MACRO-EDITING, A CASE STUDY: SELECTIVE EDITING FOR THE ANNUAL SURVEY OF MANUFACTURES CONDUCTED BY STATISTICS CANADA

Louis Boucher, Jean-Pierre Simard, Jean-François Gosselin,
Jean-Pierre Simard, Operations, Research & Development, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

KEY WORDS: Manual editing, follow ups, efficiency, critical determination, selective editing, imputation, macro-analysis.

1. INTRODUCTION

The editing of micro-data for establishment surveys requires extensive manual interventions and has proven to be an area where more significant improvement could be made in terms of achieving greater efficiencies in survey processes. In fact, recent research has tended to demonstrate that as much as 40% of all resources allocated to a survey program could be attributed to editing (Granquist, 1988), that there is a tendency to over edit on the part of statistical agencies (Granquist, 1992), and that the contribution of micro-editing to the overall data quality can be marginal (Pullum et al, 1986).

In the last few years macro-editing techniques have been developed and put forward as more efficient alternatives. Macro-editing approaches can also provide an opportunity for some significant time savings by reducing the elapsed time between the last coded questionnaire and the first tabulations and help minimizing respondent burden with optimization of edit/follow-up strategies (Latouche and Berthelot, 1990).

This paper presents results of research initiated at Statistics Canada to assess the feasibility of introducing similar methods to the Annual Survey of Manufactures (ASM). This program offered much potential for macro-editing as it is labour intensive and it involves extensive use of micro-records editing and follow-up.

2. EDITING AND VALUE ADDED

Statistics Canada conducted a study (Boucher, 1991) on the value added of editing in the 1988 Annual Survey of Manufactures (ASM). The results of the study confirmed that a significant amount of resources was devoted to edit records that had a marginal impact on the ASM estimates.

These findings constituted a very strong incentive towards rethinking current practices and led to an investigation of alternative editing strategies such as selectively editing establishments accompanied by macro-analysis.

3. BACKGROUND

3.1 The Annual Survey of Manufactures

Statistics Canada has conducted the ASM without disruption since 1917. The ASM is an **establishment** based survey covering all operations carried under one ownership at a single physical location. Both operating and commodity data are collected by the ASM. Operating data (principal statistics) are broken down into stocks, fuel and electricity, capital expenditures, salaries and wages, purchases and revenues, among others. In addition, data is collected on a large range of manufacturing materials and products.

As a result of budget cuts, the commodity portion of the program has become biennial. Every other year about 15,000 establishments are required to respond to a "long" detailed questionnaire while more than 40,000 are collected through income tax data or "short" questionnaire. Information dealing with principal statistics is available for every establishment in scope no matter the mode of collection (survey or tax data) while the commodity portion (for inputs and outputs) is only available from establishments filling out a questionnaire.

The ASM is a traditional mail-out/mail-back survey. Each record (establishment) is processed and followed-up on an individual basis and subject to all detailed editing regardless of size and relative impact. Individual follow-up takes place either to collect non-response or to react to edit failures. In the process each record is manipulated several times prior to validation. The ASM is one of the largest and most complex economic/business type survey undertaken by Statistics Canada and is a major consumer of resources.

3.2 Editing the ASM Data

The edit process is undertaken in the following manner. Once registered, records are assigned to editors for manual editing and commodity coding. These individual records are captured and submitted to the automated **Questionnaire Information Processing System (QUIPS)** for detailed editing.

In batch mode, individual records are submitted to three levels of edits, ranging from basic validity to historical ratio edits and must pass all of them in a pre-determined sequence. Follow-ups with respondents usually take place during the QUIPS stage.

Once individual records reach the 'master' status (i.e. all edits are passed or overridden) they are subjected to statistical quality control on a sample basis. This process is on-going usually from May to December of the year following the reference period. After all records of a given industry have reached 'master status', a pre-analysis exercise takes place. Such pre-analysis is rather late in the process and is the only step where records are looked at and analyzed in an aggregated way. It can result in telephone follow-ups to confirm or correct the data.

Once aggregated by industry, the data undergoes further review and analysis by Subject Matter Specialists who also decide the level of detail at which data can be published. Again, major anomalies detected as a result of these (mostly manual) interventions can be subject to follow-ups with respondents.

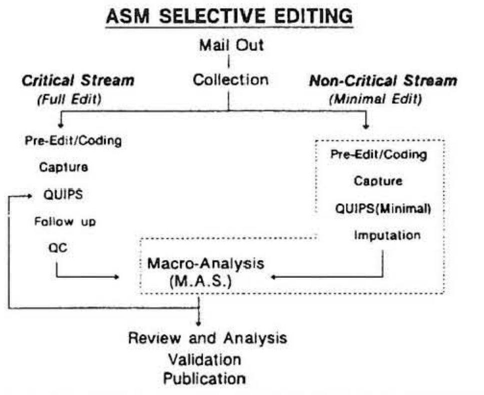
4. SELECTIVE EDITING APPROACH

4.1 Description of the Approach

The operational strategy developed involved creating two streams of processing for the ASM questionnaires: **critical** and **non-critical** streams. Critical establishments were subjected to the usual detailed edits and follow-up whereas non-critical establishments were submitted to a minimum set of edits followed by imputation and then put aside until they would be aggregated with the critical records at the industry level. Follow-up action was almost entirely eliminated from the non-critical stream.

With this strategy, the traditional detailed editing steps continued to be applied to the critical units while minimal efforts were put into non-critical records. The latter had to be complemented with some form of imputation in order to minimize the effect of curtailing the edit and follow-up process for non-critical establishments.

Figure 1



The strategy then called for both streams of records to be merged and submitted to industry level macro-editing. This consisted of scrutinizing major contributors to year-to-year change in absolute value for all variables reported in the survey. It was expected that few of the critical records would be rejected at that stage and this would provide an opportunity to detect any non-critical records having a significant impact on the estimates and follow-up on them.

4.2 The Sample

Time and resource availability forced us to restrain the test to a subset of industries from the 1990 ASM that represented the equivalent of two workloads, i.e. about 1,000 long forms.

All long form establishments in each selected industry were taken in the test. This yielded a coverage of 943 long form establishments, distributed in six industries as shown in table 1.

Table 1

SIC	Industry	# of establishment
3199	Other machinery and Equipment Industries n.e.c.	465
3711	Industrial Inorganic Chemical Industries n.e.c.	149
3712	Industrial Organic Chemical Industries n.e.c.	56
3731	Plastic and Synthetic Resin Industry	85
3751	Paint and Varnish Industry	117
3771	Toilet Preparation Industry	71
Total		943

4.3 Determining Critical Establishments

Once the sample was drawn, it became necessary to elaborate a methodology capable of determining which records would be critical.

We designed a three level determination model based on previous period shipments data;

Level 1 Select all "must" establishments and artificial splits as critical,
Level 2 Select the most significant establishments contributing to the most important commodities,
Level 3 Select the most important birth establishments.

4.3.1. Selecting "Must" Establishments

This should have usually represented the top 15 establishments per industry. The interpretation of the "must" definition seemed rather loose and the list appeared to be arbitrarily left to the judgement and knowledge of the subject matter officers. As a result the rule did not seem to be applied in a consistent manner.

4.3.2. Selecting significant commodities and establishments

The second step was to determine critical establishments based on the commodity structure of each industry. We needed to determine which commodities are most important to each industry. Within each industry, we sorted the commodity by decreasing order of shipments, based on last period's data (1989), then selected the commodities cumulating to X% of that industry shipments.

For each commodity selected, the contributing establishments were sorted by decreasing order of shipments for this particular commodity. The establishments cumulating to Y% of the total shipment value of each selected commodity would therefore be selected as critical establishments.

The following table summarizes the cut-off levels used to obtain the selected establishments;

Table 2

Industry (SIC-4)	Commodity Cut-off (X %)	Establishment cut-off (Y%)
3199	75	75
3711	85	70
3712	90	80
3731	75	60
3751	75	60
3771	75	60

The cut-offs were determined by the subject matter personnel which tailored them according to each industry's characteristics. We first ran the model at X=75% and Y=60% and analyzed the percentage of selection as well as the particular commodities and establishments that ended up being very close to the cut-off levels. It took about three simulations to finalize the cut-off levels shown in table 2.

4.3.3. Selecting Births and Transfers

New establishments in the ASM, such as births and transfers (from one industry to another), presented a problem since there was no relevant historical commodity data available from which to conduct the selection. These establishments were sorted by industry size based on whatever information was available. For each industry, the total shipment value of the smallest establishment selected from the second step (described in 4.3.2) became the cut-off for new establishments to be made critical.

Finally, some establishments could be selected more than once through the three steps. Thus the list of critical records needed to be unduplicated to account for establishments that would meet more than one criteria of selection.

5. THE RE-ENGINEERED SURVEY PROCESS

5.1 Minimal Editing for Non-critical Establishments

The idea behind the minimal edits is to reduce manual verification and follow-up and let the machine diagnose possible errors in the data reported and provide corrective measures. Given that the non-critical establishments have marginal impact on the aggregates, imputation replaces corrective actions that would otherwise have been taken as a result of follow-up. Detected errors are flagged for reference but no follow-up is undertaken at the editing stage.

Over the past years, some basic and recurring reporting errors called for immediate action from the editing staff before getting any further in the processing stream. For instances, data in straight dollars rather than thousands of dollars or data for inappropriate reporting period make any editing attempt worthless, especially in the cases of historical edits. Therefore, minimal edits consist of verifying those two elements of data with previous period data and correcting and documenting erroneous cases. Those two verifications are done with minimum time and effort from the editing staff.

Table 3

CRITICAL ESTABLISH-MENT	NON-CRITICAL ESTABLISHMENT
FULL EDITING /CODING / QUIPS	MINIMAL EDITING / CODING / QUIPS
-Balancing -Missing fields -Complete Coding -Quantities/Unit Prices Check -Ratios -Follow-up with respondents for data verification or to obtain missing data -QUIPS Corrections	-Checking for Must lines -Deleting thousands of $ -Scan data for obvious errors -Coding (no follow-up) -Same processing in QUIPS as criticals but over-rides for unit prices, wage rate ratios, range test, inter-year comparison are entered at imputation stage -QUIPS queries must be corrected without contact with respondent

5.2 Imputation

Submitting the non-critical establishments to QUIPS forced us to investigate various options regarding imputation. Our objective was to adjust some gaps and inconsistencies created by the minimum editing for this test.

Manual imputation was performed to handle infrequent cases for which the development of an automated process would not have been cost-effective.

All other cases were treated using an automated approach developed in SAS and submitted in batches once non-critical records had been submitted to QUIPS for the second time (i.e. after manually imputed).

These two sets of imputation procedures ensured that data for the non-critical establishments would reached QUIPS "master status".

6. MACRO ANALYSIS OF MERGED DATA

For the purpose of identifying significant contributor to change in the reported variables we undertook to modify an existing program (Significant Change Analysis, SCA) and tailor its output in lights of selective editing, i.e. focusing on establishments that really matter. The original SCA program was rather complex and bulky, consequently, we spent a considerable amount of time creating a new version called **Macro-editing Analysis System (MAS)**. Essentially, MAS was used by editing staff for pre-analysis as well as a tool to identify non-critical establishments that had a significant contribution. MAS was also used by subject matter staff to analyze the micro composition of the changes identified at the macro level and to document their data analysis.

The observations were sorted on the fluctuation with last period in absolute terms. Therefore, the editing staff could find the most significant changes, increase or decrease, in a top/down fashion.

In addition some indicators of the contribution of each observation toward the overall change in the aggregates were provided. These cumulative indicators were very useful in determining whether some observations shown were in fact of marginal importance.

The strength of the MAS is that it can be used to detect non-critical establishments that are significant contributors to the aggregates. They were actually flagged by an asterisk (*). Those non-criticals could be retrieved and checked for keying or response errors and followed-up if necessary.

Once the editing staff scanned the MAS and handled non-criticals, subject matter staff analyzed the industry and identified which records needed to be investigated by operations staff. This represented a significant turnover from the current way of handling SCA. Instead of applying some thresholds blindly on SCA output and documenting all outliers, the operations staff was ensuring that non-criticals were scanned and providing opportunities to subject matter staff to specifically identify which records warrant investigation and follow-up. The overall concept of top/down and significant establishments is also the backbone of these two sections.

7. ANALYSIS OF RESULTS

7.1. Shipments and Universe Coverage

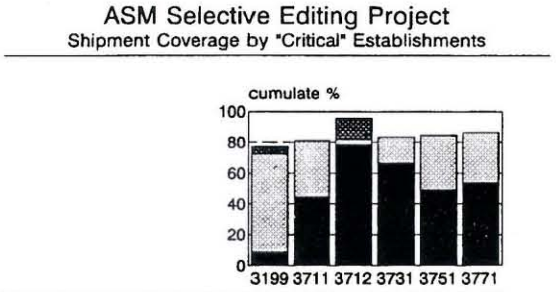
The critical determination algorithm described in Section 4.3. was programmed entirely in SAS and executed on a main frame computer. The results showed that the criteria developed for the critical determination were very effective.

Of a total of 1917 establishments in scope for the 1990 survey in the six selected industries, 943 of them were sent a long form questionnaire. After applying the critical selection model, 353 establishments met either one of the criteria and became part of the critical establishments pool. These critical establishments altogether represented **37.4%** of the total count of establishments in the long form sample. They had shipments amounting to $17.7 billion out of a total of $20.4 billion, or a share of **86.7%**.

The principal objective in determining the initial parameters (cut-offs) for the empirical selection (Level 2) was to ensure acceptable commodity coverage, as a result there was some variability in terms of the coverage when comparing one industry to another (refer to figure 2). All six recorded shipment coverage above the 75% mark. The emphasis was put on accommodating a minimum coverage of commodities manufactured, and not necessarily in meeting a predefined and uniform target in terms of the total shipment coverage. This shipment coverage however had to be relatively high.

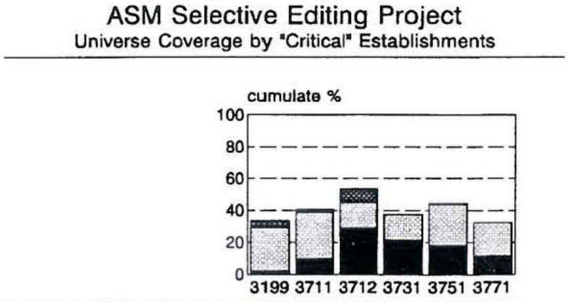
Figures 2 and 3 illustrate the shipment coverage and universe coverage. For example, in SIC 3771, by selecting **32.4%** of all the establishments in the sample, we were able to cover **86.4%** of the shipments for this industry. This meant that less than **14%** of the industry's shipments would be minimally edited under the proposed scenario even though the proportion in terms of the number of non-critical establishment was close to two thirds. At 96.1%, SIC 3712 registered the highest shipment coverage of the group while over 50% of the number of establishments in the sample were included in the critical category. Not surprisingly this was also the industry where both X and Y cut-offs were set at the highest value, respectively 90% and 80%.

Figure 2



		3199	3711	3712	3731	3751	3771
Level 3 (%)	▩	5	0.9	14.3	0	0.5	0
Level 2 (%)	▨	63.8	36.2	3.7	17.4	35.2	33.1
Level 1 (%)	■	8.6	44.2	78	66.1	48.9	53.3
Shipment Coverage (%)		77:4	81.3	96	83.5	84.6	86.4

Figure 3



		3199	3711	3712	3731	3751	3771
Level 3	▩	4.2	2	8.9	0	0.8	0
Level 2	▨	27.2	29.5	16	16.4	25.7	21.1
Level 1	■	1.9	9.4	28.6	21.2	17.9	11.3
Universe Coverage (%)		33.3	40.9	53.5	37.6	44.4	32.4

In general, the first level for "must" establishments generates a rather irregular coverage on both universe and shipments between SICs. The second level based on significant commodities seemed to be the factor having the most impact in the determination of critical establishments. The flexibility of the X/Y cut-offs enabled the subject matter staff to design a coverage that was relatively uniform across SICs and therefore ascertained greater homogeneity in the ASM estimates. The third level (births and transfers) represented a kind of safeguard. For the vast majority of industries, new records do not represent a significant impact on the estimates. In few instances, such as SIC 3199 - Other Machinery and Equipment Industries, n.e.c., the nature of the industrial activity calls for a significant number of new or transferred establishments yearly, especially in the "other, n.e.c." industry.

7.2.Commodity Coverage

The Harmonised System (HS) is used as the basis to classify goods reported in the ASM. The level of detail depends not only on the type of information being requested but also the industry. In the ASM it ranges from a minimum of HS 4 digits to a maximum of HS 9 digits. Though the test results were computed and tabulated at all six levels in use in the output section, (HS 4 to 9 digits), the results presented here were derived from the HS 6 digits table only.

Detailed results of the critical determination exercise tended to show that shipment coverage for each individual HS-6 categories stands above the cut-off in every instance, sometimes quite noticeably. This is the result of a combination of two factors. First, where the number of establishments is relatively small, the increment resulting from adding one additional establishment to the list of criticals in order to meet the cut-off (which is a minimum), may in fact bring the coverage several points above the cut-off mark. Secondly, once an establishment is selected as an important contributor to one particular commodity, all other commodity values reported by that same establishment will subsequently contribute indirectly to increase the coverage by critical establishments for those other commodities. All selected HS categories stood above the minimum with several categories at a 100% coverage rate.

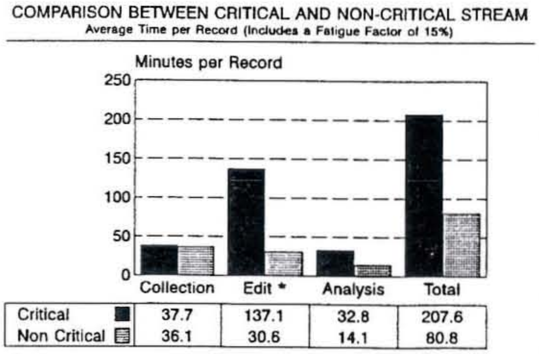
7.3. Cost Savings

During the course of processing the test, the editors kept a record of their time allocation. The time spent daily on various activities was differentiated between critical and non-critical records and allowed for comparisons. As it is often the case in work time studies, a **fatigue factor** of 15% was applied to the results to better reflect a true work environment.

As explained in great detail in Section 4, the selective editing approach had no impact on the very first tasks in the process; mailing out and filing (11.5 minutes), and collection (26 minutes). It did however for all subsequent operations in the process.

The graph shown in figure 4 summarizes the findings in terms of cost comparison. The same amount of time was required to complete the **mailing out** and subsequently the **filing** of questionnaires independently of the stream they were in.There were hardly any noticeable difference in the time required to **collect** both types of records, i.e. about 25 minutes. All together these first activities in the process accounted for 38 or 36 minutes depending whether the record is critical or not. These figures are not surprising since the strategy did not call for differentiating these tasks in any significant way.

Figure 4



COMPARISON BETWEEN CRITICAL AND NON-CRITICAL STREAM
Average Time per Record (Includes a Fatigue Factor of 15%)

	Collection	Edit *	Analysis	Total
Critical ■	37.7	137.1	32.8	207.6
Non Critical ▤	36.1	30.6	14.1	80.8

*Includes Editing, QUIPS, Follow-up and QC

On the other hand, the two labour intensive tasks of manual editing and QUIPS showed important differences. Editing non-critical records based on the new minimal procedures took **76%** less time than critical records would have required, while QUIPS took **68.4%** less time. Here the results reflect the much reduced level of intervention on the non-critical stream.

Another major difference comes from the fact that no follow-up actions are required in the non-critical pool, allowing for a 23 minutes gain over the critical.

In total over the whole range of activities covered by this process, an average of over 125 minutes per establishment could be saved on the non-critical stream. If the quality could be maintained at an acceptable level, this approach presented a serious option for higher efficiency.

7.4 Quality

In order to assess the impact of selective editing on the estimates, data from two industries (SICs 3712 and 3751) were subjected to the two different processing paths, critical and non-critical, and were subsequently compared. The application of selective editing had no noticeable impact on the estimates at the industry level. The change in the estimates as a result of editing was less than 1%.

The joint effect of the critical determination of establishments to undergo full editing and the macro-analysis seemed to provide sufficient checks to ensure the detection in an efficient manner of potential problem cases. These cases were then resolved through the usual procedures or follow-up actions while imputation did the rest.

8.RESPONSE BURDEN AND TIMELINESS

One of the added advantages of selective editing is to reduce response burden by eliminating the need to re-contact establishments in the follow-up phase. Although the study did not quantify this, it should be added that follow-up actions that were taken as a result of the MAS, would tend to be more focused, and more effective.

Despite a strong theoretical case to suggest the idea that there should be major improvements in timeliness, we unfortunately lack empirical evidence to support this. Slective editing undeniably has significantly reduced the amount of work to be done while decreasing substantially the number of follow-up interventions required.

9. CONCLUSION

The **critical determination model** proved to be effective in selecting critical establishments and showed great flexibility. The **minimum editing procedures** were effective in eliminating obvious reporting errors and preparing the documents for imputation and analysis. The staff adapted very well to them, after some adjustments in the earlier weeks of processing. The **imputation methodology** developed for this project was efficient and enabled us to get non-critical records back into the QUIPS stream. The **Macro-editing Analysis System** (MAS) was very appropriate to do pre-analysis in a selective editing environment. The significant contributing establishments were presented in a much more effective way and the additional information extended the scope of analysis.

While reducing the number of establishments requiring full editing, the approach did not noticeably impact on the quality of the ASM estimates in the test industries. There was no deterioration in timeliness but we expect that refinements based on accumulated experience will eventually lead to improvements and opportunities for the release of preliminary estimates based largely on critical units.

The pilot had its limitations there is no doubt, the sample size for one thing, the representativeness of the industries or the editors that were part of the test, the lack of empirical data to support some of the anticipated benefits, etc. It was successful however in demonstrating with a relatively limited investment in resources the real potential for efficiency gains and easy applicability of a particular macro-editing approach in the ASM; **selective editing**. Once again, one nice feature of this method was its simplicity and its ability at mobilizing human efforts to achieve a more sensible goal; correcting edit rejects only where it matters. In an operational environment where human resources still constitute the major input, that aspect has a lot of weight both in terms of future implementation and acceptance.

There is no doubt that there are some risks associated with this method, but we feel that in a context where managing resources is increasingly difficult and sometimes painful, this approach is a viable alternative and minimizes these risks.

REFERENCES:

Boucher, L. (1991). Micro-editing for the Annual Survey of Manufactures: What is the Value Added? Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference, 765-781

Granquist, L. (1992). Macro-editing - A review of some methods for rationalizing the editing of survey data. Statistical Data Editing Methods and Techniques, February 1992, Volume No.1, 117-138, Conference of European Statisticians, United Nations.

Granquist, L. (1988). On the need for Generalized Numeric and Imputation Systems, Report by Statistics Sweden. Given at the seminar on Statistical Methodology, Geneva, February 1-4, 1988.

Latouche, M., Berthelot, J.-M. (1990). Use of a score function for error correction in business surveys at Statistics Canada. Presented at the International Conference on Measurement Errors in Surveys, Tucson , Arizona.

Pullum, T., Harpman, T. and Ozsever, N. (1986). The Machine Editing of Large-Sample Surveys: The Experience of the World Fertility Survey. International Statistical Review, Volume 54, 311-326.

ALTERNATIVE IMPUTATION METHODS FOR LABOR TYPE DATA

Sandra A. West, Shail Butani, and Michael Witt, Bureau of Labor Statictics
Sandra A. West, 2 Massachusetts Ave. N.E., Washington, D.C. 20212

KEY WORDS: Regression, Bayesian, multiple imputation

1. Introduction

In this paper the results of theoretical and empirical investigations of different imputation methods for employment, wage, and ratio of wage to employment data are presented Imputation methods for item nonresponse for new establishments are also considered. The investigation began in connection with a revision project for the Bureau of Labor Statistics (BLS) program that maintains the BLS Universe Data Base (UDB). The UDB is a sampling frame of business establishments that is constructed from the State's ES-202 microdata file. The information used to maintain this file is obtained from quarterly unemployment insurance (UI) reports which each covered employer is required to submit. These quarterly reports contain, among other things, information on employment for each month of the quarter, quarterly wages, as well as a standard industrial classification (SIC) code for the establishment. Although the filing of the contribution report is mandatory under the current UI laws, each quarter there are always some reports that are filed late, delinquent accounts, as well as returns with partial data.

The goal of this project was to develop a single imputation procedure for each variable, that would work reasonably well for all SIC groups within each State. The main objective of the investigations was to compare the current ES-202 methods for imputing establishment employment and wage data with alternative procedures based on regression models.

Several types of employment and wage data were used in the studies. Most of the studies used State ES-202 microdata. Although nonrespondents were noted on the files, the actual values for the variables were never obtained. Thus nonresponse had to be simulated using the patterns of nonresponse observed on the files. For the most part, it was assumed that, within a stratum, the nonrespondents were missing at random. One study concerning employment imputation, used an alternative data source, the Current Employment Statistics (CES) Survey of establishments, conducted monthly by the BLS. With this data set a more realistic set of nonrespondents was available, so that simulation was not necessary. For the current paper the investigations will be presented only for the employment variable using this data set. In the concluding section, the results for the other variables will be summarized.

In Section 2, the data set from the CES is presented, along with a discussion of whether or not the nonrespondents are missing at random. Section 3 presents the notation used in this paper and the evaluation criteria that are used to compare the various imputation methods. Section 4 provides a description of the ES-202 Method of imputation, two hot deck procedures, and the mean imputation procedure. In Section 5, eight regression models for imputing are presented. One problem with a "best" regression-based prediction method is that all imputed values will fall on the estimated regression line and therefore, will lead to biases in estimates that involve the residual variance for nonrespondents. Simple methods that attend to this problem draw random residuals which are added to the model predictions. Details of such methods are given in Section 6. In Section 7, imputations are created under an explicit Bayesian model and multiple imputations are developed in Section 8. In a multiple imputation context, several imputed values would be created for each missing value, where ideally, uncertainty due to the estimation of the regression itself would be reflected across the imputations. Section 9 compares the results from the various imputation methods and summarizes the findings of this study. The results for the other variables are also summarized in this section.

2. Data

The purpose of this project was to develop a methodology to impute missing values for the ES-202 microdata file. Due to various reasons, it was not possible for any State to provide ES-202 microdata of the type needed. Consequently, an alternative data source, the Current Employment Statistics (CES) Survey of establishments, conducted monthly by BLS, was used for this study. The CES Survey, among other things, provides information on the monthly employment, SIC, and the closing for each establishment. The closing indicates the time frame in which the establishment responded to the survey in relation to the reference week, which is the calendar week that includes the twelfth day of the month. The first, second, and third closings normally fall, respectively, on the second, fifth, and eighth Friday following the reference week.

Most imputation procedures that are used and developed in survey sampling assume the missing data mechanism is ignorable (Little and Rubin, 1987). This issue was examined with mixed results for employment data on the CES database. Three industries were chosen and a comparison was made between those units that reported data in first or second closing (these are the respondents for this study) against those units that reported data in third closing (these are the nonrespondents for this study). The results show that there is not a significant difference in mean employment between the respondents and nonrespondents in SIC 373, but there is a difference in SICs 508 and 121. (For a definition of SICs, see Table I). Although this finding contradicts the underlying assumption of an ignorable response mechanism that is required for most of the imputation procedures examined in this paper, it does not necessarily imply that these imputation procedures are inappropriate for imputing employment values. The effectiveness of any given method is evaluated by four error measures which are discussed in the next section. Perhaps, the models could be further improved by modeling the nonresponse mechanism; this work is left for a future study.

3. Notation and Evaluation Criteria

The imputation procedures will be applied to predict the nonrespondents, by SIC group, employment size class and by month. The twelve month period ranging from November 1987 to October 1988 was considered. One, three and eight size class partitions were constructed to examine the size class effect, if any (see Table I). SIC groups 121, 373 and 508 were studied but due to the limitation of space, results are presented only for SICs 121 and 373.

Notation

Let;

t denote the current month.

$Y_{t,i}$ =Reported employment for establishment i in month t.

$\hat{Y}_{t,i}$ = Predicted employment for establishment i in month t.

$S_{12,t}$ = Set of establishments that responded by second closing for the current month, t, and have a reported value for the previous month, (t-1).

$S_{3,t,}$ = Set of establishments that responded in third closing for the current month, t, and have a reported value for the previous month (t-1).

$N_{12,t}$ = Number of units in $S_{12,t.}$

$N_{3,t}$ = Number of units in $S_{3,t.}$

$E_{t,i}$ = Error in the prediction = $(\hat{Y}_{t,i,} - Y_{t,i})$

$AE_{t,i}$ = Absolute error in the prediction = $| \hat{Y}_{t,i,} - Y_{t,i}|$

Evaluation Criteria

a. Mean Unit Error:

$$\mathrm{ME} = \sum_{size\ class} \sum_{t} \sum_{i} E_{t,i} \Big/ \sum_{size\ class} \sum_{t} N_{3,t}$$

b. Mean Unit Absolute Error:

$$\mathrm{MAE} = \sum_{size\ class} \sum_{t} \sum_{i} AE_{t,i} \Big/ \sum_{size\ class} \sum_{t} N_{3,t}$$

c. Percent Relative Error:

$$\mathrm{RE} = 100 \sum_{size\ class} \sum_{t} \sum_{i} E_{t,i} \Big/ \sum_{size\ class} \sum_{t} \sum_{i \in S_{3,t}} Y_{t,i}$$

d. Percent Relative Absolute Error:

$$\mathrm{RAE} = 100 \sum_{size\ class} \sum_{t} \sum_{i} AE_{t,i} \Big/ \sum_{size\ class} \sum_{t} \sum_{i \in S_{3,t}} Y_{t,i}$$

Note that ME (and RE) represents a macro level statistic that indicates the effect that the imputation procedure has on total employment, while MAE (and RAE) is a micro level statistic that indicates the effect on the unit.

4 ES-202 Procedure & Other Standard Methods

ES-202 Method of Imputation

Under this method, each nonrespondent's employment is imputed using its own history. The predicted value is therefore independent of size class and industry. It is computed as follows:

If $Y_{t-13,i}$, $Y_{t-12,i}$, $Y_{t-1,i}$ are nonmissing and $Y_{t-13,i} > 0$ then

$$\hat{Y}_{t,i} = (Y_{t-1,i})(Y_{t-12,i}) \Big/ (Y_{t-13,i})$$

If $Y_{t-13,i}$ or $Y_{t-12,i}$ are missing then $\hat{Y}_{t,i}$ is set equal to the most current, nonmissing $Y_{t-T,i}$ for $1 \leq T \leq 6$. Otherwise, a predicted value is not computed.

Mean Imputation Method

The mean imputation method is a common method of imputation in many surveys. If the response rate is low for a survey, then this method of imputation would not be desirable because it adversely affects the distribution of the sample units by skewing the distribution toward the mean. For any fixed SIC group, employment size class, and month t, and for all establishments in $S_{3,t}$:

$$\hat{Y}_{t,i} = \sum_{i \in S_{12,t}} Y_{t,i} \Big/ N_{12,t}$$

Thus $\hat{Y}_{t,i}$ is equal to the average employment of the respondents in the stratum.

Hot Deck Imputation Method - Random Selection

For any fixed SIC group, employment size class, and month t:

$$\hat{Y}_{t,i} = Y^{*}_{t,j}$$

where $Y^{*}_{t,j}$ is the employment of a randomly selected respondent from $S_{12,t}$. Selection was done independently within strata and with replacement.

Hot Deck Imputation Method - Nearest Neighbor

The Nearest Neighbor hot deck method is desirable because for any particular nonrespondent, it selects the respondent that appears closest to the nonrespondent in an ordered list, and substitutes the respondent's employment value for the nonrespondent's. As with the ES-202 method, this method is independent of employment size class.

For any fixed SIC group, size class, and month t, merge the respondents (from set $S_{12,t}$) and nonrespondents (from set $S_{3,t}$) into one file (set $S_t = S_{12,t} \cup S_{3,t}$), and order by $Y_{t-1,i}$ by $Y_{t-2,i}$ by State. For this ordering procedure, missing values for $Y_{t-1,i}$ and $Y_{t-2,i}$, were considered -1. Let k denote a nonrespondent and c denote a respondent such that

$$\left|Y_{t-1,c} - Y_{t-1,k}\right| \leq \left|Y_{t-1,i} - Y_{t-1,k}\right| \qquad \textit{for all } i \in S_{12,t},$$

then, $$\hat{Y}_{t,k} = Y_{t,c}$$

5. Modeling Employment by Regression

A common method for imputing missing values is via least squares regression (Afifi and Elaskoff, 1969). The following section discusses regression models for employment.

Regression Models

In two papers on estimators for total employment (West 1982, 1983), it was discovered that the most promising models for employment were the proportional regression models. These models specify that the expected employment for establishment i in the t^{th} month, given the following vector of y - values for month t-1:

$$\underline{Y}_{t-1} = [Y_{t-1,1}, Y_{t-1,2}, \ldots Y_{t-1,n}]$$

is proportional to the establishment i's previous month's employment, $Y_{t-1,i}$. That is,

$$E(Y_{t,i} \mid \underline{Y}_{t-1} = \underline{y}_{t-1}) = \beta y_{t-1,i}$$

where β is some constant depending on t.

It was further assumed that the y's are conditionally uncorrelated. That is,

$$\mathrm{cov}(Y_{t,i}, Y_{t,j} \mid \underline{Y}_{t-1} = \underline{y}_{t-1}) = \begin{cases} v_{t,i} & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases}$$

where $v_{t,i}$ represents the conditional variance of $Y_{t,i}$ which in general will depend on $Y_{t-1,i}$. Choosing a specific simple function to represent the variance $v_{t,i}$ accurately is difficult. Fortunately, knowledge of the precise form of $v_{t,i}$ is not essential, (see Royal, 1978).

The model can be rewritten as:

$$Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$$

where,

$$E\{\varepsilon_{t,i}\} = 0,$$

and

$$E\{\varepsilon_{t,i}, \varepsilon_{t,j}\} = \begin{cases} v_{t,i} & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases}$$

In West (1982), $v_{t,i} = \sigma^2 Y_{t-1,i}$ and $v_{t,i} = \sigma^2$ were considered. The model extended to two independent variables was also considered in that paper and it was found that the additional variable, Y_{t-2}, in the model was not necessary.

For the current CES data set, the following eight models were considered.

Models 1 - 4 assume $v_{t,i} = \sigma^2$:

Model 1: $Y_{t,i} = \alpha + \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 2: $Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 3: $\mathrm{Ln}(Y_{t,i}) = \alpha + \beta \mathrm{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Model 4: $\mathrm{Ln}(Y_{t,i}) = \beta \mathrm{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Models 5 - 8 are similar to models 1 - 4 respectively, except it is now assumed that $v_{t,i} = \sigma^2 Y_{t-1,i}$ for models 5 and 6, and $v_{t,i} = \sigma^2 \mathrm{Ln}(Y_{t-1,i})$ for models 7 and 8:

Model 5: $Y_{t,i} = \alpha + \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 6: $Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 7: $\mathrm{Ln}(Y_{t,i}) = \alpha + \beta \mathrm{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Model 8: $\mathrm{Ln}(Y_{t,i}) = \beta \mathrm{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

The regression model parameters were estimated using the establishments in the set $S_{12,t}$, and an imputed value was calculated for those establishments in the set $S_{3,t}$. For clarity, the subscript t was not used in conjunction with the parameters σ, α and β.

Models were fitted for the three SIC groups, twelve months of data, and three types of sample designs (1, 3 and 8 employment size classes). Based on R-squared values and other analyses, it was decided to omit models 1, 3, 5 and 7 from consideration.

Example Using Model 6

$Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$, with $v_{t,i} = \sigma^2 Y_{t-1,i}$

and β is estimated as:

$$\hat{\beta} = \sum_{i \,\varepsilon\, S_{12,t}} Y_{t,i} \Big/ \sum_{i \,\varepsilon\, S_{12,t}} Y_{t-1,i} \,.$$

For any establishment j in $S_{3,t}$ the establishment's predicted employment value at time t is:

$$\hat{Y}_{t,j} = \hat{\beta}\, Y_{t-1,j} \,.$$

Adjustments for Models 4 and 8

Consider models r, for r= 4 and 8. If it is assumed that $\varepsilon_{t,i}$ is normally distributed then $Y_{t,i}$ has a lognormal distribution with

Mean: $\exp\{\beta \mathrm{Ln}(Y_{t-1,i}) + .5\mathrm{Var}(\varepsilon_{t,i})\}$

Variance: $\{\exp[\mathrm{Var}(\varepsilon_{t,i})]-1\} \exp\{2\beta \mathrm{Ln}(Y_{t-1,i}) + \mathrm{Var}(\varepsilon_{t,i})\}$.

Therefore, an unbiased estimator of $Y_{t,k}$ is:

$$\exp\{\beta \mathrm{Ln}(Y_{t-1,k}) + .5\mathrm{Var}(\varepsilon_{t,k})\}.$$

As an estimate of $\mathrm{Var}(\varepsilon_{t,k})$, the residual mean square error, MSE_r, from regression model r was used, and the first adjustments to the regression models are:

$$\hat{Y}_{t,j} = \exp\{\hat{\beta}_r \,\mathrm{Ln}(Y_{t-1,j}) + .5\mathrm{MSE}_r\}, \quad \text{for r=4, 8.}$$

Let $Z_{t-1,i} = \mathrm{Ln}(Y_{t-1,i})$, then

$$\hat{\beta}_4 = \sum_i Z_{t-1,i} Z_{t,i} \Big/ \sum_i Z^2_{t-1,i}$$

$$\hat{\beta}_8 = \sum_i Z_{t,i} \Big/ \sum_i Z_{t-1,i} \,.$$

A second alternative adjustment to the logarithmic regression models, used by David (1986), led to the following unbiased prediction of $Y_{t,k}$

$$\exp\{\hat{\beta}\, Z_{t-1,k} + .5[\mathrm{Var}(\varepsilon_{t,k}) + Z^2_{t-1,k}]\mathrm{Var}(\hat{\beta})\}.$$

For r = 4 and 8:

$$\hat{Y}_{t,j} = \exp\{\hat{\beta}_r\, Z_{t-1,j} + .5(\mathrm{MSE}_r)(\mathrm{EMP}_r)\}$$

where $Z_{t-1,i}$ and $\hat{\beta}_r$ are defined as above, and

$$\mathrm{EMP}_4 = 1 - \{Z^2_{t-1,j} \Big/ \sum_i Z^2_{t-1,i}\}$$

$$\mathrm{EMP}_8 = 1 - \{Z_{t-1,j} \Big/ \sum_i Z_{t-1,i}\}.$$

6. Adding Residuals to the Regression Models

The methods discussed in the previous section could be thought of as imputing for missing employment by using the mean of the predicted Y_t distribution, conditional on the predictors, Y_{t-1}. As a result, the distribution of the imputed values has a smaller variance than the distribution of the true values, even if the assumptions of the model are valid. A simple strategy of adjusting for this problem is to add random errors to the predictive means, that is, draw residuals r_k, with mean zero, to add to $\hat{Y}_{t,j}$.

In this project, it was decided to consider this imputation procedure with the residuals, r_i, equalling:

1. A random normal deviate using model q.
2. A randomly selected respondent's residual, model q.
3. A randomly selected respondent's residual using model q from redefined strata Respondents and nonrespondents were restratified by $\hat{Y}_{t,j}$ using the same employment size class definitions depicted in Table I.

For each of the four models, residuals were added to the model predictions by the above three methods. For example, using model 6 and the first method described above, a prediction of $Y_{t,j}$ is:

$$\hat{Y}_{t,j} = \hat{\beta} Y_{t-1,j} + s\delta_j \qquad (6.1)$$

where δ_j is a random number from a $\mathcal{N}(0,1)$ distribution and s^2 is equal to the mean square error of the regression.

Alternatively, using the second or third method described above:

$$\hat{Y}_{t,j} = \hat{\beta} Y_{t-1,j} + r_k$$

where r_k is the residual from a randomly selected respondent k from the original employment stratum or from the redefined employment stratum.

7. Bayesian Model

In creating imputed values under an explicit Bayesian model, three formal tasks can be defined: modeling, estimation and imputation. The modeling task chooses a specific model for the data. The estimation task formulates the posterior distribution of the parameters of that model so that a random draw can be made from it. The imputation task takes one random draw from the posterior distribution of y missing, denoted by $Y_{t,BAY}$, by first drawing a parameter from the posterior distribution obtained in the estimation task and then drawing $Y_{t,BAY}$ from its conditional posterior distribution given the drawn value of the parameter.

For the modeling task, consider model 2 and $Y_{t,i}$ having a $\mathcal{N}(\beta Y_{t-1,i}, \sigma^2)$ distribution. This is the specification for the conditional density $f(Y_{t,i} \mid Y_{t-1,i}, \theta)$ where $\theta = (\beta,\sigma)$. In order to complete the modeling task, the conventional improper prior for θ, Prob(θ) proportional to a constant, is assumed.

For the estimation task, the posterior distribution of θ is needed. Standard Bayesian calculations show that:

$$f(\sigma^2 \mid Y_{t,i}) = \hat{\sigma}_1^2[n-1] \Big/ \chi^2_{n-1}$$

$$f(\beta \mid \sigma^2) = \mathcal{N}(\hat{\beta}_1, \sigma^2 v)$$

where

$$\hat{\sigma}_1^2 = \sum_i \{Y_{t,i} - \hat{\beta}_1 Y_{t-1,i}\}^2 \Big/ (n-1) = \text{MSE}$$

$$\hat{\beta}_1 = \sum_i Y_{t,i} Y_{t-1,i} \Big/ \sum_i Y^2_{t-1,i}$$

$$v = 1 \Big/ \sum_i Y^2_{t-1,i}$$

n = number of respondents.

Since the posterior distribution of θ is in terms of standard distributions, random draws can easily be computed.

The imputation task for this model is as follows:

1. Estimate σ^2 by a χ^2_{n-1} random variable, say h, and let

$$\sigma^2_2 = \hat{\sigma}_1^2 (n-1)(h)^{-1}$$

2. Estimate β by drawing one independent $\mathcal{N}(0,1)$ variate, say Z_o, and let

$$\beta_2 = \hat{\beta}_1 + \sigma_2 (v)^{.5} (Z_o)$$

3. Let n_o be the number of values that are missing, that is, the size of $S_{3,t.}$ Draw n_o values of $Y_{t,BAY}$ as

$$\hat{Y}_{t,k,BAY} = \beta_2 Y_{t-1,k} + \sigma_2 Z_k \qquad (7.1)$$

where the n_o normal deviates Z_k are drawn independently. Equation (7.1) can be rewritten as:

$$\hat{Y}_{t,k,BAY} = \hat{\beta}_1 Y_{t-1,k} + (\text{MSE})^{.5}(n-1)^{.5}(h)^{-.5}\,[(v)^{.5} Z_o Y_{t-1,k} + Z_k]$$

For model 6 an analagous Bayesian argument can be used to compute a $\hat{Y}_{t,k,BAY}$. The result will be similar, except in this case:

$$\hat{\beta}_1 = \sum_i Y_{t,i} \Big/ \sum_i Y_{t-1,i} \quad \text{and} \quad v = [\sum_i Y_{t-1,i}]^{-1}$$

8. Multiple Imputation

Multiple imputation is the technique that replaces each missing value with two or more acceptable values from a distribution of possibilities. The idea was originally proposed by Rubin. The main disadvantage that multiple imputation overcomes is that the resultant imputed values will account for sampling variability associated with the particular nonresponse model.

Multiple imputation can be obtained from the Bayesian Method by repeating the above three steps. Five repeated independent imputations were obtained by repeating the three steps. The average of these five values was taken as the imputed value.

Multiple imputation could also be obtained by using equation (6.1), adding $\mathcal{N}(0,s^2)$ residuals to the predictive mean. The error measures associated with using the average of five such repeated imputations were also considered.

9. Comparison of Imputation Methods & Conclusions

Mean Error (ME), Mean Absolute Error (MAE), Percent Relative Error (RE), and Percent Relative Absolute Error (RAE) measures were generated for the three SIC groups, each imputation method and each size class combination. However, due to space limitations, Table II presents results only for SICs 121 and 373 for ME and MAE (RE and RAE result in the same ranking of the methods).

Intuitively, it would seem that by increasing the number of size classes, greater homogeneity would be obtained and thus smaller errors would result. The data, however, showed that little or no gain in accuracy was obtained by increasing the number of size classes. This was perhaps due to the smaller number of observations within each stratum. Also, the imputation technique chosen is to be implemented for the ES-202 microdata at the state level, as opposed to the national level, such as the CES data used for this paper. This means that many State/SIC cells will have only a small number of observations. It is therefore recommended that regardless of which imputation technique is chosen, it should be employed with no more than three size classes.

Since the error measures for many of the imputation methods differ by only .01, it is very difficult to say that a Mean Error (ME) of .01 is superior to an ME of .02. While some methods, such as the Mean Imputation, can be eliminated as being the "best" imputation method, the data show that there is no one method that always yields the smallest error measures. Consequently, it was decided to search for a method that performed well on both measures and for each SIC group. As a starting point, the 96 methods, (the 32 imputation methods considered in this paper with the 3 different size class partitions) were ranked according to MAE and ME, and the top ten in each category were investigated.

For SIC 121, there were four methods that were in the top ten in both categories; three of these four involved model 6. For SIC 373, there is no method that is among the top ten in both ME and MAE. For SIC 508, the three methods that are among the top methods in ME and MAE involve logarithmic models. Next the top ten methods were examined across SIC groups for MAE and ME. According to MAE there were three methods in the top ten of each SIC. Multiple Imputation, Bayesian Model 6; Multiple Imputation, Random Normal Residual Model 6; and Regression Model 6. With respect to ME, there was no intersection of methods.

Noting the robustness of model 6, and the simplicity and intuitive appeal of Regression Model 6, it is recommended that Regression Model 6 with one size class be used. When the methods were applied to State ES-202 microdata, the same conclusion was reached.

For the wage and the ratio of wage to employment variables, the model recommended was similar to the one recommended for employment, except that the variables are transformed by the logarithm function. Also, it was recommended that the models be fit using three size class partitions. In the study for item nonresponse from new establishments in the UDB it was assumed that the wage data were always given, but the employment data were sometimes missing. The method recommended in this case was one that used a simple linear regression model through the origin, with employment as the dependent variable and wage as the independent variable. It was assumed that the variance associated with the employment variable was a function of the given wage variable. The models were fit over the all establishments, stratified by 4 digit SIC and county, that had reported both employment and wage.

Future work will include trying to model the nonrespondents, and to study estimators for total employment with a nonresponse procedure. Also a Generalized Bayesian procedure for multiple imputations using belief functions will be developed.

References

1. David, M., Little, R., Samuel, M. and Triest, R., (1986), "Alternative Methods for CPS Income Imputation", *Journal of the American Statistical Association*, vol. 81, pp. 29-41.
2. Little, R. J. A. and Rubin, D. B., (1987), *Statistical Analysis With Missing Data*, John Wiley & Sons Inc., New York.
3. Royall, R. M. and Cumberland, W. G., (1978), "Variance Estimation in Finite Population Sampling", *Journal of the American Statistical Association*, vol. 73, pp. 351-358.
4. Rubin, D., (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons Inc., New York.
5. West, S. A., (1982), "Linear Models for Monthly All Employment Data", Bureau of Labor Statistics Report.
6. West, S. A., (1983), "A Comparison of Different Ratio and Regression Type Estimators for the Total of a Finite Population", *ASA Proceedings of the Section in Survey Research Methods*.
7. West, S., Butani, S., Witt, M., Adkins, C., (1989), "Alternate Imputation Methods for Employment Data", *ASA Proceedings of the Section in Survey Research Methods*.
8. West, S., Butani, S.,and Witt, M., (1990), "Alternate Imputation Methods for Wage Data", *ASA Proceedings of the Section in Survey Research Methods*.
9. West, S., Kratzke, D., and Robertson, K., (1993), "Alternative Imputation Procedures For Item Non-response from New Establishments in the Universe", *ASA Proceedings of the Section in Survey Research Methods*.

Table I: SIC & Employment Size Class Definitions

Employment Size Class Definitions

Size class is determined by the establishment's first nonmissing employment during the time period: October 1987 to October 1988. The definition of eight, three and one size classes are as follows (table entries indicate number of employees):

EIGHT	
(1) 0 - 9	(5) 100 - 249
(2) 10 - 19	(6) 250 - 499
(3) 20 - 49	(7) 500 - 999
(4) 50 - 99	(8)1000 +

ONE: (1)-(8) collapsed

THREE: (1)-(3), (4)-(5), (6)-(8) collapsed

SIC Group Definitions

1972 SIC Code	Industry
121	Bituminous Coal and Lignite Mining
373	Ship and Boat Building and Repairing
508	Machinery, Equipment and Supplies

TABLE II: Error Measures for SICs 121 and 373

	SIC 121						SIC 373					
	Number of Employment Sizes						Number of Employment Sizes					
Imputation	1		3		8		1		3		8	
Method	ME	MAE	ME	MAE	ME	MAE	ME	MAE	ME	MAE	ME	MAE
ES 202 Method	6.6	17.4	6.6	17.4	6.6	17.4	-4.0	24.6	-4.0	24.6	-4.0	24.6
Mean	-78	206	-6.1	119	12.3	55.7	49.7	679	91.2	588	-14	275
Hot Deck:												
Rand Selection	-60	266	-3.1	161	14.5	75.4	-94	685	93.5	684	39.2	318
Near Neighbor	-2.6	17.7	-2.6	17.7	-2.6	17.7	-48	75.8	-48	75.8	-48	75.8
Reg Method												
Model 2	-.0	8.6	.0	8.7	-.2	9.0	-3.3	15.8	-3.3	15.8	-4.0	17.2
Model 4	-3.3	10.5	-1.7	9.6	-.5	9.6	-5.0	19.2	-7.5	21.9	-3.3	17.4
Model 6	-.1	8.5	-.1	8.6	.2	9.0	-3.2	16.0	-3.2	16.2	-2.7	16.6
Model 8	-3.8	10.9	-1.7	9.7	-.4	9.7	-2.5	22.1	-7.3	23.1	-3.0	17.6
Adjust Equals (.5) (MSE)												
Model 4	1.4	8.6	.6	8.7	1.2	9.4	3.1	19.5	-2.0	20.7	-1.1	17.4
Model 8	-1.8	9.8	-1.2	9.4	-.0	9.6	.6	21.8	-6.2	22.9	-2.5	17.6
Adjust Equals (.5) (MSE) (EMP)												
Model 4	1.1	8.6	.4	8.7	1.0	9.3	.1	18.9	-4.2	21.2	-1.9	17.4
Model 8	-1.9	9.8	-1.3	9.4	-.1	9.6	-.6	22.0	-6.6	22.9	-2.7	17.6
Rand Generate Normal Resid												
Model 2	-.4	18.3	.4	16.9	-.4	16.0	-4.9	61.1	-5.7	39.2	-1.6	30.8
Model 4	-1.2	37.9	-3.0	25.2	.2	18.2	20.9	70.8	5.1	57.4	-.6	37.2
Model 6	-.1	8.8	-.2	8.9	.1	9.1	-3.2	16.4	-3.2	16.5	-2.6	16.6
Model 8	-3.4	25.8	-1.0	15.4	.4	12.4	15.6	56.6	-6.8	42.7	-1.3	21.7
Rand Sel Resid												
Model 2	.4	11.4	.4	12.7	.1	11.4	-7.2	26.8	-3.1	26.2	-.8	23.1
Model 4	1.9	19.3	1.8	14.9	1.5	15.0	16.6	63.5	-2.0	29.6	-4.4	27.9
Model 6	.3	12.1	1.6	11.3	1.1	10.3	-2.6	27.1	-4.3	28.1	-3.0	22.3
Model 8	2.3	21.2	-2.2	13.7	-.8	11.8	-7.1	37.7	1.1	33.4	-2.4	26.2
Rand Sel Resid After Restrat												
Model 2	.4	11.4	1.0	11.8	1.7	10.6	-7.2	26.8	-2.1	28.8	-.7	24.4
Model 4	1.9	19.3	.5	11.5	-2.5	13.5	16.6	63.5	4.8	32.1	-13	33.0
Model 6	.3	12.1	.4	11.0	.4	11.5	-2.6	27.1	-3.9	23.4	-.3	25.4
Model 8	2.3	21.2	-1.7	13.1	-.7	11.6	-7.1	37.7	-29	53.4	3.6	32.6
Bayes Model												
Model 2	.3	17.6	-.4	16.8	-1.4	17.5	-6.2	36.5	-1.2	34.9	-2.9	35.1
Model 6	-.3	8.8	-.6	9.4	-.3	9.6	-2.6	16.3	-2.7	17.5	-2.4	19.3
Mult Imputat Bayes Model												
Model 2	-1.2	13.9	-1.7	23.1	2.3	27.8	7.9	57.3	4.0	69.6	39.4	111
Model 6	-.2	8.7	.3	8.7	-.1	9.2	-2.5	16.3	-4.7	16.7	-2.1	17.6
Mult Imp Rand Gen Norm Resid												
Model 2	-.2	12.5	-.3	11.3	-1.0	10.8	-.0	34.6	-.1	24.6	-3.8	22.7
Model 4	-2.3	18.5	-.6	13.8	-.9	12.4	9.8	37.4	-3.7	40.9	.4	23.4
Model 6	-.1	8.6	-.1	8.7	.2	9.1	-3.2	16.1	-3.1	16.2	-2.7	16.7
Model 8	-2.1	16.5	-1.0	10.9	-.2	10.5	2.9	30.0	-12	31.4	-2.6	18.2

Note: ME = Mean Error, MAE = Mean Absolute Error
Monthly Average Nos. of (Respond.,Nonrespond.): SIC 121 (337,49); SIC 373 (318,40)

VARIANCE ESTIMATION UNDER MORE THAN ONE IMPUTATION METHOD

E. Rancourt, H. Lee, Statistics Canada
and C.E. Särndal, Université de Montréal
Eric Rancourt, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

KEY WORDS: Jackknife variance estimator, ratio imputation, mean imputation, uniform nonresponse.

1. INTRODUCTION

In recent years, several papers have been written on variance estimation for data with imputed values. Different methods such as multiple imputation (Rubin, 1977;1987), the two-phase approach for ratio imputation (Rao, 1990), the model assisted approach for regression imputation (Särndal, 1990) and the jackknife method (Rao, 1992; Rao and Shao, 1992; Rao and Sitter 1992) have been proposed. All of these methods are used under the assumption that only one imputation method is used for all missing values. However, it is not uncommon to encounter surveys that make use of two or more imputation methods.

One reason for using two or more imputation methods is that they differ with respect to the auxiliary information that they require, and the more extensive information needed for a better method may not be available for all units requiring imputation. More reliable imputed values may thus be obtained for units with missing values for which there is good auxiliary information (known values of strong covariates); lacking such information, other missing values may have to be imputed by a much more elementary method, for example, by the respondent mean.

In this paper, we consider estimation of the variance of survey estimates computed from data sets containing values imputed by more than one method. For handling this type of variance estimation problem, we need a suitable tool and we consider the jackknife to be such a tool. Jackknife variance estimation for a data set with one method of single imputation was experimented in Kovar and Chen (1992). We show in this paper how the technique can be adapted to the case where more than one imputation method is used in the same data set. We concentrate on the case of two imputation methods, namely, ratio imputation using for example previous period values of the same units when available, and respondent mean imputation for the remaining units requiring imputation. Our study was motivated by the fact that some surveys conducted at Statistics Canada use this type of imputation.

In the following section, the jackknife variance estimator is described first for respondent mean imputation, and then for ratio imputation. In section 3, the technique is extended to the case where both of these imputation methods are used in the same data set. Then in section 4, a simulation study is presented. Finally conclusions are given in section 5.

2. BACKGROUND

Let $U = \{1, \ldots, k, \ldots, N\}$ be the index set of the population, and s a simple random sample without replacement (SRSWOR) of size n drawn from U. Let also r of size m and o of size l be respectively the sets of respondents and nonrespondents. Therefore, $s = r \cup o$. The variable of interest is denoted by y and we assume that $y_k > 0$ for all $k \in U$. The population mean of y is $\bar{y}_U = (1/N)\sum_U y_k$ and we are interested in finding an estimator of $\bar{y}_U$ and a corresponding variance estimator when imputation is used for nonrespondent values.

It is assumed that the response mechanism is uniform, that is, units respond according to independent Bernoulli trials. Units such that $k \in o$ are imputed by a specified single value imputation method. Let the imputed value be denoted by $\hat{y}_k$. The data after imputation are given by $\{y_{\bullet k} : k \in s\}$, where

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k & \text{if } k \in o = s - r \end{cases} \qquad (2.1)$$

Then, the usual point estimator for $\bar{y}_U$ calculated from the completed data set is

$$\bar{y}_{\bullet s} = \frac{1}{n}\sum\nolimits_s y_{\bullet k} = \frac{1}{n}\left(\sum\nolimits_r y_k + \sum\nolimits_o \hat{y}_k\right). \qquad (2.2)$$

Direct application of the jackknife variance estimation technique to the data set after imputation, $\{y_{\bullet k} : k \in s\}$, would lead to the variance estimator

$$\hat{V}_{NJ} = (1-f)\frac{n-1}{n}\sum_{j \in s}\{\bar{y}_{\bullet s}(j) - \bar{y}_{\bullet s}\}^2 \qquad (2.3)$$

or

$$\hat{V}'_{NJ} = \frac{n-1}{n}\sum_{j \in s}\{\bar{y}_{\bullet s}(j) - \bar{y}_{\bullet s}\}^2 \qquad (2.4)$$

if the finite population correction (fpc) $1-f$, $f = n/N$ is ignored, where

$$\bar{y}_{\bullet s}(j) = \frac{n\bar{y}_{\bullet s} - y_{\bullet j}}{n-1} \qquad (2.5)$$

The variance estimators $\hat{V}_{NJ}$ and $\hat{V}'_{NJ}$ are heavily biased and called "naive" since they do not account for the fact that the completed data set includes imputed values. These imputed values are treated as if they were true observations (see Rao, 1992). Rao and Shao (1992) proposed a jackknife variance estimator that corrects this problem by modifying the imputed values when the deleted unit, $y_{\bullet j}$, is a respondent. The modification reflects the fact that after deletion of the respondent, the response set is reduced by one unit and imputation should be performed using data from this reduced response set. That is, for the purpose of the variance estimation, re-imputation should be carried out when a respondent is deleted, as described in the following.

If a respondent $j \in r$ is deleted, the data set after re-imputation is given by

$$y^r_{\bullet k}(j) = \begin{cases} y_k & \text{if } k \in r(j) \\ \hat{y}_k(j) & \text{if } k \in o \end{cases} \qquad (2.6)$$

where $\hat{y}_k(j)$ is the re-imputed value based on the reduced response set, $r(j) = r - \{j\}$.

If a nonrespondent is deleted, that is, when $j \in o$, then the imputed values are unchanged. So if $k \in o$, then $y^r_{\bullet k}(j) = y_{\bullet k}$ given by (2.1) for all $k \neq j$. In summary, we have

$$y^r_{\bullet k}(j) = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k(j) & \text{if } k \in o \text{ and } j \in r \\ \hat{y}_k & \text{if } k \in o \text{ and } j \in o \end{cases} \qquad (2.7)$$

Note that while point estimation is always performed using the original imputed values, the re-imputed values are used only for variance estimation. An auxiliary file is not needed to store them.

The jackknife variance estimator of $\bar{y}_{\bullet s}$ is then

$$\hat{V}_{1J} = (1-f)\frac{n-1}{n}\sum_{j \in s}\left(\bar{y}^r_{\bullet s}(j) - \bar{y}_{\bullet s}\right)^2 \qquad (2.8)$$

with

$$s(j) = s - \{j\}$$

where

$$\bar{y}^r_{\bullet s}(j) = \frac{1}{n-1}\sum_{s(j)} y^r_{\bullet k}(j)$$

Alternatively, we can use

$$\hat{V}_{2J} = (1-f)\frac{n-1}{n}\sum_{j \in s}\left(\bar{y}^r_{\bullet s}(j) - \bar{\bar{y}}^r_{\bullet s}\right)^{2.} \qquad (2.9)$$

where

$$\bar{\bar{y}}^r_{\bullet s} = \frac{1}{n}\sum_{j \in s}\bar{y}^r_{\bullet s}(j).$$

In some cases, $\hat{V}_{1J}$ and $\hat{V}_{2J}$ are equivalent. However, in general, they are different and $\hat{V}_{1J}$ is more conservative (see Wolter, 1985, p.172). We have noted in our simulation study that $\hat{V}_{1J}$ and $\hat{V}_{2J}$ produce very close results. Following Rao, (1992) and Rao and Shao, (1992), we choose in this paper to work with $\hat{V}_{1J}$.

2.1 Mean Imputation

The mean imputation method imputes $\bar{y}_r$, the mean of the responding units, for every missing value.

When the jackknife technique is applied with this method, $y^r_{\cdot k}(j)$ in (2.7) is given by

$$y^r_{\cdot k}(j) = \begin{cases} y_k & \text{if } k \in r \\ \bar{y}_r(j) & \text{if } k \in o \text{ and } j \in r \\ \bar{y}_r & \text{if } k \in o \text{ and } j \in o \end{cases} \qquad (2.10)$$

where $\bar{y}_r(j) = (m\bar{y}_r - y_j)/(m-1)$ is the mean of the responding units after deletion of the jth unit. We then obtain the jackknife variance estimator for mean imputation from (2.8) with

$$\bar{y}^r_{\cdot s}(j) - \bar{y}_{\cdot s} = \begin{cases} \dfrac{(\bar{y}_r - y_j)}{(m-1)} & \text{if } j \in r \\ 0 & \text{if } j \in o. \end{cases} \qquad (2.11)$$

2.2 Ratio Imputation

When auxiliary information is available for all units in s ratio imputation is often used. In this case, the data after imputation are given by

$$y_{\cdot k} = \begin{cases} y_k & \text{if } k \in r \\ \dfrac{\bar{y}_r}{\bar{x}_r} x_k & \text{if } k \in o. \end{cases} \qquad (2.12)$$

where $\bar{x}_r$ is the sample mean of the auxiliary variable x for the respondents. The mean of the completed data set is given by $\bar{y}_{\cdot s} = \bar{x}_s \bar{y}_r / \bar{x}_r$.

When the jackknife technique is applied, the data set after re-imputation is given by

$$y^r_{\cdot k}(j) = \begin{cases} y_k & \text{if } k \in r \\ \dfrac{\bar{y}_r(j)}{\bar{x}_r(j)} x_k & \text{if } k \in o \text{ and } j \in r \\ \dfrac{\bar{y}_r}{\bar{x}_r} x_k & \text{if } k \in o \text{ and } j \in o \end{cases} \qquad (2.13)$$

where $\bar{x}_r(j) = (m\bar{x}_r - x_j)/(m-1)$ is the mean of the x values of the responding units after deletion of unit j from the response set. Again, we obtain the jackknife variance estimator from equation (2.8) where the values $y^r_{\cdot k}(j)$ given in (2.13) are used for the calculation of $\bar{y}^r_{\cdot s}(j)$.

3. MORE THAN ONE IMPUTATION METHOD

When more than one imputation method is used for the same data set, the idea of modifying the imputed values when a respondent is deleted can still be applied. However, special attention is required in carrying out the modification, as explained in what follows.

We consider the case of two imputation methods; ratio imputation for nonresponding units with auxiliary information and mean imputation for nonresponding units without that information.

The response set is divided into two parts; r_1 where auxiliary information is available and r_2 where it is not. Let their sizes be m_1 and m_2 respectively. Similarly, the nonresponse set is divided: o_1 of size l_1 with auxiliary information and o_2 of size l_2 without it.

The imputed values are then given by

$$y_{\cdot k} = \begin{cases} \dfrac{\bar{y}_{r_1}}{\bar{x}_{r_1}} x_k & \text{if } k \in o_1 \\ \bar{y}_r & \text{if } k \in o_2 \end{cases} \qquad (3.1)$$

Note that here the overall respondent mean $\bar{y}_r$ is imputed for $k \in o_2$. Other possibilities could be considered. If the units with and without auxiliary information are thought to be very different in their characteristics, then $\bar{y}_{r_2}$ might be better, unless m_2 is very small. An alternative which makes use of the available auxiliary information would be to impute the ratio estimate $\bar{x}_{s_1}\bar{y}_{r_1}/\bar{x}_{r_1}$ for $k \in o_2$, where $\bar{x}_{s_1}$ is the mean of x over the combined set $s_1 = r_1 + o_1$. This method was considered in Rao and Shao (1992). However, as mentioned earlier, our primary goal was to provide an appropriate variance estimator for the case of imputation method (3.1) which is often used in practice. With the imputation rule (3.1), equation (2.2) becomes

$$\begin{aligned} \bar{y}_{\cdot s} &= \frac{1}{n}\left(\sum\nolimits_r y_k + \sum\nolimits_{o_1} \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} x_k + \sum\nolimits_{o_2} \bar{y}_r \right) \\ &= \frac{1}{n}\left(m\bar{y}_r + \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} l_1 \bar{x}_{o_1} + l_2 \bar{y}_r \right) \end{aligned} \qquad (3.2)$$

The new notation introduced in these expressions is self-explanatory.

As in the case of one single imputation method, re-imputation is used when the deleted unit is a respondent.

Therefore, when the jth unit is deleted, the resulting jackknifed mean is given by:

$$\bar{y}_{\bullet s}^{r}(j) =$$

$$\frac{1}{n-1}\left(\begin{array}{ll} m\bar{y}_r + \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}}\left(l_1\bar{x}_{o_1} - x_j\right) + l_2\bar{y}_r & \text{if} \quad j \in o_1 \\ m\bar{y}_r + \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} l_1\bar{x}_{o_1} + l_2\bar{y}_r - \bar{y}_r & \text{if} \quad j \in o_2 \\ m\bar{y}_r - y_j + \frac{\bar{y}_{r_1}(j)}{\bar{x}_{r_1}(j)} l_1\bar{x}_{o_1} + l_2\bar{y}_r(j) & \text{if} \quad j \in r_1 \\ m\bar{y}_r - y_j + \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} l_1\bar{x}_{o_1} + l_2\bar{y}_r(j) & \text{if} \quad j \in r_2 \end{array}\right. \tag{3.3}$$

Then the appropriate quantity to be used in the jackknife variance estimator (2.8) is

$$\bar{y}_{\bullet s}^{r}(j) - \bar{y}_{\bullet s} = \frac{1}{n-1}\left(\frac{m\bar{y}_r + A_j l_1\bar{x}_{o_1} + B_j l_2}{n} - C_j\right) \tag{3.4}$$

where

$$A_j = \left(\begin{array}{ll} n\frac{\bar{y}_{r_1}(j)}{\bar{x}_{r_1}(j)} - (n-1)\frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} & \text{if} \quad j \in r_1 \\ \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} & \text{if} \quad j \in s - r_1 \end{array}\right.$$

$$B_j = \left(\begin{array}{ll} n\bar{y}_r(j) - (n-1)\bar{y}_r & \text{if} \quad j \in r \\ \bar{y}_r & \text{if} \quad j \in o \end{array}\right.$$

$$C_j = \left(\begin{array}{ll} y_j & \text{if} \quad j \in r \\ \frac{\bar{y}_{r_1}}{\bar{x}_{r_1}} x_j & \text{if} \quad j \in o_1 \\ \bar{y}_r & \text{if} \quad j \in o_2 \end{array}\right.$$

4. SIMULATION STUDY

4.1 Simulation Set-up

To test how well the proposed jackknife variance estimator works in a situation where more than one imputation method is used, a simulation study was carried out. For this purpose, artificial data were generated using parameters that reflect characteristics likely to be seen in reality. A population of size 400 was generated as follows. First we generated the x values from a gamma distribution with mean 48 and variance 768. Then for each value x_k, the value y_k was generated from a gamma distribution with mean $1.5x_k$ and variance d^2x_k. The constant d was chosen in order to obtain a correlation close to 0.8 between x and y. The population scatter (x_k, y_k) then follows a ratio model, that is, a linear regression through the origin, with slope close to 1.5.

The population was randomly divided into 2 sub-populations U_1 and U_2 with designated proportions; U_1 with auxiliary values x_k available, and U_2 without this information. The proportion of the population accounted for by U_1 was set to 70 % for one case and 90% for the other.

From the population, 100,000 simple random samples without replacement (SRSWOR) of size 100 were drawn. The sample size was allocated proportionally to U_1 and U_2, so that in one case the breakdown was 70% and 30%, and in the other 90% and 10%. Note that without the proportional allocation, the actual breakdowns could have been slightly different without any significant impact on the results. Nonresponse was then randomly generated using independent Bernoulli trials with a constant parameter equal to 0.3 representing the probability of nonresponse. For units with auxiliary data available, ratio imputation was performed and for the others, missing values were imputed by the respondent mean. Finally, for each sample with a realized nonresponse set and imputed values, the jackknife variance estimate was calculated.

In order to assess the performance of the jackknife variance estimator, the following Monte Carlo summary measures were calculated. Let $\bar{y}_{\bullet sm}$ be the point estimator for the population mean obtained from the mth replicate sample data after imputation and let $V(\bar{y}_{\bullet s})$ be the Monte Carlo variance of the point estimator, which is given by

$$V(\bar{y}_{\bullet s}) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{y}_{\bullet sm} - \bar{\bar{y}}_{\bullet s}\right)^2 \qquad (4.1)$$

where $M = 100{,}000$ and $\bar{\bar{y}}_{\bullet s} = (1/M)\sum_{m=1}^{M}\bar{y}_{\bullet sm}$. Now, let $\hat{V}_{1Jm}$ denote the jackknife variance estimate for the mth replicate sample. The Monte Carlo relative bias and variance of the jackknife variance estimator are given by

$$\mathrm{RB} = 100 \times \frac{\left\{\left(\frac{1}{M}\sum_{m=1}^{M}\hat{V}_{1Jm}\right) - V(\bar{y}_{\bullet s})\right\}}{V(\bar{y}_{\bullet s})} \qquad (4.2)$$

and

$$\mathrm{VV} = \sum_{m=1}^{M}(\hat{V}_{1Jm} - \bar{\hat{V}})^2 / M - 1 \qquad (4.3)$$

where

$$\bar{\hat{V}} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_{1Jm}$$

For each sample, a 95% confidence interval was also constructed using the normal distribution and the coverage of the true mean by this confidence interval was studied. The coverage rate (COVR) is defined by

$$\mathrm{COVR} = 100\frac{t}{M} \qquad (4.4)$$

where t is the number of times that the confidence interval covers the true mean.

4.2 Results

Table 1 shows the simulation results for the population generated from the ratio model as described in section 4.1. On average, 70% (or 90%) of missing values were imputed by ratio imputation and the rest by mean imputation. Two extreme cases of 100% ratio and 0% ratio were also included in the table for comparison.

Table 1.
Simulation Results for the Population with Ratio Model

Measure	Imputation Method			
	100% Ratio 0% Mean	90% Ratio 10% Mean	70% Ratio 30% Mean	0% Ratio 100% Mean
RB(%)	-2.79	-3.48	-4.77	-7.66
VV	12.38	12.49	15.74	25.89
COVR(%)	94.2	93.9	93.8	93.4

As shown in the table, the variance estimation technique appears to be well suited for cases where both ratio imputation and mean imputation are used within the same data set. It produces slight underestimation of the variance for all cases. Both the absolute RB and the variance increase with the proportion of mean imputation. While it was expected that the variance of the variance estimator would increase with the proportion of mean imputation, it is somewhat surprising to see the increasing trend of the absolute RB. Nonetheless, the coverage rates are quite good, being over 93% in all cases.

Kovar and Chen (1992) observed a positive bias with the jackknife technique. The difference between their and our jackknife variance estimators is in the use of the fpc, $1-f$. We incorporated it in our formula, whereas they did not. If in our study we had omitted the fpc, the relative bias would have been positive and in the range of 20-30%. Note that the sampling fraction they used was smaller so that the impact of the fpc was small. A more appropriate fpc was discussed in Rao and Sitter (1992). When the sample size is large however, it may be desirable to ignore the fpc in order to obtain slight overestimation rather than slight underestimation of the variance.

5. CONCLUSION

The jackknife technique seems to be an appropriate tool for variance estimation when more than one imputation method is used, at least if the response mechanism is uniform and a mixture of ratio and mean imputation is performed. In this paper, we studied only situations involving two imputation methods, but the technique can be extended to situations where three or more methods of imputation are used, as long as there is an appropriate single imputation jackknife variance estimator for each method. Extensions are also possible to cases where groups of units are deleted or where other methods than ratio and mean imputation are used in the same data set.

Further research is needed to study the jackknife variance estimation technique for the case where nearest neighbor imputation is one of the imputation methods. Also, the robustness of the variance estimator under various violations of the basic assumptions needs to be investigated.

The issue of estimating the variance in presence of more than one imputation method is of practical importance for an agency such as Statistics Canada. This paper can be seen as a first step to address the problem.

6. REFERENCES

Kovar, J.G. And Chen E.J. (1992). Variance Under Imputation: An Empirical Investigation. Presented at the 1992 Annual Meeting of the Statistical Society of Canada. Edmonton, Alberta, May 31 - June 2.

Rao, J.N.K. (1990). Variance Estimation Under Imputation for Missing Data. Unpublished paper, Statistics Canada.

Rao, J.N.K. (1992). Jackknife Variance Estimation Under Imputation for Missing Survey Data. Unpublished paper, Statistics Canada.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation. *Biometrika*, **79**, pp. 811-822.

Rao, J.N.K. and Sitter, R.R. (1992). Jackknife Variance Estimation Under Missing Survey Data. Unpublished paper, Statistics Canada.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, **72**, pp. 538-543.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. Proceedings of Statistics Canada's Symposium '90: Measurement and Improvement of Data Quality, pp. 369-380.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

COMPOSITE ESTIMATION FOR MULTIPLE STRATIFIED DESIGNS FROM A SINGLE LIST FRAME

Ismael Flores-Cervantes, William C. Iwig, and R. Ronald Bosecker, USDA/NASS
William C. Iwig, Research Division, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

KEY WORDS: Post-stratification, non-probability survey, county estimates

Abstract

The National Agricultural Statistics Service (NASS) uses stratified list frame sample designs for almost all surveys it conducts. The frame is stratified based on control data obtained from previous surveys or other sources. The County Estimates Survey uses multiple stratified designs, each based on a single control variable for each major item of interest. Currently, these data are summarized in a non-probability fashion. A composite approach for post-stratified data is proposed in this paper for summarizing County Estimates data in a probability fashion. This could strengthen the State, district, and county level estimates provided by the County Estimates Survey. A State level composite of direct expansion estimates for total hogs from eight original commodity designs in the 1991 Ohio survey provided a CV of 5.7. A composite of the eight post-stratified estimates produced a CV of 2.3.

Introduction

The National Agricultural Statistics Service (NASS) conducts many surveys covering multiple items. Most of these surveys employ stratified list frame sample designs. The List Sampling Frame contains names, addresses, and control data obtained from previous surveys or other sources for all known farming operations in each State. The control data are used as stratification variables for the different surveys conducted by NASS.

The Quarterly Agricultural Survey (QAS) is a multiple-frame probability survey that covers multiple crop acreage, stored grain, and hog items. The iist frame is stratified based on a priority scheme involving cropland, grain storage capacity, and total hog control data. A single design is developed to cover all items of interest. An area frame component accounts for the incompleteness of the list frame and ensures survey coverage for the entire farm population.

Alternatively, the County Estimates Survey is a large non-probability survey conducted in each State designed to provide estimates of crop acreage, crop yield, crop production, and livestock inventory in each county. The survey uses multiple stratified list frame designs, each based on a single control variable for each major item of interest. This design is intended to ensure adequate coverage from the list frame for all commodities of interest.

There are three basic areas of concern with the County Estimates Survey. The first is that the response rate is typically 30% or less, so nonresponse adjustments are problematic. The second is that the County Estimates Survey only covers the list frame population, which typically contains about 60% of the actual farms in a State and about 80% of the production. The third concern is that the data are currently summarized without regard to the probability of selection. If the data were summarized in a probability fashion, then NASS could possibly use these data in helping to set official USDA State level estimates.

Currently, official USDA estimates are based primarily on estimates from the QAS. If the list frame domain of both the County Estimates Survey and the QAS were the same, then the two independent estimates could be composited and provide improved State estimates for the list domain. Improved district and county level survey estimates, including variance estimates, would benefit the published series of county estimates which are continually coming under scrutiny.

A composite approach for post-stratified data is proposed in this paper for summarizing the County Estimates Survey in a probability fashion. This could strengthen the State, district, and county level estimates provided by the County Estimates survey. Analysis of the 1991 Ohio County Estimates survey data investigates the effect of this approach on State level estimates.

The County Estimates Program

Each NASS State Statistical Office publishes annual county estimates for most major agricultural commodities. Current year data are collected using primarily a mail survey in the fall of the year with some selected telephone follow-up. In addition to providing county estimates, the data are used to update the control data on the list frame in order to provide for efficient stratified sampling for all other NASS surveys. State sample sizes are dependent on the number of farms in the State, but typically range from 15,000 to 20,000 with usable record counts around 200 for major items in major counties. However, for minor crops in minor counties, sample sizes are frequently less than 10.

A key feature of the current system is the sample design which involves selecting sampling units from multiple stratified designs. For instance, there may be specific designs stratified on corn, wheat, soybeans, barley, oats, hogs, cattle, sheep, and total cropland. Typically, States will use ten or more separate designs for their survey. Individual population units on the list sampling frame would likely be included in multiple designs. The goal of this approach is to provide adequate coverage of each agricultural item of interest. This is relatively easy for major crops in a State since a sample design including all known operations with cropland would represent most major crops adequately. However, in order to provide adequate representation for rare crop and livestock items, separate stratified sample designs are developed for each agricultural commodity as needed.

The sample design strata for each commodity frame are based on the positive control data for that particular item. Table 1 illustrates the sample design that might be developed for barley in a particular State, covering all known operations that have positive control data for barley. The sample design would only include strata 10 - 40. Stratum 99 contains all population units that do not have a positive control value for barley, and so is not sampled specifically for barley.

Table 1: An Example Stratified Design for Barley

Stratum	Population Count	Boundary (acres)
99	36,000	0
10	2,500	1 - 49
20	1,000	50 - 99
30	400	100 - 299
40	100	300+
Total	40,000	

A single sample unit may be selected from multiple commodity designs. The system identifies which records are duplicated in multiple samples so that only one questionnaire is sent to each sampled unit. The same questionnaire, containing all items of interest, is used regardless of the commodity design (barley, corn, hogs, etc.) from which the record was selected.

For estimation, all survey records from all commodity designs are post-stratified together to the design strata for the commodity of interest. Direct expansion estimates are calculated based on usable sample counts within each post-stratum, not on the original sampling weight. Various ratio estimates, such as using the ratio to previous year, are also created. While this approach makes full use of the available data, the unknown quality of these non-probability survey estimates is a concern to NASS.

Alternative Post-Stratification and Composite Estimation Approach

The alternative approach investigated in this study post-stratifies the survey data from each commodity design separately to districts within the design strata employed when sampling for the commodity of interest. For example, to estimate total hogs from the soybean sample, sample records are post-stratified to cells representing districts within the hog design strata. The district refers to a group of geographically contiguous counties within a state with similar climates and agricultural practices. There are usually five to ten counties in a district and nine districts in a State. The data are post-stratified to the district level rather than to the county level to help ensure adequate sample counts in each post-stratum. Post-stratification to the district level will help provide some added control for county estimates. If only State estimates were desired and data were similar across districts, post-stratification to the State level might be satisfactory. Some commodities are very localized, and the sample may be very sparse in certain districts, so a district post-stratification would frequently be advantageous.

The post-stratified estimate of a commodity total for a particular district (d) and stratum (h) from an original design f is expressed as follows.

$$\hat{t}_{fhd} = N_{hd}(\hat{Y}_{fhd}/\hat{N}_{fhd})$$

where:

N_{hd} = known population count in post-stratum hd

$\hat{Y}_{fhd}$ = direct expansion estimate of commodity total within post stratum hd from original design f

$\hat{N}_{fhd}$ = direct expansion estimate of population count within post-stratum hd from original design f.

A key component of this estimator is the population count for each post-stratum. This value is available from the List Sampling Frame in each NASS State Statistical Office. These estimates are then summed over the strata and over the districts to provide State level estimates for each original commodity design (f) as follows.

$$\hat{t}_f = \sum_d \sum_h \hat{t}_{fhd}$$

The composite State level estimate using all the original commodity designs is expressed as:

$$t = \sum_f \lambda_f \hat{t}_f \;/\; \sum_f \lambda_f$$

Where λ_f is the inverse of the estimated variance of $\hat{t}_f$.

Data

The proposed estimator was applied to data from the 1991 Ohio County Estimates Survey. Unfortunately, the survey data file did not indicate from which design(s) each record was originally selected. Approximately 30,000 records were mailed, with 11,178 records returned with usable data. These 11,178 records were stratified according to the original sample designs, and samples were selected for analysis with sampling rates similar to those actually used.

Table 2 presents the sample design stratified based on soybean planted acreage. Other designs included in the study were for corn, hay, oats, wheat, cattle, hogs, and sheep, and are similar in nature. The "Other" stratum contains all records in the population that do not have a soybean control value. This stratum was not sampled originally for soybeans, but is sampled for this study so all commodity designs cover the same population. The sample sizes for each design should provide reliable State level estimates for the commodities of interest, but may not provide reliable county estimates. The sampling weights used for estimation are based on the population and analysis sample size, even though the sampling was actually conducted from the 11,178 records available. For example, a sampling weight of 5080/684 is assigned to units in stratum 1 from the soybean design shown in Table 2. Consequently, the direct expansion and post-stratified estimates utilize pseudo design-based weights.

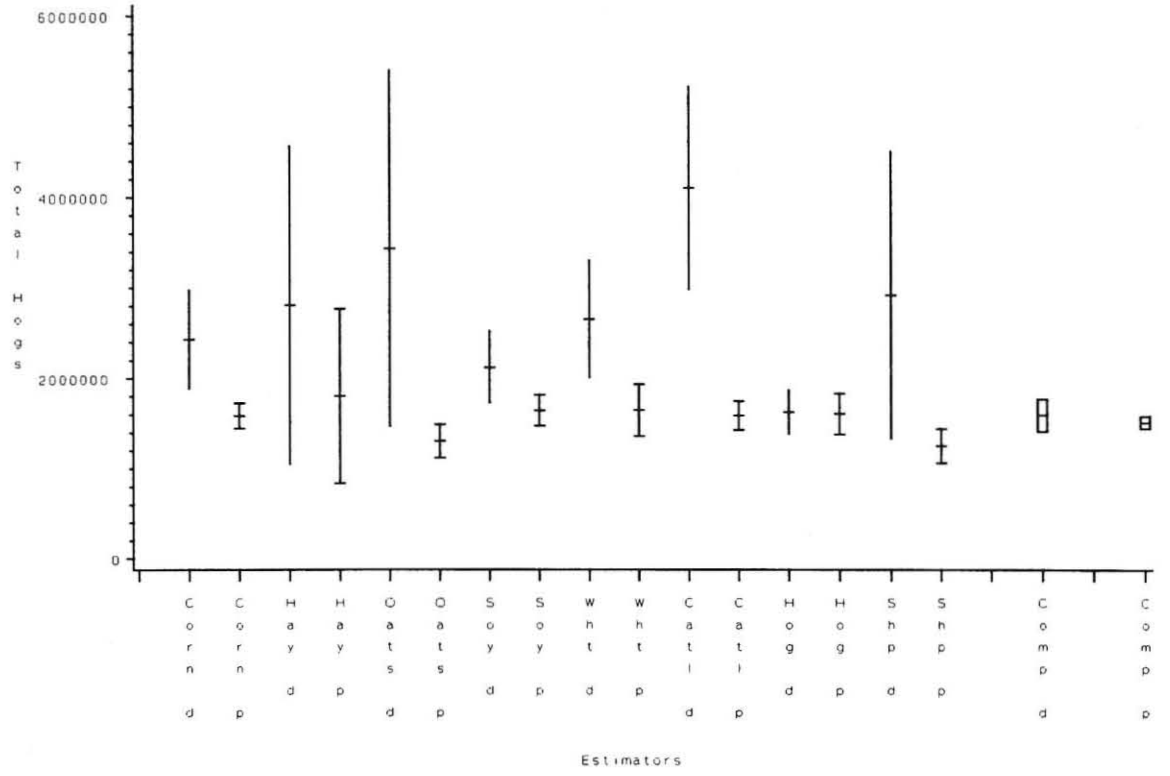
Table 2: Soybean Stratified Design, Ohio 1991

Stratum	Stratum Boundary	Population	Records Available	Analysis Sample Size
99	Other	32708	5162	135
1	01 - 24	5080	1287	684
3	25 - 99	10665	2379	683
5	100 - 249	6142	1424	587
6	250 - 499	2602	658	448
7	500 - 999	777	214	214
8	100 +	121	54	54
Total		**58095**	**11178**	**2805**

Results

Illustration 1 presents State level estimates of total hogs and the associated 95% confidence intervals for the direct expansion (d) estimates and for the post-stratified (p) estimates. Each of the eight original commodity designs (corn, hay, oats, soy = soybeans, wht=wheat, catl=cattle, hogs, and shp=sheep) were included in the study. The composite estimates over both groups (comp d and comp p) are also indicated. The

Illustration 1: Hog Estimates and Confidence Intervals



composite weights are inversely proportional to the estimated variances. A Taylor's Series approximation was used to estimate the variances of the post-stratified estimates. The individual variances were treated as constants when estimating the variances of the two composites.

The illustration shows the pseudo design-based direct expansion estimates have large variances and tend to be biased upwards compared to the post-stratified estimates, which are relatively consistent. This bias is due to an overrepresentation of large agricultural operations among the 11,178 records which were sampled for this analysis. Specifically large hog operations contributed to the biases shown in Illustration 1. The direct expansion from the hog design is not affected by this overrepresentation since the hog sample is stratified by hog control data. Although this is an artificial data problem unique to this data set, the robustness of the post-stratification approach is apparent.

The confidence intervals for the post-stratified estimates of hogs are much smaller than for the direct expansions. The largest reduction is from the original oats design where the confidence interval for the post-stratified estimate is about 10% as large as the direct expansion confidence interval. The resulting approximate confidence interval for the composite of the post-stratified estimates is about 40% as large as for the composite of the direct expansion estimates. The estimated CV of the post-stratification composite is 2.3 compared to an estimated CV of 5.7 for the direct expansion composite.

Discussion and Conclusions

NASS is interested in applying composite estimation to data collected from the County Estimates Survey to improve State, district, and county level estimates. State level estimates for the list frame domain could possibly be used in conjunction with list frame estimates from the probability QAS. This would strengthen the USDA official estimates of various commodities and make full use of the County Estimates data base.

Initial analysis presented in this paper indicates that a State level composite of post-stratified estimates from multiple stratified designs would provide more reliable estimates than a composite of direct expansion estimates. The post-stratification approach exhibited robust characteristics and may also help address nonresponse bias due to the large nonresponse problem in the County Estimates Survey. The variance approximation of the composite estimate needs to be further evaluated.

This composite estimation approach at the district level should also be evaluated. Reliable district estimates benefit the county estimation process since county values must add to the district. The variance approximation of the composite at the district level, which is based on a much smaller data set, also needs to be closely evaluated.

References

Bass, J., B. Guinn, B. Klugh, C. Ruckman, J. Thorson, and J. Waldrop (1989), "Report of the Task Group on County Estimates," National Agricultural Statistics Service, U.S. Department of Agriculture.

Bethlehem, J.G. (1988), "Reduction of Nonresponse Bias Through Regression Estimation," Journal of Official Statistics, 4, 251-260.