Coordinated Collection and the Quality of Paradata for CAPI Surveys at Statistics Canada

Lihua An[1], Guy Laflamme[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, Canada

Abstract
Collection cost represents a large portion of the budget allocated to carry out surveys using Computer Assisted Personal Interviewing (CAPI). To improve efficiency, the samples of three major ongoing Statistics Canada CAPI surveys using continuous collection are coordinated to ensure that the same geographical areas are targeted during the same collection period. Using this approach, we hope to reduce the interviewer travel time and distances and thus decrease collection costs. This paper focuses on the paradata quality issues encountered when we try to assess whether these expected savings are materialised. The quality and "fitness for use" of the available paradata are evaluated.

Key Words: cost model, data collection, paradata quality, sample coordination

1. Introduction

In recent years, the high cost of collecting survey data has drawn much attention at Statistics Canada (StatCan). Many initiatives have been launched to improve the efficiency of survey data collection. In particular, travel cost was identified as a major cost component for surveys using CAPI collection and new methods to reduce travel cost were sought. Such an initiative is the coordination of CAPI collection for household surveys at StatCan.

The Labour Force Survey (LFS), Canadian Community Health Survey (CCHS) and Survey of Household Spending (SHS) are the three main continuous CAPI household surveys that share the same area frame at Statistics Canada. These surveys use a two-stage design in the vast majority of the country. At the first stage, Primary Sampling Units (PSUs) are selected with probability proportional to the number of households according to Census 2001. At the second stage, a list of dwellings for the selected PSU is selected using a systematic sampling mechanism. To reduce listing costs, the sampled PSUs are shared by all surveys. Within a sampled PSU, different systematic samples are used by each survey. As a result, there is no overlap between the dwellings contacted by each survey.

In the past, each survey determined independently when the dwellings selected from a given PSU should be contacted. Under this independent random process, two surveys rarely contacted dwellings from the same PSU at the same time. As a result, interviewers would need to travel to a PSU frequently: during the first collection period for the first survey, and again in a different month for the second survey.

The coordination of CAPI collection for the LFS, CCHS and SHS started in January 2009. By coordinating the samples over time and geography, we hope to reduce travel time and the number of kilometres driven by the interviewers.

An analysis study is planned to determine the impact of the coordinated CAPI collection on the travel cost. The production data (Case Management files) and the payroll data (Survey Operations Pay System files) need to be brought together in order to model the travel cost. Questions such as whether these CAPI paradata are consistent and reliable and whether both sources paint the same picture need to be answered before further investigations into the impact of CAPI coordination can be conducted and concluded. The objective of this paper is to assess the quality and limitations of the paradata and to evaluate its fitness for use in our future researches.

This paper is organized as follows. Section 2 gives an overview of the sample coordination. Section 3 provides a definition of paradata and a brief description of the three CAPI paradata sources at Statistics Canada. Section 4 focuses on the evaluation of the paradata quality from two aspects: the internal quality of each paradata source and the coherence between the two data sources involved in our study. Conclusions and discussions are given in Section 5.

2. Overview of the sample coordination

2.1 Description of the surveys involved

The LFS is a monthly household survey with an approximate monthly sample size of 56,000 households. All members aged 15+ of the sampled households are interviewed for six consecutive months. Every month, one-sixth of the sample is renewed (i.e. one-sixth of the sample rotates out of the survey). After the initial interview (birth) is performed (typically, by a CAPI interviewer), a case is generally transferred to the computer assisted telephone interviewing (CATI) environment and is interviewed for the next five months from a CATI call centre. Overall, there are about 17,000 CAPI cases on a monthly basis. The LFS collection period is 10 days.

The redesigned SHS is an annual CAPI survey of about 9,000 households when it was first done in 2009. 1/12 of the households are interviewed each month. A sample of around 20,000 households is planned for 2010 and upcoming years. SHS interviews involve at least two visits, the first one to complete a questionnaire and drop off a diary and the second one to pick up the completed diary. The data collection period for each survey month extends to six weeks.

The CCHS is a large cross-sectional annual survey that samples about 8,000 CAPI households and 9,500 CATI households in every two-month collection period. The collection is continuous through out the year.

The analysis in this paper was performed for the July to August 2009 collection periods. Only the CAPI cases for these surveys are included in this study.

2.2 Overview of the CAPI work force

Over the 2-month collection period, approximately 700 CAPI interviewers worked on the three surveys investigated. About 20% of the interviewers worked on only one of these three surveys, while the rest (80%) of the interviewers worked on more than one surveys during the same collection period. These interviewers worked 16,000 days during which they did collection for at least one of the three surveys. The interviewers did collection in more than one cluster (i.e. PSU) about 60% of the days. It should be noted that often there are other smaller surveys going on in the field at the same time. Therefore many interviewers do not work on the collection of these three surveys exclusively. Approximately 600,000km were travelled for the three surveys.

2.3 Some basic numbers regarding the coordination

Coordinating CAPI collection relies on three cornerstones, which correspond to three stages where coordination is involved:

- The three surveys should sample dwellings from the same PSU and assign them to the same collection period as much as they can under their sampling design.
- Once the samples have been coordinated, assignment planning should ensure that all dwellings selected from a shared PSU are assigned to the same interviewer.
- If we have coordinated the samples and performed the assignment planning accordingly, the interviewers need to contact the dwellings in a coordinated fashion.

Table_1. Percentage of coordination at design, assignment, and interview stages

Stage		**Percent of coordination**
Design stage	A: design coordinated clusters	27%
	B: interviewers with design coordinated clusters	86%
Assignment stage	C: clusters with potential for coordination (out of A)	73%
	D: interviewers with potential for coordination (out of B)	60%
Interview stage	clusters with coordinated collection done (out of C)	64% (15% overall)
	interviewers with coordinated collection done (out of D)	72% (37% overall)

Table_1 summarizes how the coordination was done at each of the above three stages. A cluster is design coordinated if two surveys have sampled dwellings from the cluster in the same collection period. A cluster could also be shared by two surveys during different collection periods. In this case, it would not be a design coordinated cluster. A design coordinated cluster has potential for coordination if any interviewer has sampled households for more than one survey from this cluster. An interviewer has potential for coordination if he/she has sampled households for more than one survey from one cluster

in his/her assignment. An interviewer performed coordinated collection on a given day if he/she did collection for more than one survey in one cluster on that day. Overall, about 15% of the clusters and 37% of the interviewers had coordinated collection done at least once during the 2-month collection period.

3. Paradata sources

In addition to capturing survey responses from the interviewing process, StatCan's data collection systems also retain additional information on the collection process. This information, which can include data on contact attempts, interview flow, edits triggered and other aspects of collection, is referred to as paradata.

CAPI paradata at StatCan come primarily from three sources: 1) Case Management Application (CaseEvent files), 2) Survey Operations Pay System (SOPS) and 3) Audit trail files (Laflamme, 2010).

The CaseEvent files contain attempt transaction history records for CAPI surveys. A record is automatically created every time a case is opened, for either data collection or other purposes. It contains detailed information about every call or visit made to contact each sampled unit during the data collection process. The record also includes information on the survey and case identification, the date and time the case was open and closed, the interviewer who worked on it, the result of the call plus additional relevant information about each call or visit. Based on the transaction information, some important production concepts can be derived. For example, Complete Interview System Time represents the amount of time devoted only to conducting interviews (or calls that were recorded as completed interviews, e.g. successful calls) while Total System Time represents the total time logged onto the system (successful and unsuccessful calls) once a case is open (Laflamme, 2008).

The SOPS files contain interviewer pay claims for all collection activities (both interviewing and administrative). A SOPS record is generated every time an interviewer enters a claim for a task performed for a particular survey on a given day, either for direct data collection activities (e.g. interviewing or travelling to respondent locations) or for other purposes (e.g. supervision and specific training). Each claim includes interviewer identification, type of interviewer (regular, senior), survey name, date, activity task code (interview, training, traveling, etc.), number of hours claimed and corresponding interviewer fees, as well as additional information such as kilometres driven for CAPI surveys. It must be noted that payroll information is not recorded on a per case basis. For example, an interviewer likely combines the time spent on interviews for multiple cases on the same day and for the same survey together, and only enters one record for the total time spent on this task. The time entered in the SOPS is always rounded to 15 minutes.

The Audit Trail files contain keystroke level information that allows an analyst to observe which questionnaire fields were changed during an interview, when the changes were made, and what the changes were. In some ways it is a richer source of information than the resulting microdata file, since it allows an analyst to see how the interview progressed, not simply what the results were at the end.

In this study, we focus on the data quality of two paradata sources: the CaseEvent files and the SOPS files.

4. Assessing paradata quality

In this section, the quality of CAPI paradata is assessed from two aspects: the internal quality of a paradata source and the coherence between the two independent data sources.

4.1 Internal quality of CaseEvent data and SOPS data

4.1.1 CaseEvent data

The collection application records automatically the vast majority of the paradata. Hence, there is no additional collection cost to obtain the paradata. This also does not add to the interviewer's workload and hence limits the risk of capture errors. However, there is one important caveat for CAPI surveys: unlike the applications for CATI survey, applications for CAPI do not include call scheduler features and facilities that automatically select and assign cases. CAPI interviewers have to open and close the case each time a successful or unsuccessful attempt is made in order to record that attempt. In other words, the internal quality of production information depends on how accurately interviewers record attempts and enter and exit a survey application (Laflamme and Karaganis, 2009).
In this paper, the internal quality of the CaseEvent files is assessed from the following four aspects: under-reporting of attempts by interviewers, patterns in recording of attempts (lag between consecutive attempts), self-reporting of interviewing mode, and potential outliers in the duration of the interviews (attempts that were too short or too long).

Under-reporting of attempts: as mentioned previously, paradata for each attempt is recorded only when an interviewer opens the survey application and records that attempt. When an unsuccessful attempt is made, an interviewer may not bother to open the application, which causes under-reporting of unsuccessful attempts. Laflamme (2009) compared the average number of attempts per case in the three surveys to results from the U.S. National Health Interview Survey. There was no indication of under reporting.

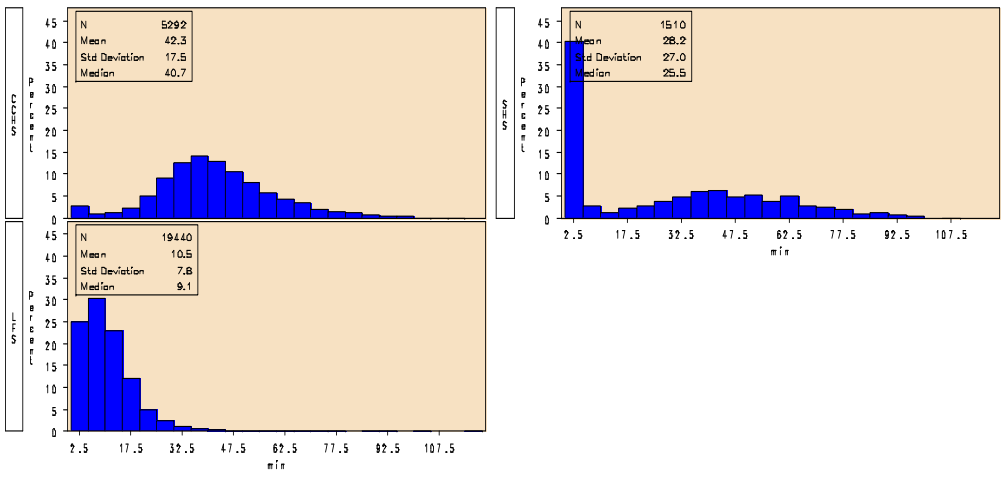
Patterns in recording of attempts: the accuracy of information can also be compromised if an attempt is not recorded at the time it was made. For example, an interviewer makes several unsuccessful attempts in a specific geographic area, returns home, and records the unsuccessful attempts all at once. Some short attempts that are clustered together are observed in the analysis, but in general, the lag times between recorded attempts were in line with the expected time required to move between cases.

Personal visits vs. telephone attempts: depending on the availability of a telephone number, CAPI interviewers can contact a household in person or by telephone (including for the interview itself) in accordance with survey procedures. Though the information on whether an attempt was made in person or over the phone is not directly available from CaseEvent files, it can be extracted from the survey application. The collection mode is self-reported by the interviewers. Analysis showed that about 25% to 40% of the CAPI attempts were made by telephone depending on the survey and interviewer. The percentage is much higher among senior interviewers. This is due to the fact that senior interviewers mostly do follow-ups and much of this can be completed over the phone.

Potential outliers in the duration of an attempt: for various reasons, an interviewer can keep a survey application open at the end of any type of attempt, which creates an attempt with an abnormally long duration. Alternatively, an interviewer can record an extremely short interview.
The unsuccessful attempts (the outcome code indicates a non-response) have an average length of 1.4 minutes. About 0.5% of the attempts lasted more than 30 minutes, and very few lasted more than 2 hours. Such extreme cases should be excluded from further studies using the available paradata. Telephone attempts tend to be shorter and have lower variance compared to personal visits. There is no significant difference across the three surveys.

Among the successful attempts with interviews completed (fully or partially), only a few lasted for more than 2 hours. The distribution of the interview duration is substantially different for the three surveys as shown in Figure_1. To have a clear view of the distributions, the interviews that lasted more than 2 hours are excluded in the graph. The duration of CCHS interviews has a nice bell-shape distribution with the longest average interview time. About 5% of the interview durations were more than 2 times standard deviation away from its mean. LFS has a skewed distribution with a long right tail and much shorter average interview time. About 2% of the interviews lasted longer than the mean plus 3 times its standard deviation. The distribution of the SHS interviews is the most complex. It follows a bimodal distribution: about 40% of the interviews are very short and the rest follow a flat bell curve. This is due to the two different visits during the SHS data collection: the first visit includes an interview with the respondent to complete a questionnaire and a diary is given to the respondent. Since the SHS questionnaire is long, the first visits takes a long time. The second visit is only to pick up the completed diary and no questionnaire is involved. Therefore it lasts much shorter. In general, the SHS interview length has a higher variance.

Figure_1. Distribution of the duration of completed interviews (in minutes)
By survey



Based on the above analyses, we conclude that the internal quality of the CaseEvent data is good.

4.1.2 SOPS data

On the pay side, interviewers are responsible for submitting their pay claims in the SOPS. These claims go through certain quality control processes, e.g. review by the supervisor to maintain the accuracy of self-reporting and the quality of the pay paradata. In this study, we are mainly interested in direct collection costs, especially those related to travel. Therefore, we only keep the records in the SOPS file with the following task codes in our analysis:

'9' - personal interview,
'10' - telephone interview, or
'13' - travel for personal interview.
The internal consistency of the SOPS data is first assessed by the consistent use of different codes.

Consistency between codes '9' and '13': It is logical that a task '9' (personal interview) must be accompanied by a task '13' (travel for interview). For some reason, some interviewers claimed time for personal interviews (code '9'), but did not claim any travel time under code '13'. A quality indicator Intra_SOPS_quality is defined to measure the degree of consistency when the interviewers report these two tasks in SOPS.

$$Intra\_SOPS\_quality = \frac{\#days\ \ for\ which\ both\ tasks\ 9\ and\ 13\ were\ reported}{\#days\ \ for\ which\ task\ 9\ was\ reported} \times 100\%.$$

Table_2. Consistency between task codes '9' and '13' by regional office

Regional Office	*Intra_SOPS_quality*	*Intra_SOPS_quality (weighted by hour)*
1	45.3%	46.4%
2	88.1%	92.6%
3	94.7%	96.9%
4	88.9%	92.7%

Table_2 shows the degree of consistency in terms of the use of codes '9' and '13' by regional office (RO). The last column is the percentage of days the interviewers reported both personal interview and travel tasks, weighted by the number of hours worked on collection during that day. Analysis showed that the consistency is very low in RO 1 but reasonable in other offices, especially when the days are weighted by hours of work. Further investigation discovered that many interviewers in RO 1 claimed kilometres under task code '9'. We suspect that these interviewers claimed travel time under code '9' as well. The finding was discussed with the regional office and measures have been taken to correct the situation.

Task '10' with kilometres: It also occurred that an interviewer recorded kilometres on a day when personal interviews were not reported. Only a few cases with this type of inconsistency were observed. We assume these are coding errors and can be ignored due to the small amount of incidences.

4.2 Coherence between the two data sources

The consistency and coherence of the two independent paradata sources has to be evaluated to assess the degree to which production and payroll data can be successfully

and correctly merged for further analysis. Records from CaseEvent and SOPS were merged by survey, day and interviewer. Again, only the records with task code showing interview or travel for interview in the SOPS file were included in the analysis. Overall, 77% of merged records contained values from both data sources. This represents 95% of the hours in the CaseEvent files or 96% of the hours in the SOPS files. In practice, interviewers who worked on a given survey on a given day may not claim their time on that day for various reasons. For example, attempts were unsuccessful and an interviewer added the time spent to the next day or an interviewer could not connect to the pay system that day. This results in a record with production time but no payroll hours. On the other hand, interviewers can claim payroll hours on a specific survey for work done outside the survey application (e.g. travel time), which explains the opposite situation (i.e. payroll hours and no production time). It may also occur that the interviewers claim time actually spent on survey A under survey B.

Coherence based on the ratio of CaseEvent hour/SOPS hour: an indication of the magnitude of the difference between production and financial information can be provided by the system time and payroll hour ratio.

Figure_2. Ratio of CaseEvent_hour/SOPS_hour
By day and survey

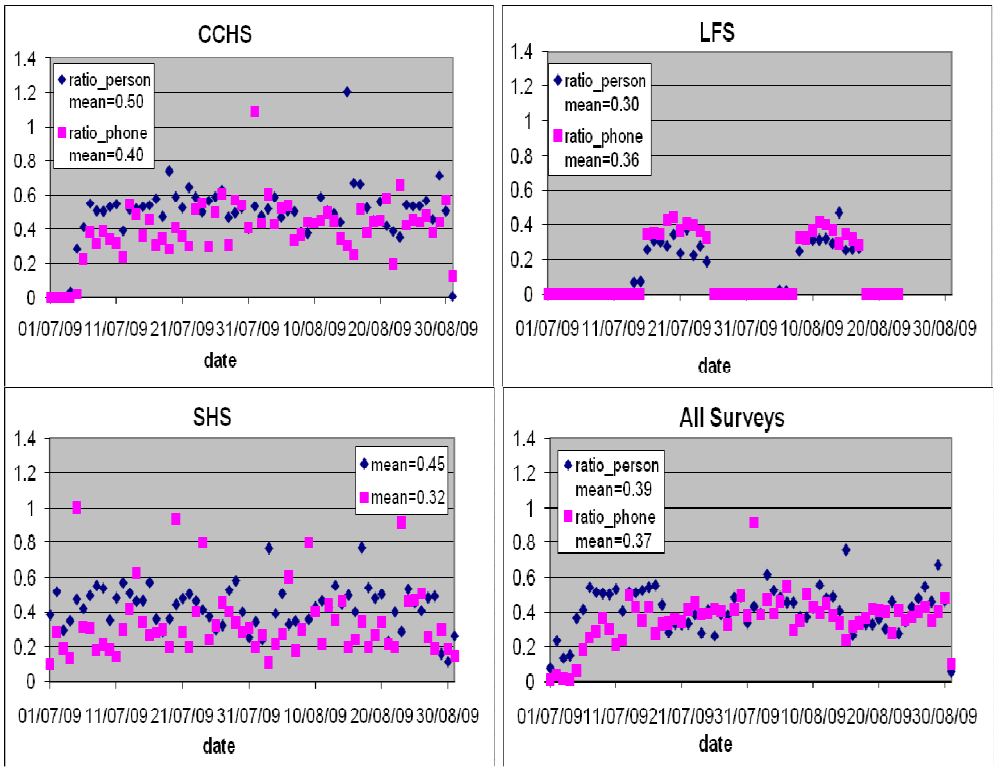


Figure_2 shows the distribution of this ratio on each day of the collection period for each survey. Personal visits and telephone calls are analysed separately to detect different patterns. The CaseEvent_hour is the total interview time recorded in the CaseEvent file

on a particular day (for personal visits or telephone attempts, separately). The SOPS_hour is the total interview hours reported in SOPS on the same day, task '9' for personal visits and task '10' for telephone attempts. CaseEvent almost always records only a fraction of the time reported in SOPS though they both record the time spent on interviews (including successful and unsuccessful attempts). This can be due to several reasons. First of all, CaseEvent records the system time. It starts when the case is open and ends when the case is closed in the system. There is much time spent on other relevant activities that is not captured in the system. For example, the interviewers often knock and wait at the door, and greet the respondents before opening the case. In addition, part of the travelling time, especially short travel between interviews, may be reported as interview time. There is also the end of day effect, e.g. time is left after an interview at the end of the day, but not enough to move to the next case to conduct another interview. Moreover, all the claims in SOPS are rounded to units of 15 minutes compared to the seconds recorded in CaseEvent. These all contribute to the difference between the hours recorded in SOPS and in CaseEvent. In general, the ratio ranges from 0 to 1 depending on survey and day.

LFS has a 10-day data collection period each month, while collection is continuous for the other two surveys. There are usually some collection activities going on outside LFS collection periods, which are not recorded in the system (CaseEvent), resulting in a "0" ratio for these days. During the collection, the ratios basically fluctuate around their respective average over the entire collection period, with a slightly lower ratio at the beginning and end of each period. This ratio shows the percentage of time spent interviewing. In general, LFS has the lowest ratio, followed by SHS while CCHS has the highest ratio. This is consistent with the survey characteristics, i.e. while the non-interviewing time (travel and other relevant activities) is about the same for all surveys, LFS requires the shortest interview time, followed by SHS and CCHS. This is reassuring about the quality of the data. It is also interesting to note that LFS has a higher ratio for telephone attempts (37% on average) than for personal attempts (30% on average), whereas CCHS and SHS show the opposite.

Coherence based on interview mode: In this study, we are particularly interested in factors related to travel cost. While travel occurs only when personal interviews are attempted, it is important to assess the consistency of reporting interview modes in the two data sources.

The two files are merged by survey, day and interviewer. We defined personal interview only day, telephone interview only day, and mixed day according to the types of attempts made for a given survey on a given day by a given interviewer. The combinations of survey, interviewer and day, are denoted by 'day' for simplicity. If the two data sources cohere perfectly, each day should be defined the same based on SOPS and CaseEvent. Quality issues can be discovered by the inconsistency of the types of days based on the two files.

The result of the analysis is summarized in Figure _3.

There was a total of 17847 days on the files, 13792 of which (77%) were present on both. The percentage increased significantly when the days were weighted by hours, i.e. this 77% of the days corresponded to 95% of the CaseEvent hours and 96% of the SOPS hours.

Among the common records in both files, the same collection modes were reported in both files for 75% of the days. This is represented by the green cells (cells on the diagonal in the bottom three rows of the flow chart) in Figure_3. The overall percentage stays the same when hours are used instead of days. The off diagonal cells (in yellow and red) represent days that were reported differently in the two files. For the non-coherent days, the biggest inconsistency occurred when the production file reported both telephone and personal interviews, but only one type of interviews, either personal, or telephone, was claimed in the financial file. This accounts for about 20% of the 13792 days in both files. This percentage is quite high when we use the number of days as a base. However, if we look at the actual interview hours recorded in CaseEvent file for the days defined differently based on SOPS and CaseEvent, they are mostly very short. For example, only 8% (146 out of 1751 hours) of the total CaseEvent hours are telephone interview in the days defined as mixed based on CaseEvent, but personal only based on SOPS. It is assumed that the interviewers decided not to claim these attempts due to the tiny amount of time involved.

Analysis showed discrepancies in the reporting of collection modes between the CaseEvent and SOPS files. However, the inconsistency was minor and can be explained when taking into account the specifics of the CAPI collection process. The overall data coherence and quality are good.

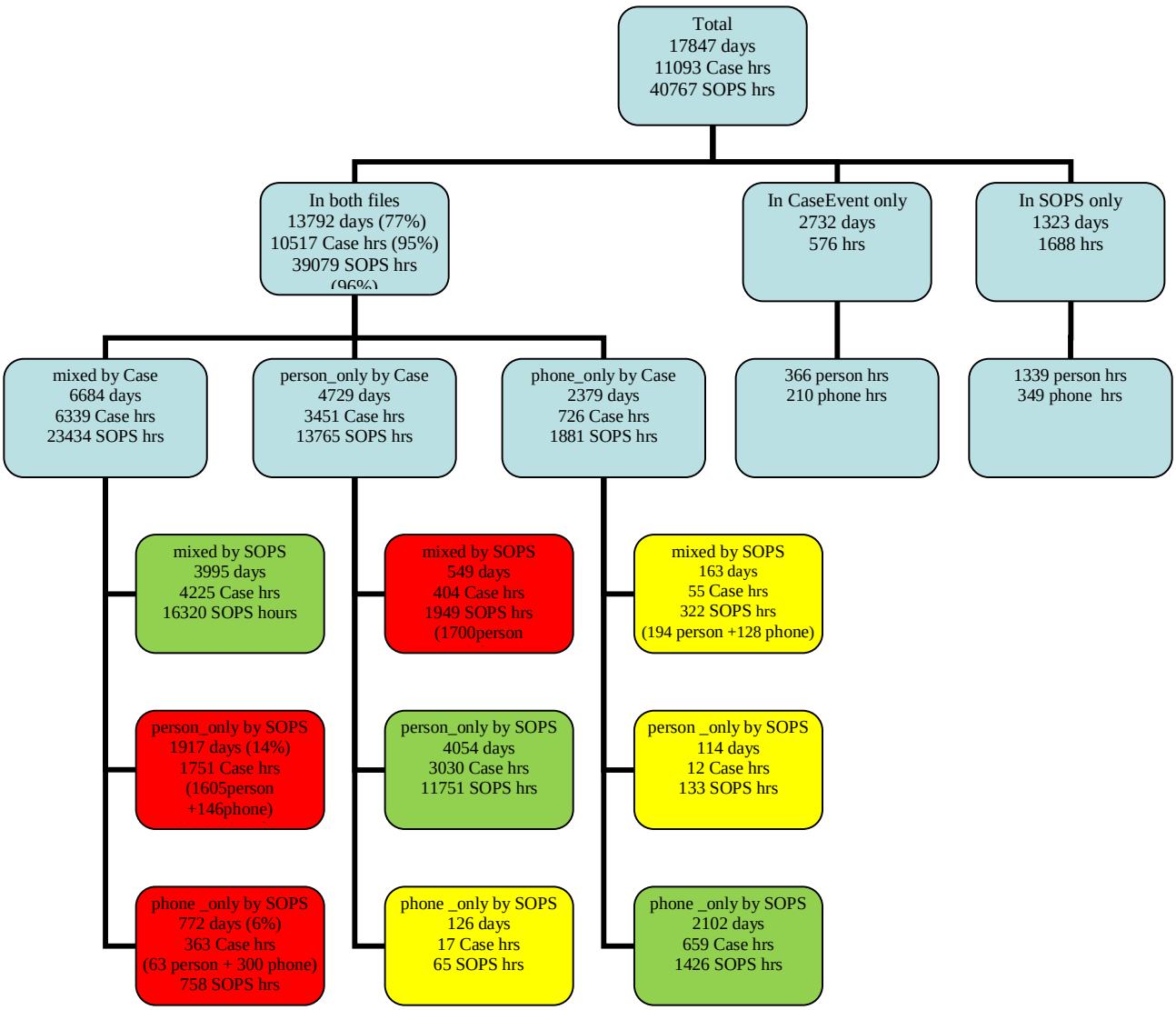
5. Conclusion and discussion

The objective of this paper was to present the initial findings in the investigation of the quality of two CAPI paradata sources at StatCan and to better understand the data collection process for CAPI surveys. In this paper, the quality of the production and financial data was assessed from two aspects: the internal quality of a paradata source and the coherence of the two paradata sources. Most of these analyses were presented to and discussed with regional office representatives to validate the findings. Based on these results and discussions, it was concluded that CAPI paradata reflects data collection processes reasonably well in the Statistics Canada environment. There is a large amount of "noise" in CAPI paradata due to the nature of the collection procedures; however, overall, CAPI paradata could also be advantageously used for further analysis, and to better understand the collection processes.

The ultimate goal of our study is to assess the cost saving due to sample coordination. Our next step is to model the travel cost to evaluate the impact of coordination. Initial analysis showed some advantages when coordination was done in terms of reduced travel time and kilometres per attempt (An and Laflamme, 2009). It also revealed the complexity of CAPI collection process. Travel cost depends on the interviewer's assignments (e.g. number of surveys and clusters to work with, how far the clusters are from each other and from the interviewer's residence), experience (senior or regular interviewer), interviewing habits (work within the same cluster or travel between clusters during a day), geography (rural or urban area), and numerous other variables. A question about the fitness for use of the available paradata is raised. The overall quality of the paradata is good. However, do they provide enough information to effectively model the cost? In other words, do the data enable us to successfully identify and include in the

model all (or most) important variables? At this point, we do not have a definite answer to this question.

Figure_3. Coherence between CaseEvent and SOPS based on interview mode



Acknowledgements

The authors would like to acknowledge François Laflamme and Milana Karaganis for their helpful discussions and valuable input in this study. We would also like to thank Hélène Bérard and Jean-Louis Tambay for reviewing the paper and providing constructive feedbacks.

References

An, L. and Laflamme, G. (2009), "Coordinated CAPI collection: Results for January to June 2009". Internal document. Labour Force Survey Section, Household Survey Methods Division, Statistics Canada.

Laflamme, F. (2008), "Data Collection Research using Paradata at Statistics Canada". Proceedings from the 2008 Statistics Canada International Symposium on Methodological Issues.

Laflamme, F. (2009), "Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada". Proceedings from the 57th International Statistical Institute Conference.

Laflamme, F. and Karaganis, M. (2009), "Experience in Assessing the quality of paradata information at Statistics Canada". Working paper for the 20th International Workshop on Survey Household Nonresponse – Paradata Panel.

Laflamme, F. and Karaganis, M. (2010), "Using Production and Cost Paradata to Better Understand the Data Collection Process and Practices for CAPI Social Surveys at Statistics Canada". European Quality Conference, Helsinki, Finland.