Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

EVALUATING THE IMPACT OF ALTERNATIVE EDIT PARAMETERS ON DATA QUALITY

Katherine Jenny Thompson and Samson Adeshiyan[1]

ABSTRACT

This paper describes a test of two alternative sets of ratio edit and imputation procedures, both using the U.S. Census Bureau's generalized editing/imputation subsystem ("Plain Vanilla") on 1997 Economic Census data. We compare the quality of edited and imputed data -- at both the macro and micro levels -- from both sets of procedures and discuss how our quantitative methods allowed us to recommend changes to current procedures.

KEY WORDS: Ratio edit; tabulation; blind testing

1. INTRODUCTION

The U.S. Census Bureau conducts an economic census in years ending in 2 and 7, mailing out over four million census forms to business establishments that provide commercial services to the public and other businesses. For the 1997 Economic Census, the Census Bureau developed and used a generalized editing and imputation subsystem, called Plain Vanilla (PV). The PV edit subsystem consists of three separate (generically written) edit and imputation programs: a ratio edit module; a balance edit module; and a verification module. Program areas customize these programs by developing edit script files that describe how PV processes a particular program's edits. Examples of script file functions include listing ratio edit items and associated reliability weights, providing imputation model sequencing in specified order of preference, and describing balance edits.

For the services sectors portion of the Economic Census, the use of PV in 1997 was a significant change in editing and imputation methods from those used for their previous censuses. In particular, the ratio edit module's methodology was quite different (see Section 2). Consequently, at the conclusion of their 1997 production processing, we conducted a quality audit of each services sector's PV implementation. Based on the audit results, we recommended several modifications to the production ratio edit procedures. Additionally, we recommended developing separate edit scripts for each trade area; for 1997, one production script was used to process all five services-sector trade areas.

To evaluate the effect on data quality using these two alternative edit implementations, we conducted a test on a subset of industries and basic data items using data from the 1997 census. For this test, we developed new edit scripts for each trade area using the ratio edit and item imputation methods recommended by the quality audit (Thompson et al, 2001), then submitted test decks for these industries through both the revised (new) scripts and the original (production) scripts (hereafter referred to as the "new" and "old" scripts). After processing, we had three competing values for each edited data item in each industry: the final published value of the item in the production database (assumed "correct"); the old-script-imputed/old parameter value; and the new-script-imputed/new parameter value.

This paper describes how we compared the quality of edited and imputed data -- at both the macro and micro levels B from the 1997 production processing to that from the modified procedures recommended by the quality audit. Section 3 provides background on the evaluation study. Section 4 presents the methods used to compare

[1] Economic Programs Directorate, U.S. Bureau of the Census, Washington, D.C, U.S.A, 20233. The authors would like to thank Scot Dahl, Ruth Detlefsen, Nash Monsour, Kenneth Sausman, Richard Sigman, and Michael Walkup for their valuable contributions to earlier versions of manuscript. This paper presents the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties and to encourage discussion of work in progress.

the macrodata (tabulations) and the results of these comparisons. Section 5 presents the microdata review methodology and associated results. Section 6 discusses these results, and Section 7 provides our conclusions.

2. PV RATIO EDIT MODULE METHODOLOGY

A ratio edit compares the ratio of two highly correlated items to upper and lower bounds, called tolerances. Reported items that fall outside of the tolerances are considered edit failures, and one or both of the items in an edit-failing ratio are either imputed or flagged for analyst review. The PV ratio edit module utilizes the Felligi-Holt model of editing which means that the complete set of ratio edits is tested **simultaneously**. The program determines the minimum number of reported data fields that need to be changed to satisfy the complete set of edits (Greenberg, 1986). The complete set of edits is defined as the user-specified edits provided in the script (the explicit edits), plus the other ratio tests **implied** by the explicit set. [Note: any pair of ratio edits with a common data item implies another ratio edit]. This methodology has been used successfully at the Census Bureau by other economic programs since the early 1980s (Greenberg et al, 1990).

For example, each of the services-sectors collect data on annual payroll (APR), 1st quarter payroll (QPR), and employment (EMP). To guarantee that the imputed value of APR is never smaller than the imputed value of QPR and that the ratio of APR to QPR is never “far from” the industry average of four, the edit-developer specifies that $1 \leq APR/QPR \leq 6$. Since employment is usually a good predictor of annual payroll, an edit developer defines an explicit test between those two variables with industry-specific tolerance limits ($LB \leq APR/EMP \leq UB$). These two tests imply a third relationship, namely 1st quarter payroll to employment tested by $LB/6 \leq QPR/EMP \leq UB$. The edited/imputed record must satisfy **all three** edits.

Reliability weights for each item can be used to influence the probability of deleting a given data field, with lower weights indicating **higher** reliability. Failure counts for each edit-failing item are multiplied by its reliability weight, so that minimizing the number of items to be deleted is equivalent to maximizing the weighted failure count.

Missing reported data items are automatically imputed by the PV ratio edit module. If only one data item is reported, it is not edited (two non-missing data items are required for a ratio edit). However, a complete record is imputed from the single (unedited) reported item.

3. BACKGROUND

The services sectors portion of the Economic Census is a mail-out/mail back census that comprises five trade areas: Retail Trade; Wholesale Trade; Service Industries; Transportation, Communication, and Utility Industries (Utilities); and Finance, Insurance, and Real Estate (FIRE). Data are collected on approximately one hundred fifty different industry-specific questionnaires. Some trade areas further classify the establishments within industry by legal form of organization, type of operation, and tax status. We used these editing-processing classifications for our evaluation, but refer to each classification as an industry.

Trade area subject-matter-experts provided the industries used for this test. These industries were selected because they were particularly problematic in 1997 and were not meant to be representative of the trade area as a whole. A “side effect” of this criterion was that some of these industries could be very intractable in terms of edit and imputation parameter development. We had a small number of industries per trade area: four in Retail; 14 in Wholesale; seven in Services; four in Utilities; and four in FIRE. We performed our evaluation by industry within trade area and used only active full-year reporter records. The data items used in this study varied slightly by trade area. Besides annual payroll, 1st quarter payroll, and number of employees, all trade areas collect sales/receipts (SLS). In addition, Wholesale collects operating expenses, purchases, beginning inventories, and ending inventories, and Services collects operating expenses in tax-exempt industries.

There are two key differences between the sets of ratio edits employed by the new and old scripts. First, the old scripts contained more ratio-tested variables, including trailer data and administrative data tests. Second, the old scripts provided tolerances for the complete sets of ratio edits (explicit and implicit), resulting in very tight edit-acceptance spaces. The new scripts dropped all trailer data tests (often, these data items were poorly correlated with the basic data items, causing edit failures/imputations based on tenuous relationships) and specified a very limited set of explicit ratio edits (APR/QPR, APR/EMP, and SLS/APR in all trade areas, with a few additional tests in one Services industry and in Wholesale Trade). Although we recommended including administrative data ratio tests in the new scripts, they were not included in our test scripts because of operational concerns. Different item reliability weights were used (for the same items) in each script. Finally, there was some discrepancy in edit and imputation parameter quality. The old-script parameters had undergone several revisions; while our parameters were reviewed once. Consequently, we viewed equally good results from the old and new scripts as evidence of improved methodology in the new scripts.

4. MACRO-LEVEL EVALUATION (TABULATION COMPARISONS)

Our first set of analyses compares data item tabulations from the old and new edit results to the tabulations based on final 1997 publication data (our "gold standard"). We did this by producing three data item tabulations within industry/establishment size category (small, large, total): one per script (new/old), plus the tabulation from the final edited data. We then computed ratios of the old and new script tabulations to the final data tabulation, selecting the "better" tabulation as the one with the ratio closer to 1. When **both** ratios were within five-percent of the final value, then the two scripts tied. Table 1 summarizes our data item comparisons for **total** establishments summed over industries within a trade area. The small and large establishment tabulations showed similar patterns.

Except for Retail, the new edit script tabulations are generally closer to the final published values than the corresponding old scripts tabulations for all data items, often by a wide margin. The two inventory items in Wholesale Trade are notable exceptions: both scripts performed equally poorly. In the six industries that collected inventories data, the average ratio of new script to final beginning inventories and new script to final ending inventories were 13.3 (min = 1.8, max = 35.8) and 15.4 (min = 1.9, max = 54.9) respectively, and the average ratio of old script to final beginning inventories and old script to final ending inventories was 12.03 (min = 2.03, max = 35.9) and 13.4 (min = 1.9, max = 54.9) respectively. Inventories items are difficult to ratio edit; although they have high pairwise correlation, they are poorly correlated with the other basic data items.

Table 1: Data Item Comparisons

Trade Area	Data Item	New Better	Old Better	Tie
FIRE	Sales	3	1	0
	Annual Payroll	3	1	0
	1st Quarter Payroll	3	1	0
	Employment	4	0	0
Retail	Sales	2	2	0
	Annual Payroll	2	2	0
	1st Quarter Payroll	2	2	0
	Employment	2	2	0
Services	Sales	2	2	3
	Annual Payroll	2	1	4
	1st Quarter Payroll	3	1	3
	Employment	3	1	3
	Operating Expenses	0	1	0
Utilities	Sales	2	1	1
	Annual Payroll	1	1	2
	1st Quarter Payroll	1	1	2
	Employment	2	0	2
Wholesale	Sales	9	5	0
	Annual Payroll	8	4	2
	1st Quarter Payroll	7	2	5
	Employment	9	1	4
	Operating Expenses	6	5	3
	Purchases	6	0	0
	Beginning Inventories	2	4	0
	Ending Inventories	3	3	0

Next, we compared total results by industry within trade area. In each industry, we summed up our ratio classifications (new better/old better/tied) for each data item to get an industry-level determination.

Table 2 summarizes the industry level comparisons for **total** establishments. Again except for Retail, the new scripts generally produce results closer to the final tabulated results than the old scripts. Unfortunately, in the two Retail industries where the

sets of edits in terms of effect on the tabulations. However, large establishments are very influential in this type of comparison. Although all Economic Census forms are machine edited, analyst review of edit-failing cases is generally restricted to large establishments because of time-constraints. Subsequent stages of data review usually focus on large establishments because they most impact the tabulations. So, a limitation of the macro-level evaluation is that while it does well on the whole and for large establishments in particular, it does not necessarily do well for detecting edit problems with small establishments.

Given these limitations, we did not want to use the published microdata for our micro-level comparison. Moreover, the subject-matter-expert analysts wanted to review the microdata from both edit scripts. So, we conducted blind testing in all of our industries. The original motivation for the blind testing was to (hopefully) make the analysts comfortable with the revised procedures on a case-by-case basis. We also planned to use blind test results to either confirm the macro-level results presented in Section 4 or to uncover systematic edit/imputation problems for certain classes of establishments (e.g., small establishments).

For the blind test, analysts from each trade area were provided with basic information for each edited data item[2] for 200 randomly selected cases (100 cases per size category) per trade area and were asked to select which -- if either -- edit outcome (edit A or edit B) was acceptable. The label for edit A and edit B was randomly assigned, so that neither the analysts nor the evaluators knew which script was used to obtain either outcome. To avoid potentially biasing the outcome, analysts were **not** to be able to identify a particular establishment. They also were not given any edit flags. In effect their data review tools were more limited than they would be in a production system. Also, the tabulated results described in Section 4 were not provided until blind testing was completed. Analysts were asked to review at least 50 of the 100 cases per size category.

We were interested in gaining insight into two following questions. First, at the micro-level, did the analysts have a preference for one script outcome over another? Second, were the results from the micro-evaluation (analyst preference) consistent with those from the macro evaluation? And if not, can we explain the reasons for the differences and what can we do to correct these differences in the future? We addressed the first question using the standard categorical analyses presented in Section 5.2 and examined the second set of questions by conducting an exploratory review of the records in which the analysts clearly preferred the old script results.

Our ability to analyze the blind test results was greatly handicapped by the sample design. Ideally, we would have selected a sample that was stratified within industry and establishment size cell by “magnitude of edit difference.” Also, our population would have been the set of all establishments whose edit results (new script/old script) contained at least one basic data item changed by either script. Because of processing timing concerns, we were not involved in the sampling plans. Instead, the analysts were provided with a stratified random sample of establishments selected within trade area (not industry) and size category (large, small) that had at least one imputed data item from either script that differed from the published value. Consequently, the blind test data contained a high percentage of cases with nearly identical edit outcomes.

5.2 Statistical Analysis Tools: Tests for Association and Analyst Preference

Test 1. Test for Association

We first tested whether the analysts have a preference between the two edit scripts with

H_0: Choice of the new or old edit is independent of each other (i.e., no preference between edits).
H_1: Choice of the new or old edit is dependent on each other (i.e., one type of edit is preferred over the other).

using the standard Pearson chi-squared test (Agresti, 1990) with the count data shown in Table 3.

[2] Reported value, administrative value (if available), 1992 census edited value and associated edit flag, and the two different script edit outcomes.

Table 3: Tabulation of Blind Testing Data

		New PV Edit		
Old PV Edit		Acceptable	Not Acceptable	Total
	Acceptable	Both Acceptable (N_{11})	Only Old Acceptable (N_{12})	N_{1+}
	Not Acceptable	Only New Acceptable (N_{21})	Neither Acceptable(N_{22})	N_{2+}
	Total	N_{+1}	N_{+2}	N_{++}

Rejecting the hypothesis of independence allows us to conclude that the analysts tend to prefer one edit over the other, but it does not tell us **which** edit is preferred. To determine the preferred edit, we focused on the highlighted cells in Table 3, where the analyst made a clear choice: "Only New Acceptable" (N_{12}) or "Only Old Acceptable" (N_{21}). Formally, we tested

<u>**Test 2.** Test for Analyst Preference</u>

H_0: $p_{21} <= p_{12}$ (or p_{21}-p_{12}<=0, i.e., the proportion of "only new acceptable" is less than or equal to the proportion of "only old acceptable.")

H_1: $p_{21} > p_{12}$ (or p_{21}-p_{12}>0, i.e., the proportion of "only new acceptable" is greater than the proportion of "only old acceptable.")

where proportions $p_{12} = N_{12}/N_{++}$ and $p_{21} = N_{21}/N_{++}$. The difference in proportions is estimated by ($p_{21} - p_{12}$), and the standard error (se) is estimated by $se(p_{21} - p_{12}) = [(p_{21}(1-p_{21})/N_{++}) + (p_{12}(1-p_{12})/N_{++}) - (2p_{21}p_{12}/N_{++})]^2$, yielding a test statistic distributed as $t_{\alpha,N_{++}-1}$. Rejecting H_0 provides evidences that the analysts prefer the new edit script over the old edit script. Since this is a one-tailed test, any t-statistic with negative value implies that we cannot reject H_0.

5.3 Results

5.3.1. Summary Data

Table 4 presents the counts of the blind test data by trade area. As expected, "Both Acceptable" was the most common choice. Wholesale had a high percentage of "Neither Acceptable" cases, likely attributable to the poor inventory edit results (see Section 4).

Table 4: Counts from Analyst Review By Trade Area

Analysts' Choice	Trade Area				
	FIRE	Retail	Services	Utilities	Wholesale
Both Acceptable (N_{11})	137	294	509	176	765
Only New Acceptable (N_{21})	73	115	183	73	263
Only Old Acceptable (N_{12})	149	55	276	92	116
Neither Acceptable (N_{22})	45	41	82	13	744

Table 5 provides the results of our tests for association (Test 1) and analyst preference (Test 2) using the count data shown Table 4. Except for Services, Table 5 provides evidence of association in analysts' old and new edit choices at the 5% significance level. For FIRE and Utilities, the old edit is preferred to the new edit; for Wholesale and for Retail, the new edit is preferred to the old edit; and we were unable to make a conclusion about direction of preference for Services.

These results are very interesting because they appear to conflict with the macro-level results discussed in Section 4. However, before interpreting these results, we wanted to confirm

Table 5: Results of Tests for Association and Analysts Preference

	Test 1	Test 2	
	Reject H_0	t-statistic	Reject H_0
FIRE	Yes	-4.75	No
Retail	Yes	4.00	Yes
Services	No	-4.50	--
Utilities	Yes	-1.25	No
Wholesale	Yes	7.00	Yes

that these test results were truly indicative of the analysts' preferences and not merely a function of a poor sample. For example, Table 5 provides evidence that the analysts preferred the new Retail script. In two Retail industries, the new script was clearly better than the old script at the macro-level (Table 2). If the majority of Retail "Only New Acceptable" (N_{21}) cases were sampled from those two industries, then the blind test results would actually be consistent with the macro-level results. Further examination of the industry distribution of sample data was required.

Table 6 post-stratifies the blind test cases where analysts clearly preferred one script over the other (i.e., the N_{12} and N_{21} cases) by the Table 2 industry-level classifications. Similar distributions of N_{12} and N_{21} counts in Table 4 and Table 6 distributions would provide evidence that the analysts could predict the macro-level results from their micro-review (i.e., that the consistently preferred script would have the better tabulated results). Dissimilar results are more difficult to interpret. For example, if the difference in both sets of macro-level results quality was caused by a few establishments (none of which were included in the blind test), then the sampled cases in an industry characterized as "Old Script Better" in Section 4 could actually have the same or even more consistent results with the new script.

Table 6: Distribution of N_{12} and N_{21} Cases by Macro-Evaluation Industry Classification

Trade Area	New Script Better from Macro Evaluation	Old Script Better from Macro Evaluation	Both Old & New Scripts Tied
FIRE	66.22%	0.00%	33.78%
Retail	39.41%	61.59%	0.00%
Services	39.00%	30.50%	30.50%
Utilities	23.64%	17.58%	58.79%
Wholesale	58.84%	20.58%	20.58%

For Retail, only 39.41% of the cases where analysts clearly preferred one script (N_{21} and N_{12}) were selected from industries where the new script did better overall. However, the analysts preferred the new script over the old script in approximately 68% of the N_{12} and N_{21} cases (115/115+55). In the other trade areas, a high percentage of these N_{21} and N_{12} cases were sampled from industries where both the new and old scripts tied. In these cases, it is tricky to draw parallels between analyst preference and macro-level results: for example, analysts might have preferred the edit that preserved more reported data or might have a preference for changing the value of one data item over another. With the Utilities data, the majority of N_{21} and N_{12} cases are from industries where both scripts tied, making it impossible to draw any parallels between macro- and micro-level results. For both FIRE and Wholesale, most of the cases came from industries where the new scripts did better. The majority of FIRE's test cases (66.22%) were selected from industries where the new scripts did better, but the analysts tended to prefer the old script. The majority of Wholesale's test cases (58.84%) were selected from industries where the new scripts did better, and the analysts also tended to prefer the new script.

5.3.2 Exploratory Review of Analysts' Preference

We found the apparent contradiction between macro-level and the micro-level results perplexing. Tabulations from section 4 showed marked improvements in edit outcome with the new scripts, but the blind test results showed that analysts tended to prefer the old script results at the micro-level. To understand this preference, we conducted an exploratory review of the records where the analysts clearly preferred the old script results (N_{12} cases). This led to two major findings (both confirmed by the analysts). First, the analysts usually preferred the script that changed fewer reported values or used administrative data for imputation, even when final edited data contained unusual ratios (e.g., a payroll to 1st annual payroll ratio of 12, far from the industry average of four). Second, there were several cases where analysts failed to provide sufficient trade-area tolerance limit rules. For example, based on the analysts' specifications, the new script tolerances guaranteed that sales had to be greater than or equal to annual payroll. When reviewing the blind test results, we learned that the FIRE analysts, in addition, preferred that sales should not exceed five times the annual payroll.

6. DISCUSSION

When we originally planned this evaluation study, our goal was to prove that the PV ratio module - if properly implemented - could achieve excellent edit results for the services-sectors portion of the census with little or no human intervention. For the most part, we succeeded. We are not finished, however. This evaluation revealed two major systematic problems - common to all trade areas - with the new edit scripts: poor detection of rounding errors and an increased probability (compared to the old script) of imputing a complete record from a single unedited data item. We will reduce occurrences of these edit problems in the 2002 Economic Census by including ratio tests to administrative data in our edit scripts. We will also safeguard against these situations by data-filling blank items with other reported sums of details (and/or administrative data) and correcting rounding errors prior to ratio-editing.

We were initially disappointed by apparent contradictions between the macro-level and blind test results. We knew that the new procedures improved overall data quality, but the analysts reviewing the blind test data concluded differently. The deficiencies in the sample itself made it difficult for us to understand the implications of the blind test results. Even so, the blind testing was a useful analysis tool. First, it revealed a “disconnect” between analyst preference and industry-level ratio requirements (e.g. preferring edit outcomes that retained more reported or administrative data, even if results contradicted industry level tolerances). And, by reviewing the cases where the analysts accepted only the old script results (N_{12}) cases, we found some systematic problems with edit parameters.

We view the blind test as a “dress rehearsal” for the production micro-review. The analysts are clearly knowledgeable about their subject-area. They are not, however, as knowledgeable about ratio edit implementation. We must address this by developing training that conveys the connection between ratio edit tolerances and edit outcomes so that the analysts can build all program requirements into their 2002 edit scripts.

The macro-level analyses described in Section 4 were quite effective at both evaluating the edit results overall and indicating systematic edit implementation problems (especially when combined with a micro-review of “problem” records). We do not really expect the analysts’ micro-review to be as revealing. Micro-review checks individual records for internal consistency, not for conformance with the industry norms. Usually outliers cannot be detected in the absence of the full distribution. Recognizing this, the 2002 PV edit implementation should include ongoing summary audits of ratio edit failures within industry to reveal problem ratio edits or tolerances, rather than relying on the analysts ability to recognize and articulate such problems.

7. CONCLUSION

This paper describes a test of two alternative sets of ratio edit and imputation procedures on data from the service-sectors portion of the 1997 Economic Census: the production processing procedures and a modified set of procedures resulting from a quality audit. We showed that the modified procedures resulted in improved edited data quality. Moreover, the evaluation process revealed further necessary enhancements to the modified procedures, which will be implemented in the 2002 census.

A less measurable - but equally important - deliverable from this evaluation study was the dialog between edit-implementor (methodologists) and subject-matter experts. Discussing both the macro-level results and the blind test results began a long-overdue edit-implementation training process. As a consequence of this study, we are establishing workgroups that consist of methodologists, production specialists, and subject-matter-experts to develop edit parameters and scripts for the 2002 census. Also with the work groups in place, we can develop training courses for the analysts that address the deficiencies revealed by this evaluation.

REFERENCES

Agresti, Alan (1990), *Categorical Data Analysis*, New York: Wiley.

Greenberg, B. (1986), “The Use of Implied Edits and Set Covering in Automated Data Editing,” Unpublished report, Washington, DC: U.S. Bureau of the Census.

Greenberg, Brian; Draper, Lisa; Petkunas, Thomas (1990), “On-Line Capabilities of SPEER,” *Proceedings of the Statistics Canada Symposium*, *Statistics Canada*, pp. 235-243.

Thompson, K., Sausman, K., Walkup, M., Dahl, S., King, C., Adeshiyan, S. (2001), “Developing Ratio Edits And Imputation Parameters For the Services Sector Censuses (SSSD) Plain Vanilla Ratio Edit Module Test,” unpublished report, Washington, DC: U.S. Bureau of the Census.