals coupled with the increase in no contact and other nonresponse[5] households might also point to a movement away from refusal toward refusal by avoidance.

We have also witnessed an increase in the percentage of out-of-scope (out-of-scope is not a nonresponse type) cases. It is not clear if this is indicative of a demographic phenomenon or simply improved tracing of out-of-scope households.

[4] 2008 collection ended after 80 days; the 2007 collection period was extended to 102 days.
[5] Other nonresponse includes missed appointments, household overlaps with other surveys, households where an interview could not be conducted because of a language barrier, unusual or special circumstances or where the longitudinal respondent(s) were absent for the duration of the survey.

Figure 3.1-3 also shows a reversal of trend over the last three years for tracing (or unresolved) cases. This can be explained in large part by a modification of the two-year rule in 2007 and 2008, which saw households that were in tracing for two consecutive years returned to collection. This change in procedure, which yielded more refusals than respondents and made little improvement to the response rates, will be revisited for 2009 collection. Figure 3.1-4 shows that in the absence of the additional cases, the percentage of tracing cases remains relatively stable over the last three years, with 2008 showing a slight improvement, perhaps due to improved tracing information

Figure 3.1-2
SLID nonresponse by type: 2003 to 2008

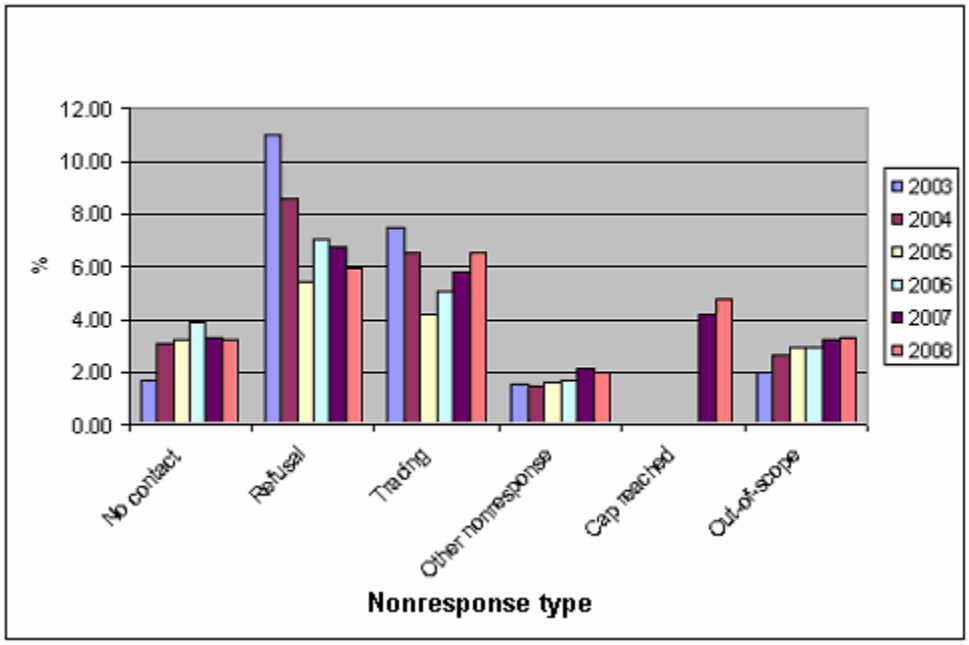
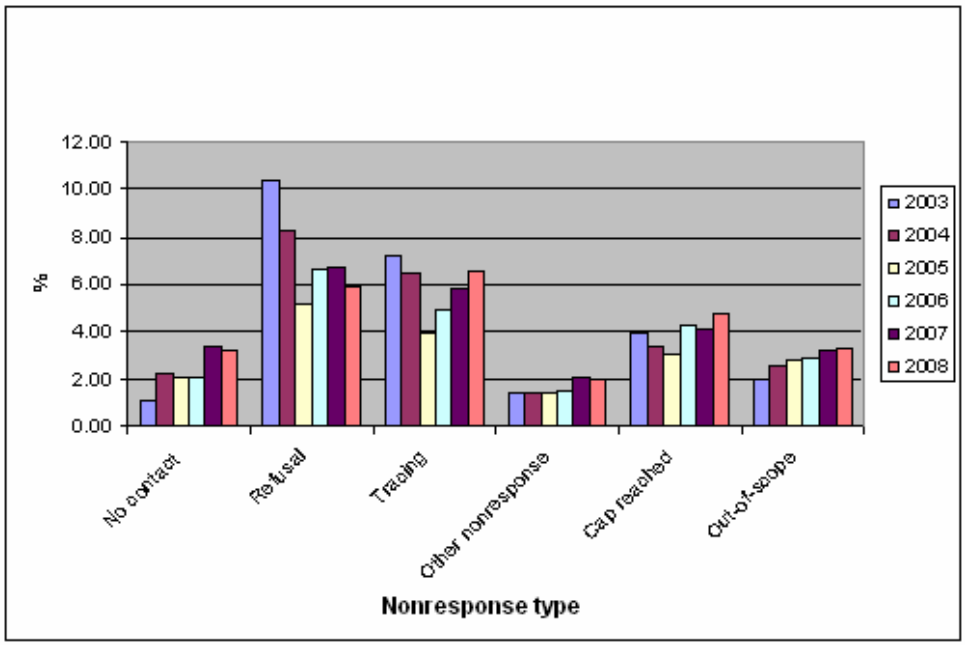


Figure 3.1-3
SLID nonresponse by type, assuming a 40 cap in all years



To get some idea of how the observed nonresponse trends might affect our ability to make inference about certain subpopulations consider the percentage of longitudinal persons associated with 2008 nonrespondent households, by age presented in Figure 3.1-5. Refusals seem to impact all age groups equally. However, young adults seem to be more often associated with no contact, cap reached and tracing households than other sample members and represent the group most affected by nonresponse. While this is not indicative of bias, it certainly highlights an area of potentially growing concern. In the next section, we consider using the paradata for a nonresponse adjustment model.

Figure 3.1-4
SLID tracing, adjusted for 2-year rule

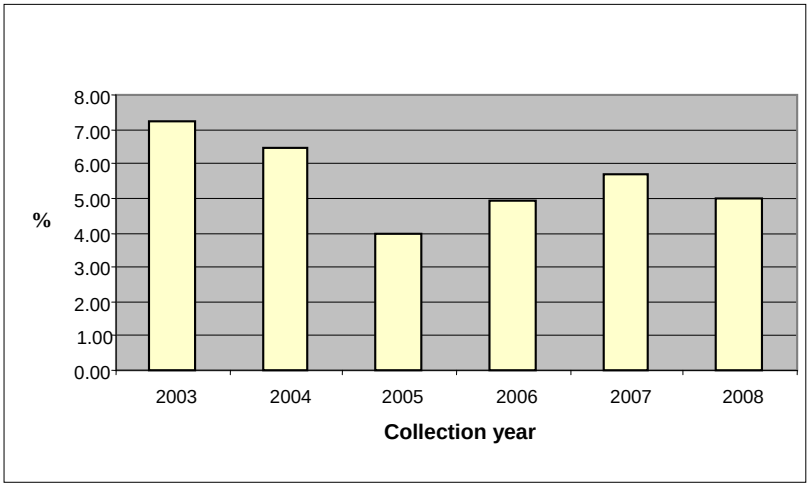
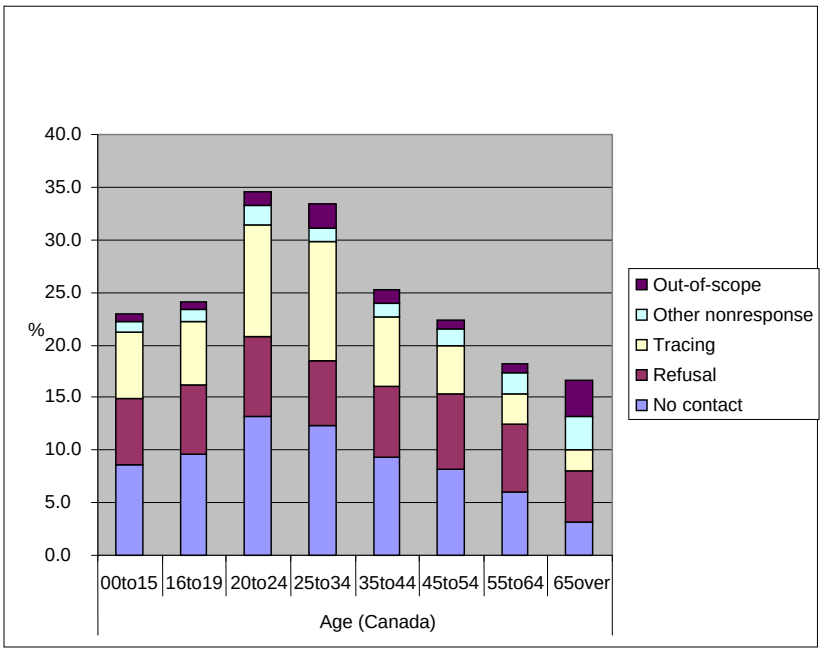


Figure 3.1-5
SLID 2008 nonresponse, by age of longitudinal respondents



4. Using paradata to model nonresponse

As mentioned in section 2 above, the SLID sample is a sub-sample of the respondents to the Canadian Labor Force Survey. We therefore have basic initial characteristics for these respondents (demographic and otherwise) which are currently used to model the nonresponse in SLID. The drawback of this process is that as the panel is aging, some of the initial information is becoming more and more outdated.

In the last few years a wealth of information obtained at the collection stage has been made available through the BTH files which is more up-to-date information and may be closely correlated with the survey variables of interest. In what follows, we propose the use of data obtained during collection (Paradata), to improve the nonresponse modeling process for SLID. This section is organized as follows: in section 4.1 to 4.3 we describe the current non response adjustment methodology for SLID. Section 4.4 provides a brief description of a study conducted to assess the use of paradata in modeling nonresponse. Section 4.5 contains a summary of the results.

4.1 Nonresponse adjustment and weighting

The nonresponse adjustment is part of the weighting process of SLID. Each year, different weights are produced: a longitudinal set of weights by panel, a combined panel longitudinal set of weights, and two cross-sectional sets of weights. The nonresponse adjustment is done at the longitudinal level and separately for each panel. Two different nonresponse adjustments are made: one for the longitudinal weighting and one for each panel hence respondents may also represent nonrespondents. These weights form the basis for the cross-sectional weighting scheme. Combining the panels, we obtain the cross-sectional sample and a weight share method is used within the household to derive weights for the co-habitants new to the household. The nonresponse adjustment is the first step in the weighting process. It is always conducted independently by panel.

4.2 SLID's current nonresponse methodology

The longitudinal units are classified as respondents, nonrespondents or ineligibles. An ineligible unit would be someone who had move out of the country or was institutionalized or was deceased. Ineligible units are excluded from the nonresponse modeling exercise. Nonrespondents from a responding household are excluded as well, even though they are treated as respondents in the nonresponse modeling process.

The following variables collected by LFS, are used for the nonresponse modeling: Age of the respondent, marital status, revenue class, employment status, household size, disability, and so on. These are binary variables for the most part and they describe the individual at the time of selection. The technique used for the non response modeling is dubbed 'modeling by segmentation' and is based on the Chi-Square Automatic Interaction Detection (CHAID) algorithm. This method operates by picking the most significant variables that explain response to the survey. This is done by comparing the values from a Pearson Chi-square among the variables and by retaining the variables which yield the highest values. Once a variable is selected, the set of individuals is divided into two groups according to the variable selected. The same procedure is applied to each subset and to subsets of subsets. No group is created if the variables are uncorrelated with the response or if the resulting group is too small. This process clearly generates a tree and is stopped once the group at the bottom of the tree branches can not be split any further. These groups are called Response Homogeneity Group (RHG).

Within each RHG the weighted response rate (WRR) is evaluated. The respondent weight is multiplied by the inverse of WRR while the nonrespondent weight is set to zero.

4.3 Nonresponse adjustment method

The current nonresponse methodology uses data gathered at the beginning of the life of the panel, another approach might have been the use of the last wave of collection data for modeling nonresponse for the current wave. Another method closely related to the second method mentioned, is the introduction of paradata into the current model.

As in Watson and Wooden (2006), we consider variables that influence contact and variables measuring the probability that once contacted the individual will provide an interview. Practically, we will consider variables such as change of address since the last wave, number of attempts to reach the respondent, Total number of contacts made, Total number of calls made, the Time Periods in which the case was accessed (day, time of the day), and demographic variables from LFS and current model such as age, sex, marital status, size of the household, province of residence, age, presence of youngster, employment, home ownership (Owner or Renter), type of dwelling and so on.

The equation below shows the logistic regression model used. This model, along with the 'stepwise' option as in Beaumont (2005), was used to incorporate the most influential variables into the model.

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \sum_{k=1}^{t} \beta_k x_{ik}$$

Where θ_i the estimated probability of response for unit, while the i 's and the β_k 's are regression parameters to be estimated and auxiliary variables respectively. The same logistic regression model was fit on three different sets of variables yielding what we denote by Model I, Model II and Model III. In Model I **only** the variables from LFS were used. Paradata variables **only** were used for Model II and finally, both paradata and LFS variables were combined for Model III. The Cluster procedure of SAS along with the Tree procedure was then used to form the RHGs.

4.4 Assessment of the models

Various statistics such as Pearson Chi-Square tests, the Hosmer-Lemeshow test, were calculated to assess the validity of the models. To measure the impact of the nonresponse model on the estimates, we assume, following Singh et al (1995), that the first wave estimate is the benchmark. Then, using only the longitudinal units of the survey, for each data source (Model I, Model II and Model III), we compare the weighted estimates of each successive wave calculated using the current's wave weight and the first wave *y*-values to that of the first wave as in the formula below. This quantity is dubbed pseudo relative bias for obvious reasons.

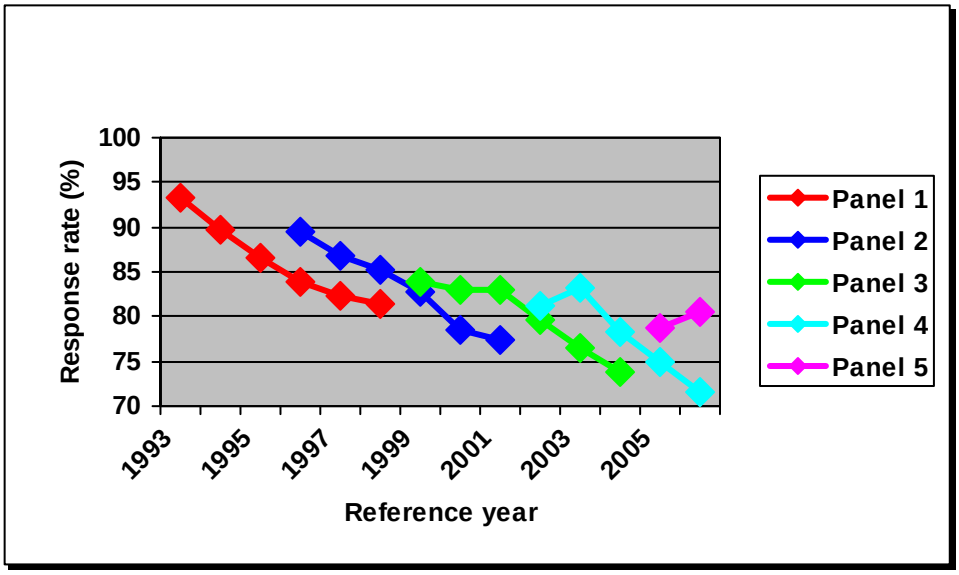
$$pseudo\_rel\_bias = \left(\frac{\hat{Y}_{1,S}^{(wi)} - \hat{Y}_{1,LFS}^{(w1)}}{\hat{Y}_{1,LFS}^{(w1)}}\right) * 100$$

$\hat{Y}_{1,S}^{(wi)}$ is an estimate for a particular variable of interest (Income, Salary, Rent and so on) and a given source ***S*** (LFS, Paradata, Combined) using weight from wave $i\ (wi)$, and *y-values* from wave 1.

4.5 Data used

Panel 4 data from SLID was used as it was first introduced in 2002 and the last wave of data for the panel was collected in early 2008. Sample size for panel 4 went from around 34,000 for wave 1 to 30,000 at wave 5. Figure 4.5-1 shows the longitudinal response rates over the years. We can see that response rates not only decline over the duration of the panel but also over the years.

Figure 4.5-1
Longitudinal response rate by wave and panel



4.6 Results

Results for Total Income and Paid rent are presented in Figures 4.6-1 and 4.6-2 below. The pseudo-bias is shown for wave 2 to 5. We can see that the combined and the paradata sources have consistently smaller pseudo-biases. For the variables considered, the nonresponse model using the paradata variables seem to bring us closer to first wave's estimates (which is arguably the closest we can get to the population parameter to be estimated).

Figure 4.6-1
Pseudo relative bias by data source: total income

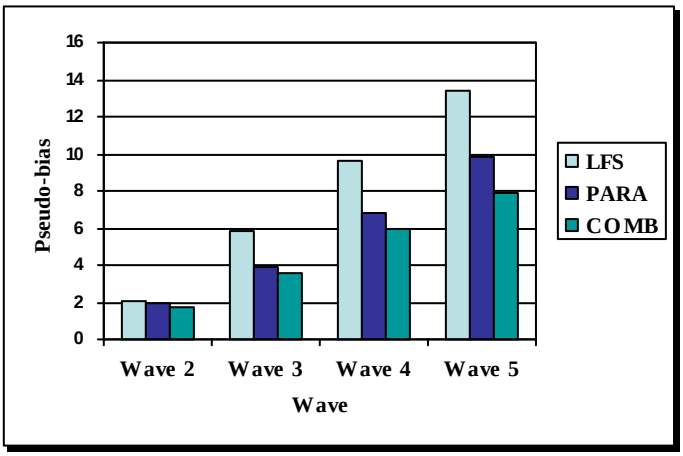
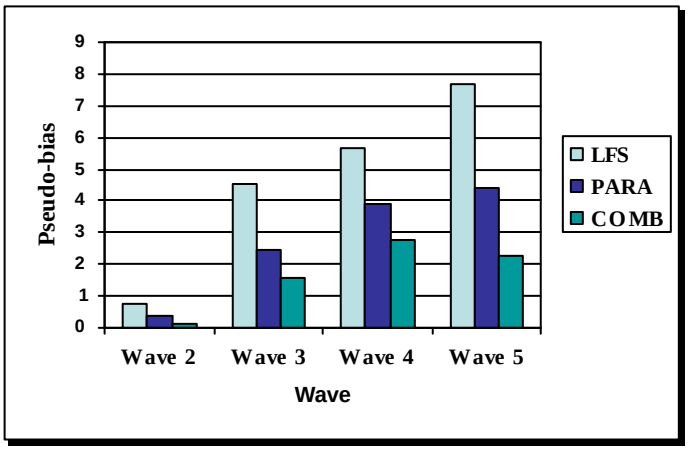


Figure 4.6-2:
Pseudo relative bias by data source: paid rent



5. Conclusion and future work

The examination of paradata has led to improvements to the active management of SLID collection activities and potentially to the nonresponse modeling as well. Based on the results of the 2008 collection, we must continue to improve our understanding of, and our ability to adapt, the collection process to make more efficient use of collection resources.

New measures, such as additional tracing sources were introduced for 2008 and will be evaluated shortly. Additionally, we will begin collecting in 2009 cell-phone numbers and e-mail addresses, two pieces of information that are not likely to change in a move. On the nonresponse front, the current study will be extended to include other SLID variables and to assess the potential impact of the new methodology on the estimates (both longitudinal and cross-sectional). Alternative models will be considered, for the different types of nonresponse (Untraced, hard refusal and so on).

This fall, Statistics Canada will pilot a new survey – the Living in Canada Survey (LCS) – to study the interplay of health status, family formation and dissolution, labour dynamics, the evolution of human and social capital and geographic effects in an effort to better understand "life-courses". The plans are to expand the content of the current SLID and to follow the longitudinal panel for an indefinite period of time. In that context, the understanding and mitigation of the effects of nonresponse will become a major preoccupation.

References

Beaumont, J-F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse through Weight Adjustment. *Survey Methodology*, 31, 227-231.

Chapman, B. (2007a). BTH File Analysis of SLID Collection 2003-2007 by Panel, internal document, Statistics Canada.

Chapman, B. (2007b). BTH File Analysis of SLID Collection 2005-2007, internal document, Statistics Canada.

Hunter, L. and Carbonneau, J.-F. (2005). An Active Management Approach to Survey Collection, *Proceedings: Symposium 2005, Methodological Challenges for Future Information Needs*. Statistics Canada

Laflamme, F., Maydan, M. and Miller, A. (2008), Using Paradata to Actively Manage Data Collection. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Laflamme, F., (2008), Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada. *Proceedings of the American Association for Public Opinion Research Section on Survey Research Methods*, American Statistical Association.

Lévesque, A. and Poulin, J. (2007). Collection Activities Management for the Survey of Labour and Income Dynamics: 2005 to 2007 Evaluations, internal document, Statistics Canada.

Scheuren, F. (2005). Paradata from Concept to Completion, *Proceedings: Symposium 2005, Methodological Challenges for Future Information Needs*. Statistics Canada

Singh, A.C., Wu, S., and Boyer, R. (1995). Longitudinal Survey Nonresponse Adjustment by Weight Calibration for Estimation of Gross Flows. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 396-401.

Watson, N. and Wooden, M. (2006). Modelling Longitudinal Survey Response: The Experience of the HILDA Survey, Hilda Project Discussion Paper Series No.2/06, Australia.