

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (i): Measuring the impact of editing in various phases of statistical survey processing

THE IMPACT OF EDITING ON DATA QUALITY

Submitted by Statistics Canada¹

Contributed paper

I. Introduction

1. The Project to Improve Provincial Economic Statistics (PIPES) is one of the largest and most important initiatives at Statistics Canada (Royce, 1998). It arose out of discussions in 1996 on sales tax harmonization between the Governments of Canada and several of the provinces. In three of the ten provinces, a common sales tax with a single collection authority was adopted, reducing the administrative burden on business. A formula to allocate the pooled revenues among the participating governments was determined and Statistics Canada was asked to provide detailed provincial economic data. In order to do so, Statistics Canada would need to substantially improve the quality of its statistics on provincial and territorial economies. By the end of 2000, Statistics Canada programs will be restructured and expanded to provide detailed and reliable economic accounts for the provinces and territories (Enterprise Statistics Division, 1998).

2. Approximately 200 separate business surveys are conducted regularly to gather data on different industries and commodities. Most of these will be integrated into a single master survey program called the Unified Enterprise Survey (UES). The new integrated approach focuses on enterprises: it will ensure that financial data collected from the enterprise's head office will be consistent with production and sales data received from its different establishments. The UES will collect more industry and commodity detail at the provincial level and avoid overlap between different survey questionnaires.

3. The UES is being designed to incorporate quality improvements in four areas: (i) better consistency of the methods used across industries, (ii) better coherence of the data collected from different levels of the business, (iii) better coverage of industries and (iv) better depth of information, in the sense of more content detail and estimates for more detailed domains. All of these objectives, but especially consistency, are easier to achieve with an integrated approach.

4. Presently, Statistics Canada encourages the consistent use of concepts, methods and procedures among the different surveys, through a number of mechanisms. These include the standard conceptual frameworks (such as the System of National Accounts (SNA)), standard classification systems (North American Industrial Classification System (NAICS)), a common business register, common staff pools for methodology, operations and systems development, and corporate policies related to survey-taking procedures.

¹ Prepared by Patricia Whitridge and Julie Bernier.

5. The objective of this paper is to present the results of a study that was conducted to examine the impact of data editing on the quality of the final survey data and estimates. This first section has provided an introduction and a context for the study. The second section will discuss editing in general and at what stages different edits are applied. The Unified Enterprise Survey (UES) is presented in the following section, with more information about how editing is performed for the survey. The fourth section discusses in detail the study that was conducted. Descriptive information is given about the sample and what happened to it as the data moved through the different processes. The results section presents and compares several different sets of estimates that can be produced under different processing scenarios. Finally, some general conclusions are drawn.

II. The Editing Process

6. Non response is an ongoing problem in surveys dealing with economic data. Such data are often highly skewed and quantitative in nature, so it may be difficult to fill in the data holes caused by non-response. In addition, some survey responses may be invalid when considered either within a questionnaire, or as part of a set of questionnaires. Such errors happen because methods of creating records in files are not consistent, because questions are not fully understood, or because of transcription or coding problems (Kovar and Winkler, 1996). Edits are used to identify fields, parts of questionnaires, or, in some cases, entire questionnaires that need to be corrected.

7. Typically, edits are applied at several stages during the survey process. Initially, edits are applied during data capture, as part of a CATI application, or on the keying process. These edits seek to identify missing or erroneous information within a record. The result produced by these edits may involve follow-up or confirmation questions with the respondent, turning back to a completed questionnaire form, or referring a record on to imputation.

8. The second major application of edits occurs during the edit and imputation stage. Here, records that are incomplete after all attempts to complete them with information provided by the respondent are further verified. Additional checks are applied that will look at the data confirmed by the respondent in different ways. Statistical edits that examine a record in terms of many records, such as within an imputation class or group may be specified. Edits that examine the relationships between different fields within a record (correlational edits) are also used for quantitative data. The result produced by the edits at this stage is to send the record either for manual review or to an automatic imputation system.

III. Editing the Unified Enterprise Survey

9. For the UES, certain integrating principles were adopted, including the harmonization of concepts, definitions and questionnaires. This logically extends to the objective of developing common methodology for all steps of the survey, of which the processing systems are a part.

10. In terms of data collection, all sampled businesses were sent a mail-back paper questionnaire (Communications Division, 1998). Interviewers then telephoned if they needed to clarify any aspects of the reported information, or to follow-up with businesses that had not yet returned their questionnaires. Information could be adjusted or even collected directly over the phone, using a computer assisted telephone interviewing (CATI) system. The UES also offered a basic electronic reporting option for some of its questionnaires.

11. The processing systems for the UES are divided into logical steps. At each step, different edits are applied, and different actions are taken. During collection, edits are applied to identify missing 'key' variables, such as totals of revenues, expenses, and employment. Arithmetic edits that verify totals within a 2% tolerance are also applied at collection. Any unit failing an edit, be it arithmetic or for a 'key' variable, or a unit where a change in the industry class is suspected, is followed-up. This is done either as confirmation during a CATI interview, or as a separate contact with the respondent for paper

questionnaires. During the pre-processing step of edit and imputation, some basic edits are applied that will result in some blanks being changed to zeroes, if it is obvious that such an action should take place. After these edits have been applied and the appropriate actions taken, the file will be called the 'RAW DATA' file, and is seen to represent the collected data. Given the collection system used for the UES Pilot, it is not possible to see the original data as reported by the respondent; it is then impossible to evaluate the impact of the follow-up procedure on the estimates.

12. The second series of edits are applied as part of the edit and imputation system. Different types of edits are applied, with different actions. If a record fails an edit in such a way that there is only one possible value that can be used to fix the record, then the record is corrected. This is known as a deterministic edit failure. In some cases, there are default values that are assigned, depending upon the situation. For example, in some industries, if no inventory is supplied, then it is assumed that there is no inventory and the field is set to zero. Lastly, any records with remaining edit failures are passed to donor imputation for correction. This imputation system is comprised of a pre-processing program that sets the stage for the donor imputation module. During the pre-processor, the data may be transformed (for example, the distribution of components of a total may be calculated) and some flags set for use by the donor module. After donor imputation, a pro-rate program is run, which adjusts imputed values to ensure that all totals are indeed equal to the sum of their parts, as well as calculating exact dollar amounts where distributions were imputed. Once edit and imputation has been completed, all records are clean and would pass the original edits as specified. Once the data have passed through edit and imputation, they are known as the 'PROCESSED DATA'.

13. The third step in the process is manual review, where subject matter experts manually examine the records and try to find a solution. The following records are reviewed manually: outliers, records that fail statistical edits, cases that the edit and imputation system cannot solve, and what are called 'critical imputed units'. This last category includes records that are very large or influential within their industry and province. Once the data have passed through manual review, they are considered to be 'FINAL DATA'.

IV. The Study

14. For the first year of the UES, seven industries were chosen to form the Pilot survey – industries previously without ongoing surveys. These industries are: aquaculture, taxis, couriers, lessors of real estate, real estate agents, food services and construction. There are four data collection vehicles for the UES, defined by the type of enterprise and the type of data (enterprise-level or establishment-level) required. For this study, analysis has been limited to one industry – taxis – and one type of questionnaire – that which collects establishment-level financial data. Specifically, the impact the editing process has on the quality of the final data will be analyzed, and this for revenues and expenses.

Table 1: The Sample

Sample size	320		
Out of scope (dead, inactive, etc.)	102 (32%)		
In scope	218 (68%)	Total non response	7 (3%)
		Questionnaires to process	175 (80%)
		Fully completed questionnaires	36 (17%)

15. As can be seen in Table 1 above, there were 320 units selected to be sent taxi questionnaires. After the field collection, 102 units were returned with response codes out of scope, for a variety of reasons, be it wrong industry, dead, inactive, unable to contact... Of the 218 that responded, and were in

scope, 7 units provided no data at all, but it was possible to discern that they are still active in the taxi industry. 175 units were identified as needing some processing to complete the record. The remaining 36 questionnaires were fully completed and required no imputation at all even though they could require some missing fields to be changed to zeroes.

16. The first batch of edits was developed to provide a coherent set of checks that could be used for preliminary analysis for both the data collection process and the edit and imputation process. Additional correlational and statistical edits were applied at edit and imputation. Information that could be useful for evaluation purposes was maintained throughout the edit and imputation process, as well, some basic information was retained after collection. This information can be used to evaluate the efficiency of the process, and the impact it would have on the final data quality.

17. For taxis, counts and basic performance measures about the collection and edit and imputation processes can be produced. In addition, a number of series of estimates can be calculated, using different hypotheses each time. The estimates can be calculated using simple domain estimation based on the raw data. A second set of estimates can be produced by eliminating records that are total non-respondents and adjusting the expansion weights for each stratum in consequence; this set is also based on the raw data. This re-weighting method adjusts the weights at the stratum level. Due to the small number of units involved in the taxi survey, the strata are quite broad and heterogeneous in terms of the size of the units. A third set is based on the final data that have been fully edited, imputed and reviewed. For the data from taxis, there were not enough records changed at manual review to warrant producing separate estimates from the processed data and the final data, so we have elected not to examine the estimates calculated from the processed data. Conclusions can be drawn about the impact of editing by comparing and analyzing the different sets of estimates.

18. It should be noted that the estimates prepared for this paper ARE NOT those that will be published from the survey. For the official estimates, a two-phase estimator will be used, and tax data will be added to the estimate to account for the smallest businesses that were excluded from the survey. In addition, the estimates will be post-stratified to a more current version of the frame, since several months elapsed between the time when the sample was drawn and the time when the questionnaires were completed by the respondents. The estimates presented here are simply to help understand the impact of editing.

19. For the purpose of this paper, two sections of the questionnaire will be studied: revenues and expenses. The revenue section is composed of 14 variables (R1-R14). Respondents are asked to report R12 as the sub-total of variables R1 to R11 and to report R14 as the sub-total of R12 and R13. In addition, two sub-totals were calculated and added to the table of results: $SR1 = R1 + \dots + R11$ and $SR2 = R12 + R13$ in order to verify that the rules are met for estimation. There were 23 different edit rules involving these variables in the collection system. Some of the rules are validity checks or balancing rules, while other more complicated rules involve variables from other sections of the questionnaire and even some administrative information.

20. The expense section is composed of 40 variables (E1-E40). The variable E37 is supposed to be the sum of E1 to E36 and E40 the total of E37 to E39. In addition, two calculated sub-totals were added to the table of results: $SE1 = E1 + \dots + E36$ and $SE2 = E37 + E38 + E39$. There were 21 edit rules applied to the expense section during collection. Even though there were less edits on the expense section, it was a much more complicated section. Often, respondents were unable to provide the level of detailed information that was requested on the questionnaire. In the collection system, a number of temporary sub-totals were planned, to make it possible to collect as much partial information (usually at a higher level of aggregation) as could be supplied. For example, salaries and employee benefits were often specified as one value, rather than the two fields requested on the questionnaire.

Table 2: Follow-up

Revenue section

# records in scope	218		
# records without follow-up	184 (84%)		
# records through follow-up	34 (16%)	All fields confirmed	9 (27%)
		At least one field changed	25 (73%)

Expense section

# records in scope	218		
# records without follow-up	152 (70%)		
# records through follow-up	66 (30%)	All fields confirmed	30 (46%)
		At least one field changed	36 (54%)

Both sections

# records in scope	218		
# records without follow-up	143 (66%)		
# records through follow-up	75 (34%)	All fields confirmed	22 (29%)
		At least one field changed	53 (71%)

21. Table 2 shows the distribution of records as they moved through the collection system, in terms of the revenue section, the expense section, then both sections considered together. The follow-up strategy that was implemented for UES 1997 established a list of 'priority' units that should receive preferential follow-up, starting with the largest units and those with the greatest contribution to the estimates. In practice, since 1997 was the first year for the survey, virtually all records with an edit failure at collection were followed-up. In future years, a detailed strategy for follow-up will need to be developed.

22. From the table, it can be seen that 71% of the records followed-up resulted in a change being made to at least one variable on the record. It appears that there was some over-editing of the expense section, as almost half of the follow-ups of edit failures resulted in the original data being confirmed by the respondent.

23. It should be noted that the collection system marked records where changes had been made due to follow-up, but it did not track how many variables were changed, nor the quantity by which the values were changed. This information would be interesting for a more in-depth analysis and evaluation of the collection process.

Table 3: Edit and Imputation Flow: last change in the record

Revenue section

Collection	8 (4%)
Pre processing (0 in)	203 (93%)
Processing (imputation)	5 (2%)
Manual review	2 (1%)
Total	218

Expense section

Collection	3 (1%)
Pre processing (0 in)	184 (84%)
Processing (imputation)	28 (13%)
Manual review	3 (1%)
Total	218

Both sections

Collection	3 (1%)
Pre processing (0 in)	184 (84%)
Processing (imputation)	28 (13%)
Manual review	3 (1%)
Total	218

24. Table 3 presents the final status of records after all processing, including manual review, has been completed. For each field, only the final status is kept on the database. The imputation rate is derived from data coming out of processing and manual review. Records changed at manual review may or may not have been imputed previously. Unfortunately, this information is not kept by the system.

25. In terms of the revenue section, it can be seen that the majority of changes made to this section (93% of records) involved filling zeroes into fields where it was clear that zero was the only feasible answer. There were 7 records where this section was imputed as a result of total non-response, so there was no partial imputation for this section. The expense section required more interventions by the imputation system for partial non-response.

26. It should be noted that there are very few records that were 'perfect' after collection. The most common correction was to fill a blank field with zeroes – the respondents seemed reluctant to specify all zeroes that applied. The strength of the processing system was to determine when it was appropriate to fill a field with zero, depending upon the inter-relationships of the variables and the actual data supplied by the respondent.

27. As can be seen from Table 1, there were seven records that were total non-response and required complete imputation. In addition, there were 24 records (28 changed at processing + 3 at manual review - 7 total non-response) that required at least partial imputation. In the next section, the estimates will be calculated and compared for individual fields.

V. Results

28. Table 4 shows that, for the revenue section, the imputation rate is the same for all fields (3%) and equals the total non-response rate. This means that no partial imputation was required for this specific section. All of the records were perfectly balanced after collection, this explains why the estimates of sums are also balanced to the totals after collection ($SR1=R12$ and $SR2=R14$). As shown in the table, the re-weighting provides higher estimates than the two other methods. The re-weighting is not the best

method here because the small sample size combined with the high out of scope rate and the out of date frame information did not leave enough units over which to re-distribute the weight. For the 218 respondents, there are 40 strata, with very small effective sample sizes. In general, it is the smallest units that tend not to respond; hence their weight is re-allocated to larger units, causing the over-estimation. For example, the fourth biggest unit (according to sampling information) had its weight re-distributed among the three biggest units in the sample.

29. The percentage change for a specific method has been calculated as:

$$\frac{\text{estimation with the method} - \text{estimation after collection}}{\text{estimation after collection}}$$

This percentage change after imputation appears elevated for fields R10 and R13. This is due to the fact that these items were almost always reported as zero and thus the choice of a non zero donor had a high impact on the corresponding estimate. This is aggravated when the recipient has a large weight itself. The choice of donors was approved by experts during manual review and changes were made as needed. In general, the percentage change after re-weighting seems high. For the total revenue item (R14), the non-respondents represent 10% of the total revenue in the sample, according to tax revenue information. A re-weighting effect of 24% seems exaggerated, but this is due to the high out of scope rate combined with the small sample size, as stated previously.

30. As a general comment, re-weighting is based on the sample design, which relies on frame information. The frame information can be quite out of date, if there have been no previous surveys, as is the case for taxis. However, the donor imputation is based on survey data and some administrative data that are current. This accounts for some of the discrepancies between the weighted totals.

31. For the expense section, the imputation rate varies from 3% to 14%. This section does require partial imputation. Field E2 (employee benefits) is the most often imputed; this was expected. At collection, although a temporary sub-total was available to be used when salaries and employee benefits were provided together, it appears that the sub-total was not used adequately. Field E35 is the only field to show a significant decrease – this field collects “other expenses” and often imputation distributed this amount among other expense items. The imputation rate for total expenses (E40) is 6%. This is low, due to the fact that all respondents who did not supply a value in field E40 were subject to follow-up. For this section, estimates of totals based on the sum of the raw data do not balance with respondent-specified totals (SE1≠E37 and SE2≠E40); this holds true under re-weighting. Fields E33 and E38 show an elevated rate of change for the same reason as R10 and R13, as explained above. Again, the 24% re-weighting effect on total expenses (E40) seems exaggerated since the non-respondents represent 11% of the sample according to tax expense information. This can be explained in the same way as the revenue section – the high out of scope rate, the small sample size, and the out of date frame information all contribute to the problem.

Table 4: Domain Estimation with and without Imputation

Revenue section

Variable	Imp. rate *	Weighted total after collection	Re-weighted total	Weighted total after E&I	% of change when re-weighted	% of change in E&I
R1	3%	157216187	202026637	163098957	29%	4%
R2	3%	28662555	31893946	28806555	11%	1%
R3	3%	26691494	30845122	31624061	16%	18%
R4	3%	3848843	4076987	3848843	6%	0%
R5	3%	3161258	4224939	3161258	34%	0%
R6	3%	629263	874003	629263	39%	0%
R7	3%	7470327	8703264	7470328	17%	0%
R8	3%	690478	770281	690478	12%	0%
R9	3%	28593	40458	28593	41%	0%
R10	3%	3512792	4586222	4844997	31%	38%
R11	3%	750131	969781	750131	29%	0%
SR1		232661920	288063876	244953464		
R12	3%	232661920	288063876	244953464	24%	5%
R13	3%	425521	504961	969086	19%	128%
SR2		233087441	288568838	245922550		
R14	3%	233087441	288568838	245922550	24%	6%

* including manual review

Expense section

Variable	Imp. rate *	Weighted total after collection	Re-weighted total	Weighted total after E&I	% of change when re-weighted	% of change in E&I
E1	4%	96040721	121853298	103033141	27%	7%
E2	14%	5735944	7246292	5863093	26%	2%
E3	3%	7007716	9014703	7697905	29%	10%
E4	3%	4656028	5707240	4854152	23%	4%
E5	4%	852951	984756	852608	15%	0%
E6	4%	5965173	6932015	5978866	16%	0%
E7	4%	114500	115270	114500	1%	0%
E8	4%	2468151	3348184	2492200	36%	1%
E9	4%	182249	190847	182260	5%	0%
E10	4%	5611360	6976462	6037325	24%	8%
E11	4%	452490	502145	452490	11%	0%
E12	4%	8992293	11485634	9208109	28%	2%
E13	4%	3043	3043	3043	0%	0%
E14	4%	4302516	5217615	4384020	21%	2%
E15	4%	2102995	2543025	2104984	21%	0%
E16	4%	155385	217327	155405	40%	0%
E17	4%	1410369	1741278	1408617	23%	0%
E18	4%	486453	701105	482993	44%	-1%
E19	4%	864582	1006969	866850	16%	0%
E20	4%	14447910	17091199	14964906	18%	4%
E21	4%	1054388	1344969	1053882	28%	0%
E22	4%	151247	157996	151187	4%	0%
E23	4%	2717348	3331349	2814581	23%	4%
E24	4%	786890	841032	821598	7%	4%
E25	4%	1144209	1421657	1212718	24%	6%
E26	4%	11304722	13364746	11787760	18%	4%
E27	4%	7102299	9777538	7191967	38%	1%
E28	4%	482869	528567	503900	9%	4%
E29	4%	1820252	2172021	1835002	19%	1%
E30	4%	1571982	1933032	1579415	23%	0%
E31	4%	893382	1050913	893181	18%	0%
E32	4%	8177241	9978843	8329484	22%	2%
E33	4%	3956338	4667113	5949648	18%	50%
E34	4%	2764652	3581070	2764291	30%	0%
E35	4%	8049724	8642331	4874384	7%	-39%
E36	4%	2323295	2720276	2355994	17%	1%
SE1		216153670	268391860	225256458		
E37	4%	174963281	225586915	225256458	29%	29%
E38	4%	845804	902799	2339447	7%	177%
E39	3%	1815829	2274091	1898228	25%	5%
SE2		177624913	228763805	229494133		
E40	3%	215999779	268729489	229494133	24%	6%

* including manual review

VI. General Conclusions

32. Based on this study, it has been seen that editing during collection/capture and follow-up procedures provides data of high quality, as shown by the generally low imputation rate required. By examining the weighted total based solely on the data available after collection and follow-up had taken place, it can be seen that the estimates are almost always negatively biased (underestimated) – imputation increases the estimates by approximately 5%. This is as would be expected, since imputation or re-weighting are methods used to account for non-response in a manner different than simply assuming zero data. If the solution to total non-response is to re-calculate the weights, distributing the weights of the non-respondents over the respondents, then the estimated totals tend to be overestimated, often by a significant amount. On average, the estimates were increased by about 25%, although the amount ranged up to 177%. Again, this could be expected, since the sampling fraction was greater for larger units, as is often the case in business surveys, so the weight of larger units is increased, resulting in higher estimates.

33. The principle of minimum change as applied here tended to generate a variety of imputed records that can be expected to be similar to the natural variety of records in the population.

34. Re-weighting does not provide final estimates of high quality when the sample size is small and there are not enough units to provide a good re-distribution of the weight. This is exacerbated by an elevated out of scope rate and out of date frame information for this particular survey. As well, this example involved broad strata that were not as homogeneous as one would hope. Here, re-weighting, which is equivalent to imputing the mean does not work as well as donor imputation, which imputes small units from other small units.

35. When donor imputation is used in a survey with small sample sizes and hence limited choices of donor, it is important to have a manual process of review to allow experts to identify situations where an inappropriate donor might have been used, leading to unexpected estimates.

36. It is important to ensure that adequate performance measures or descriptive information about the collection, editing, and imputation systems are maintained for evaluation purposes. Unfortunately, the next year's survey will use a different collection system that is based on a package that does not make this information available at the end of the collection and follow-up process.

37. This study has helped to highlight the importance of interviewer training. This is especially important when the interviewers are being asked to use collection systems that include more sophisticated tools, such as the temporary sub-totals available for this survey.

38. Based on this study, it is evident that further examination of the data should take place. It would be worthwhile to repeat the study using data from the other six Pilot industries, especially since some of the other industries involve a significantly higher volume of data.

REFERENCES

Communications Division, STC (1998) The How and Why of Business Statistics: The Respondent's Perspective. Statistics Canada Internal Document.

Enterprise Survey Division (1998) PIPES Information Package – Outline. Statistics Canada Internal Document.

Kovar, J. and Winkler, W. (1996) Editing Economic Data. Proceedings of the Section on Survey Research Methods of the American Statistical Association.

Royce, Don (1998). Project to Improve Provincial Economic Statistics. Proceedings of the Joint IASS/IAOS Conference in Mexico.

