

Developing Ratio Edit Parameters with SAS® for the Economic Census

Richard W. Graham, Bureau of the Census, Washington, DC

Abstract

Developing ratio edit parameters for the editing of census data has always been an arduous task at the Census Bureau. Prior to each quinquennial census undertaking, subject-matter specialists and programming staffs worked countless hours developing specifications, computer programs and generating and reviewing edit parameter output. For the 1997 Economic Census a team of specialists endeavored to throw out the old technology of developing edit parameters and developed a SAS/AF®, SAS Screen Control Language (SCL), and SAS/Macro program for the creation of ratio edit parameters.

Introduction

In 1995, the Bureau of the Census commissioned a team of individuals with varying expertise and backgrounds to develop a generalized editing system for the Economic Censuses. The team consisted of subject-matter specialists, data processing and programming specialists, and mathematical statisticians. Out of that endeavor, the team developed the Economic Census Generalized Edit and Imputation System, a.k.a. Plain Vanilla (PV). The PV system consists of three main editing components which may be used independently or in various combinations. The three components are the ratio edit, the balancing edit and the verification edit. The team also developed various tools for developing editing and imputation

parameters, for integrating the generalized edit with the various legacy edit processing systems, and audit trail and audit information systems for evaluation purposes.

This paper gives a brief description of the Economic Censuses and the ratio editing concept. The paper then specifically details the methodological strategy used to develop the ratio edit parameters, including the SAS® products and procedures employed to develop these parameters.

The Economic Census

The Economic Census is a complete enumeration of businesses and their economic activity. Conducted quinquennially in years ending in 3 and 8 for the census years ending in 2 and 7, the economic censuses provide the most comprehensive source of information on virtually all sectors of the economy, approximate 98 percent of all economic activity is covered.

The Economic Census is comprised of three major divisions: manufacturing and mining; construction; and services. This paper focuses on the services division which is comprised of retail and wholesale trades; service industries; transportation, communication, and public utilities industries; finance, insurance and real estate industries; auxiliary establishments, and outlying areas (including manufacturing and construction).

Within the services division, data are collected for nearly 2,900 industry, type of operation, tax status and legal form of organization classification combinations. Data are collected by mail-out/mail-back procedures utilizing some 150 different questionnaires. Data collected include an SIC-based (standard industrial classification) industry classification - later converted to a NAICs-based (North American Industrial Classification) industry classification, type of operation, legal form of organization, tax status, sales/revenues/receipts - herein referred to as sales, payroll items, employment items, and numerous industry-specific items (e.g., gallons of gasoline, hotel rooms, restaurant seats).

In late 1997, the Bureau of the Census mailed over 4 million questionnaires to businesses covered by the Services Division. Returned questionnaires are microfilmed, data entry keyed, matched against the business register (Standard Statistical Establishment List - SSEL), and released to the subject-matter division for editing, tabulation and publication. During the editing phase, the data are subjected to a number of consistency, completeness, and validation checks to ensure the data's accuracy. One of the edit procedures to which the data are subjected is the ratio edit.

Ratio Editing

Ratio editing is used in the Economic Censuses as a primary tool for locating suspicious and aberrant data. This is accomplished by subjecting the data to relationship tests between data items. That is, each numeric data item is ratio compared against all other correlated data items and tested to determine if the relationship falls

within a prescribed set of boundaries. Relationships that fall outside of the relationship boundaries are then analyzed to determine which of the items is most likely causing the record failure. These identified items are then imputed.

The PV system's ratio module is a one-pass mathematically-based editing algorithm developed originally by Brian Greenberg (Greenberg, 1981). Greenberg's work was based on the work of Fellegi-Holt (Fellegi-Holt, 1976). Their theory stated that given a set of explicit conditions to which an edit record is subjected, one can derive additional implicit conditions such that a record must pass all conditions (both explicit and implicit) to ensure a complete and consistent record. Further, the theory specifies that given the complete set of editing conditions, when edit failures occur, the field, or fields, causing the error condition(s) may be systematically determined by examining those field(s) which are common to the edit-failure conditions.

The work of Greenberg led to the initial development of an automated generalized editing technique for quantitative (arithmetic) continuous data utilizing a ratio methodology between related data fields. Under this methodology, a set of explicit ratio boundaries (explicit edit conditions) is provided by subject-matter experts to define specific relationships between correlated items. From the supplied explicit ratio boundaries, implicit relationships (implied editing conditions) are developed among all interrelated fields which then become the basis for determining edit-failing records and for imputing items on the failed record.

The Plain Vanilla ratio module processes census or survey data against the complete set of editing conditions (ratio boundaries) established by the subject-matter specialists.

The module evaluates the survey data relative to the appropriate parameters, and on determining that certain edits have failed, will continue processing the record through an error localization routine which determines the minimum number of fields to delete and mark for adjustment. On determining those items requiring adjustment to bring about consistency vis-à-vis the edit parameters, the procedure continues with the imputation of deleted items.

Ratio Edit Parameters

The ratio edit module requires several inputs for defining what and how the census or survey data are to be processed. These inputs, known as parameter files, tell the system what items are to be edited, the criteria by which the data are to be edited, and the methods by which the data are to be altered or imputed. The ratio module requires a minimum of three parameter files—they are the script parameter file, the implicit parameter file, and the imputation parameter file.

The script parameter file identifies those items that are to be edited and defines the methodologies to employ for imputing missing and/or aberrant data. The script file defines 150 separate items that are to be ratio edited. However, it must be pointed out that not all items apply to each industry. That is, gallons of gasoline apply only to petroleum industries and gasoline service stations, hotel rooms only to hotels and casinos, etc.

The imputation parameter file provides the category averages (for census purposes this is based on the industry, type of operation, legal form of organization, and tax status classification combinations) used for imputing missing and/or aberrant data. (Throughout this paper, the term industry average will be used, but refers to the category average). The imputation file contains the industry code, the numerator and denominator mnemonics, the explicit lower and upper boundaries, the industry average, and an inflation factor. Although the file contains upper and lower boundaries, these boundaries are not directly used in the ratio edit. Rather, these explicit boundaries are used as input to a bounds generation program which develops the complete set (explicit and implicit) of editing boundaries. For the service division, the imputation parameter file contains nearly 138,000 parameter records.

The implicit parameter file is the at the heart of the ratio edit system. Although the script file defines the complete set of items to be edited by the ratio edit, the implicit parameter file serves as the basis for defining those items to be edited for a particular industry. The file provides the upper and lower tolerances against which each data pair is tested to determine its accuracy and validity. The file contains an industry code, a numerator and denominator sequence number, upper and lower boundaries, and the numerator and denominator mnemonics. For the service division, the implicit parameter file contains over 432,000 parameter records segmented by 2,866 classifications.

Explicit Ratio Edit Parameter Development Methodology

Along with the development of the ratio edit module, the team also developed some tools for developing ratio edit parameters. After investigating several methods, the team adopted a resistant fences approach to creating parameters and developed a SAS® software application to implement the methodology.

The resistant fences methodology is an exploratory data analysis (EDA) technique for outlier detection. This method defines outlier ratios from a set of rules based on quartiles. The quartiles are calculated by setting the cumulative probability level for the i^{th} order statistic $x_{(i)}$ equal to $i/(n+1)$, as recommended by Hoaglin and Iglewicz (1987), Thompson and Sigman (1996).

The methodology defines outliers as ratios whose values are less than $(q_{25} - k*H)$ or greater than $(q_{75} + k*H)$, where q_{25} = the first quartile (the lower fourth of the distribution), q_{75} = the third quartile (the upper fourth of the distribution), $H = q_{75} - q_{25}$ (the inter quartile range of the distribution), and k = a constant - the inner fences rule sets k = to 1.5, while an outer fences rule sets k = 3. For census parameter development purposes, we use a compromise rule, or middle fences rule, where k = 2.

One of the assumptions of the resistant fences methodology is that the distribution of ratios is symmetrical. Although symmetry is not required for the development of tolerances, symmetrizing a distribution brings about consistency with the main distribution those ratios initially

appearing as outliers in an unsymmetrized, or skewed, distribution.

Most ratio distributions are positively skewed, generally approaching a Chi-square distribution. For economic census ratio data, we also find this to be true, generally in a highly skewed sense. As such, we use an EDA approach for symmetrizing skewed distributions because each set of ratio distributions display a unique behavior in its form of skewness. Although a general form for symmetrizing a distribution exists in taking the natural logarithm of the ratios, this approach is not always appropriate (e.g., when developing boundaries or tolerances for a distribution of percent change containing negative values).

The EDA approach we use for determining the "appropriate" symmetrizing (transformation) factor for a skewed distribution of ratios is defined by Hoaglin et al. (1983) and modified by Thompson and Sigman (1996). This approach (the power transformation) uses a transformation plot for symmetry by placing

$$\begin{aligned} x_{vi} &= [(x_{ui} + x_i)/2] - M && \text{on the vertical axis (v),} \\ &\text{and} \\ x_{hi} &= [(x_{ui} - M)^2 + (M - x_i)^2]/[4/M] \\ &\text{on the horizontal axis (h), such that} \\ \text{slope}_i &= x_{vi}/x_{hi}. \end{aligned}$$

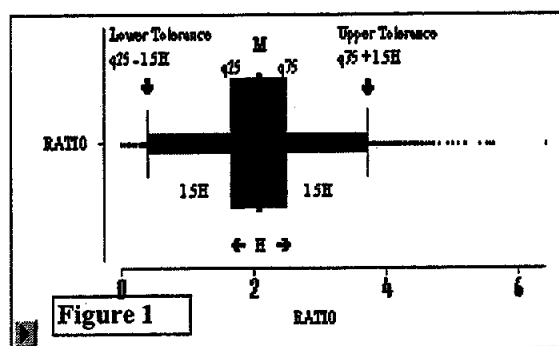
If the resultant graph is nearly linear (i.e., $\text{slope} \approx c$), then $p = 1 - c$, where p is the power transformation factor used to transform the distribution.

Rather than graphing all of the ratio distributions for each of the 2,866 industries, we determined a value of p by setting this value equal to $1 - \text{median}(x_{vi}/x_{hi})$. Although

Hoaglin recommends rounding p to the nearest power of $\frac{1}{2}$, we did not employ this rounding methodology.

Our approach then for determining the appropriate symmetrization factor was to first determine p , apply the natural logarithm transformation to the same ratio distribution, then select the transformation resulting in the smallest absolute value of the skewness coefficient. This methodology also used the untransformed distribution in the comparison of the skewness coefficient and occasionally resulted in the selection of the non-transformed distribution for setting tolerances.

Once the power transformation factor is determined (note: p may be 1 indicating selection of an untransformed distribution), we then apply a resistant-fences rule to the "symmetrized" distribution to obtain initial bounds, and then apply the inverse power transformation to the initial bounds to obtain the initial tolerances or upper and lower explicit boundaries. Figure 1 depicts a graphical illustration of this methodology.



To develop the ratio tolerances, Thompson (1996) of the PV Edit Team, developed a SAS® application using the SAS Macro Language and various Base SAS procedures

to implement the above methodology. Within that program the following processes occur:

1. Test the distribution of ratios for statistical skewness,
2. Determine power transformation to symmetrize the distribution, if skewed,
3. Apply a resistant-fences rule to the symmetrized data to obtain initial bounds,
4. Apply the inverse power transformation to the initial bounds to obtain tolerances,
5. (Optional) Compare lower tolerance value to user-specified allowable minimum value.

Ratio Edit Parameter Development

Developing initial ratio edit parameters was quite an arduous undertaking. The first step in the development was to obtain the data that would ultimately feed into the resistant fences program. This entailed the collection of 1992 census data from several sources. In 1992, each of the five trade areas had its own separate database which was comprised of a base table containing identifier information and general statistics items (i.e., sales, payroll, employment, inventories, and operating expenses) and a trailer data table containing industry-specific items (e.g., hotel rooms, gallons of gasoline, etc.).

The second source of data was the Standard Statistical Establishment List (SSEL) which is used as the control file for the census and the source from which we obtain Internal Revenue Service administrative data. This source also is comprised of several databases split between multiunit companies and single unit businesses. The records obtained from this source were then matched with the 1992 census records.

As mentioned, there were five trade areas for which data were collected in 1992. These were the wholesale trade; retail trade; service industries; finance, insurance, and real estate (FIRE) industries; and the transportation, communication and utilities industries. From the wholesale trade, we obtained some 470 thousand records; from retail, 1.5 million records; service, 2 million records, FIRE, 500 thousand records, and utilities, 220 thousand records.

To create SAS® data sets, we used SAS/Access software, version 6.09, to connect to the databases. The databases resided on a DEC VAX 7620. Early on we learned that only one database could be opened at a time using SAS/Access software. As such, our strategy was quickly revised from opening three databases, the trade area database, the single unit SSEL database, and the multiunit SSEL database, to opening one at a time. Fortunately, the trade area databases contained both tables containing the bulk of the data. Nevertheless, once an initial 1992 census data set was created, each file was matched to the SSEL databases to extract the appropriate administrative data.

The next step in the process was defining those items to be ratio edited via the ratio module. In previous censuses, the services area only ratio edited the general statistics items mentioned above. Trailer data items were edited using special coding techniques.

For 1997, we selected 150 data items to be edited within the ratio module, although on an industry-by-industry basis generally no more than 35 items apply. For all industries, the following four core items apply: sales, annual payroll, first quarter payroll, and

employment. Within the wholesale area the following additional three items apply, beginning and ending inventories and operating expenses; and for service tax exempt industries, operating expenses was a common item. The remaining items pertained to specific industries. After defining which of the items were to be edited by the ratio module, explicit ratio pair definitions were developed. Over the course of parameter development, we defined 555 separate explicit ratio pairs for which upper and lower boundaries and industry averages would be generated using the resistant fences parameter development software.

The maiden production process of the resistant fences program was run on a DEC VAX 6400 using the utilities data set. This set contained nearly 220 thousand records and 173 variables and would be segmented initially by 279 industry classifications and calculating 85 separate ratios for each industry. Initially, the resistant fences program did not utilize indexes on the data set and resulted in a processing run lasting some 30 hours before it was system terminated due to memory constraints. The developer was consulted and modifications were made to the program which included the use of indexes on the data set. The second run, again on the VAX 6400, completed in 15 hours.

Using the performance statistics from this completed run, we extrapolated these figures to provide us with performance expectations with respect to the services file. The services file contained nearly 10 times as many records, two million records with 162 variables, segmented by 479 industry classifications, and calculating 67 separate ratios for each industry. Our extrapolation

suggested at least 15 days processing on the VAX 6400.

Given these prospects, and the general requirement that these procedures would be run several times, we moved the entire process over to an SGIXL running SAS 6.11. The maiden run on the SGI was the same as the second run on the VAX, however, the performance was drastically improved. Running on this platform yielded performance efficiency by a factor of 4. That is, the entire utilities file was processed in three and a half hours. When it was time to process the service file, the performance on the SGI bore out our extrapolations. The service file completed in three solid days of processing time.

As mentioned, our expectation was that we would not just run this process one time against the data sets and be done with them.

Rather, initial parameter development was an iterative process requiring several runs of the process against each data set. Although there is no strict requirement that a certain number of records containing each of the paired items must be present, the developer instituted a minimum of 16 records per ratio per classification. And through our own study and decision, instituted a minimum of 30 records per ratio per industry. As such, a number of industries had fewer than the required minimum and were subsequently combined with other like industries.

The initial program while fairly simple to use, was cumbersome in providing all of the input variables to run the procedure. For each trade area, the program had to be modified to include appropriate LIBNAMEs, input and output files and data

sets. This being an unwieldy process lent itself to the development of an SAS/AF front-end application.

The SAS/AF application consists of four screens. The first screen solicits such information as host system (platform) on which the application will be run, physical location of data sets and input and output files, and selection of data set to use. Regarding host systems, the application provides the user with access to four platforms -- DEC/ALPHA via SAS/Connect, SGI UNIX -- via SAS/Connect, DEC UNIX -- via SAS/Connect, and the PC platform (See Figures 2 and 3 and following code).

Figure 2

Figure 3

```

PLATFORM:
  call notify('PLATFORM','_get_last_sel_',row,issel,text1);
  if text1='UNIX?SGI' then do;
    submit continue;
    options comamid=TCP remote=SGIXL;
    filename rlink "c:\sas\connect\saslink\tcpunix.scr";
    signon;
    endsubmit;
    REM_OUT2='Y';
  end;

  else REM_OUT2='N';
  cursor inlib2;
RETURN;
INLIB2:
  if inlib2 ne '' then do;
    if text1 = 'UNIX/SGI' then do;
      submit continue;
      libname in remote "ainlib2" server=SGIXL;
      rsubmit;
      libname in "ainlib2";
      run;
      endrsubmit;
    endsubmit;
  end;
RETURN;

```

Upon selection of the non-PC platforms, the user is prompted for a user id and password for logging onto the host platform. The user is then requested to identify the physical location of the input data set, select a data set once a LIBNAME has been created, identify the physical location of the output data set and file names, and solicits the location and filename containing the appropriate defined explicit ratio definitions.

The second screen then requests whether the output data sets are permanent, if all three industry average statistics are to be generated, and if summary and univariate statistics are to be printed. (See Figure 4).

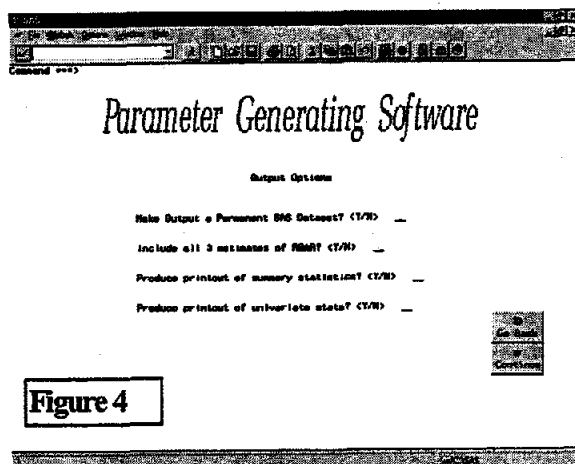


Figure 4

The third screen of the application displays within a listbox all of the variables within the data set. From the listbox, the user is requested to select a classification variable (in this application, the user would select the corresponding industry classification variable). Once the variable is selected, the application requests whether to index the data set - if already indexed, the Y is ignored. The next request pertains to the use of historical symmetrization factors. The resistant fences program provides for using historic symmetrization factors rather than recalculating these factors each time the program is run (See Figure 5).

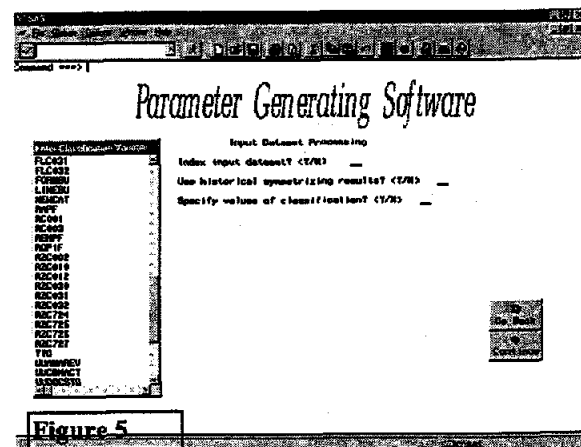
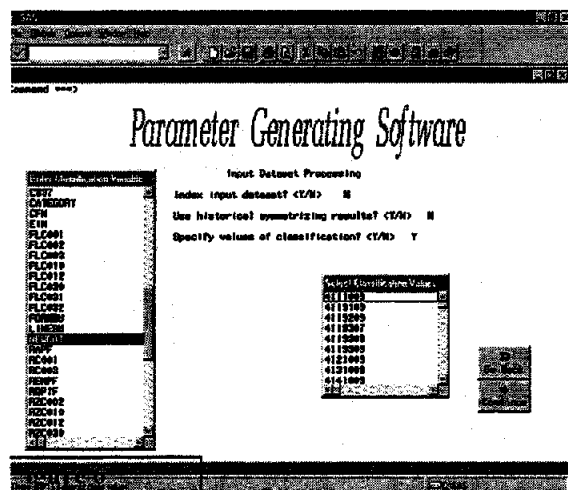


Figure 5

The user is then requested to specify values of the classification variable for which bounds are to be created. If the user selects no, boundary parameters for each unique classification value will be generated. If the user selects yes, a listbox of all the unique classification values is displayed (See Figure 6). The user may select one or more classification values.

The final input screen (Figure 7) solicits information regarding the k constant to be



used in the analysis, whether to symmetrize the distribution, whether to compute industry averages only, the type of average to compute, what action to take when a negative lower bound is encountered, the minimum number of observations to calculate bounds for each ratio, and whether the process is to be run in batch or interactively. Upon completion of inputs, the execute button is ungrayed. Clicking on this button launches the resistant fences program on the selected platform. Behind the scenes, as information is being input, an ASCII file is created. Once all the information is input, the file is then uploaded to the host platform and the program launched, as illustrated in the following code.

```
fn = 'c:\globfile.dat';
rc = filename('globfile',fn);
rc = fdelete('globfile');
FID = fopen('globfile','O');
RECD = '%let inlib = '||inlib2||';;';
rc = fput(FID,rc);
rc = fappend(FID);

RECD = 'libname out '||outlib2||';;';
rc = fput(FID,rc);
rc = fappend(FID);
rc = fclose(FID);
if text1 = 'UNIX/STE' then do;
```

```
submit continue;
options remote=steps4;
rsubmit;
data _null_;
call system('rm /users/graha005/globfile.dat');
run;
proc upload infile = 'c:\globfile.dat'
outfile = '/users/graha005/globfile.dat';
run;
endsubmit;
endsubmit;
end;
```

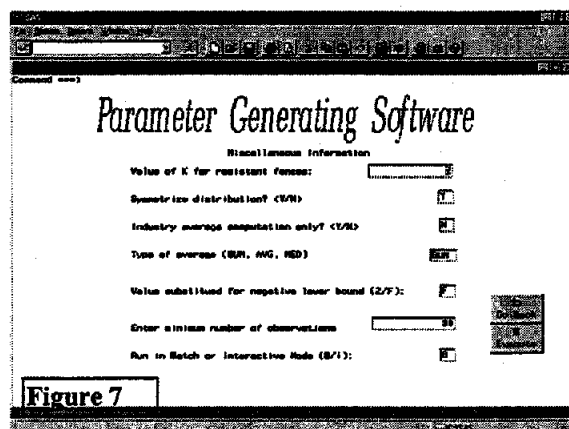


Figure 7

Conclusions

The frame- and macro-based programs described in this paper are fully functional SAS applications. These applications have been developed with an eye toward generalization. The applications may be run against any data set containing numeric data requiring some sort of ratio edit methodology.

The parameter generating interface program and the resistant fences program code will be available at the conference.

SAS, SAS/ACCESS, SAS/AF, SAS/CONNECT, SAS/INSIGHT, and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

® indicates USA registration.

References

Amadeus Software Ltd.(1997), Building Frame Entry Applications in SAS® Software. Oxfordshire, UK

Destiny Corporation (1996) Macros is SAS® Software, Wethersfield, CT.

Fellegi, I.P. and Holt, D. (1976), A Systematic Approach to Automatic Editing and Imputation. Journal of the American Statistical Association, 71, 17-35.

Graham, Richard and Anderson, Tammy (1991), Imputation Parameter Study and Specifications for Parameter Development. Bureau of the Census: Services Division. Internal documentation.

Greenberg, Brian (1981), The Use of Implied Edits and Set Covering Approach in Automated Data Editing. Washington, DC: Bureau of the Census Technical Report Census/SRD/RR-86/02.

Hoaglin, D.C. and Iglewicz, B. (1987), Fine-Tuning Some Resistant Rules for Outlier Labeling. Journal of the American Statistical Organization, 83, 1147-1149.

Hoaglin, D.C.; Mosteller, F.; and Tukey, J.W. (Eds.) (1983), Understanding Robust and Exploratory Data Analysis. NY: Wiley.

Office of Management and Budget (1987). Standard Industrial Classification Manual.

SAS Institute Inc.(1987), SAS/AF Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute.

SAS Institute Inc.(1990), Procedures Guide, Version 6, Third Edition. Cary, NC: SAS Institute.

SAS Institute Inc.(1994), SAS/Connect Software: Usage and Reference, Version 6, Second Edition. Cary, NC: SAS Institute.

SAS Institute Inc.(1993), SAS/Insight User's Guide, Version 6, Second Edition. Cary, NC: SAS Institute.

SAS Institute Inc.(1990), SAS/STAT User's Guide, Version 6 Fourth Edition. Cary, NC: SAS Institute.

Thompson, Katherine J. (1996) Ratio Edit Parameter Development Software. Bureau of the Census: Economic Statistical Methods and Programming Division. Unpublished internal documentation.

Thompson, Katherine J. and Sigman, Richard S. (1996), Evaluation of Statistical Methods for Developing Ratio Edit Module Parameters. Washington, DC: Bureau of the Census. Technical Report #ESMD-9601.

Thompson, Katherine J. and Sigman, Richard S. (1996), User Guide for Generalized SAS Ratio Edit Parameter Development Programs. Washington, DC: Bureau of the Census. Technical Report #ESMD-9602.

Questions or comments to:

Richard W. Graham
Department of Commerce
Bureau of the Census
Room 2517 Bldg. 3
Washington, DC 20233

Phone: (301) 457-2606

Fax: (301) 457-1777

e-mail: richard.w.graham@ccmail.census.gov