

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 5 of the provisional agenda

**DEVELOPMENT OF A "PLAIN VANILLA" SYSTEM
FOR EDITING ECONOMIC CENSUS DATA**

Submitted by the U.S. Bureau of the Census¹

¹ Prepared by Richard S. Sigman.

I. INTRODUCTION

1. The U. S. Census Bureau is re-engineering its data processing systems for economic surveys (and the Annual Survey of Manufactures (ASM)). This is being done by replacing much of the edit-and-imputation portions of different, census-specific systems with a general-purpose subsystem called "Plain Vanilla" (PV). Plain Vanilla is so named because it provides basic editing and imputation capabilities that a system designer can augment with survey specific code (i.e. toppings) to suit particular tastes. Section II describes our editing and imputation operations for economic censuses prior to re-engineering and Section III describes the re-engineering effort. Section IV describes the components of PV. The final section discusses some of our experiences in developing and implementing PV.

II. EDITING AND IMPUTATION OPERATIONS IN THE 1992 ECONOMIC CENSUSES AND 1995 ASM

2. The 1992 economic censuses were managed by three different Census Bureau divisions. Construction Division conducted the Census of Construction Industries; Industry Division conducted the Censuses of Manufactures and Mineral Industries; and Services Division conducted the Censuses of Retail Trade, Wholesale Trade, Finance and Real Estate, Services, and Utilities. As a result, the Census Bureau used three independent editing and imputation systems to conduct the 1992 economic censuses. The 1995 ASM used an additional editing and imputation system, which was related to the system used by the Census of Manufactures. Each of these four editing and imputation systems was very large, custom designed, and custom programmed with large quantities of custom code to edit each item. These four editing and imputation systems each performed the following functions:

- F1. Validated and supplied survey codes for industries, commodities, and geographic areas,
- F2. Imputed for nonresponse (This included mass imputation for "small" companies which were not required to file reports.),
- F3. Checked all data items and automatically produced "clean" records, and
- F4. Screened for potential reporting problems and identifying questionable records for analyst review.

3. Functions F2 and F3 were performed by the SPEER system and associated, census-specific satellite routines. SPEER (Structured Program for Economic Editing and Referrals) uses Fellegi and Holt (1976) methodology to perform ratio edits (Greenberg (1984); Greenberg and Surdi (1984); Greenberg, Draper, and Petkunas (1990)). The satellite routines, on the other hand, do not use Fellegi and Holt methodology and can be used to perform other types of edits, such as balancing (i.e. additivity) edits. At the time of the processing of the 1992 economic censuses (and the 1995 ASM), census-specific edits and imputations were coded directly into the system code of both SPEER and the satellite routines.

4. Table 1 summarizes the processing characteristics of the editing and imputation operations for the 1992 Economic Censuses and the 1995 ASM. The different Census Bureau divisions kept somewhat different statistics on how many records required corrections. In general, the statistics in Table 1 focus on analyst corrections rather than computer corrections.

For Services Division, statistics are available for corrections made through their correction system. These include analyst corrections but they also include some generalized adjustments done on large numbers of records. These general adjustments were monitored by the analysts but the actual corrections were made by machine. Industry Division had a two-step approach to editing and imputation. The first step identified write-in entries that had to be clerically coded and identified certain data deficiencies that were considered to be critical. In the second step, cases were referred for analyst review and correction. Construction Division also used a two-step approach of clerical screening followed by analyst review.

5. Table 2 categorizes the data items in the 1992 Economic Census by the type of edits the data items were subjected to. Of the 662 data items across the eight censuses, 515 of the items were subjected to balance edits. Table 2 also lists the number of "balancing complexes" present in the edits for each economic census. For simple-1d balancing, a balancing complex is a single equation, and for nested-1d and 2d balancing a balancing complex is a dependent set of equations.

III. RE-ENGINEERING THE EDITING AND IMPUTATION SYSTEM FOR ECONOMIC CENSUSES (AND ASM)

6. In 1994, managers of the Census Bureau's economic censuses became concerned about the large amounts of time and resources that were required to develop and maintain four separate editing and imputation systems. These four systems perform similar functions but for different trade areas. Consequently, management formed a working group to study the needs of the economic areas to see if common needs could be handled by a common system. The working group concluded that the following types of edits were common to all four existing systems:

- (a) testing of item relationships through ratio tests;
- (b) the testing and balancing of groups of items that were defined to be additive; and
- (c) validating codes assigned to industries, geographic locations, and kinds of business.

7. The group recommended that a general-purpose editing and imputation subsystem be developed. The group also recommended that edit designers reevaluate the need to include all the item-specific tests currently being performed. Management decided to form a dedicated team of technical specialists to develop the recommended subsystem. The team consisted of two subject-matter specialists, a mathematical statistician-programmer, two programming specialists, a full-time and a part-time mathematical statistician, and a team leader with subject-matter and edit design and processing backgrounds. In addition, the team had two

consultants with significant background in editing and imputation methodology.

Table 1. Processing Characteristics for Editing and Imputation in the 1992 Economic Censuses and 1995 ASM

Census Bureau organizational unit	Survey	Records processed [thousands]	Number of “analyst” corrections (see text) [thousands]	Number of analysts [FTEs]
Construction Division	1992 Census of Construction Industries	573	35 (step 1) 50 (step 2)	5
Industry Division	1992 Census of Mineral Industries	31	75 (step 1) 200 (step 2)	40
	1992 Census of Manufactures	382		
	1995 Annual Survey of Manufactures	57	6 (step 1) 40 (step 2)	20
Services Division	1992 Census of Utilities	244	approx. 200	55
	1992 Census of Wholesale Trade	495	approx. 400	
	1992 Census of Retail Trade	1,562	approx 1,100	
	1992 Census of Finance and Real Estate	586	approx 400	
	1992 Census of Services	2,034	approx 900	

8. The development team recently completed the development of the Plain Vanilla subsystem for editing and imputation. PV provides basic editing and imputation capabilities that subject-matter programmers can augment with custom code, as needed. For example, PV performs only deterministic imputations--the hot-deck imputations used to handle the Census of Construction’s unit nonresponse must be handled outside of PV. PV is not a stand-alone system but must be incorporated into survey-specific processing systems.

9. The PV subsystem consists of the following software described in more detail in Section IV:

- Modules for performing verification edits, ratio edits, and balance edits (written in FORTRAN so that they can be linked with FORTRAN and COBOL legacy systems);
- Program for generating implicit ratio-edits (also written in FORTRAN);
- Program for processing survey-specific script files, which specifies the edits and imputations to be performed (written in SAS²); and

² SAS

is a registered trademarks of SAS Institute Inc.

- Programs for determining ratio-edit parameters from historical data (also written in SAS).

Table 2. Item-Level Characteristics for Editing and Imputation in the 1992 Economic Censuses

Survey	Number of data items						Number of balancing complexes	
	Total	Subject to only ratio edits	Subject to only balance edits	Subject to both ratio & balance edits	Subject to neither ratio nor balance edits	Subject to SPEER ratio edits	Do not involve ratio edits	Do involve ratio edits
1992 Census of Construction Industries	63	19	17	26	1	10	0	3+
1992 Census of Mineral Industries	78	8	26	43	1	14	0	8+
1992 Census of Manufactures	74	7	27	39	1	14	0	7+
1992 Census of Utilities	93	4	66	4	19	5	11	3
1992 Census of Wholesale Trade	81	2	28	29	22	6	2	6
1992 Census of Retail Trade	59	1	27	13	18	7	2	8
1992 Census of Finance and Real Estate	56	1	34	3	18	4	4	3
1992 Census of Services	158	2	127	6	23	5	2	4
TOTAL	662	44	352	163	103	65	21	42+

9. These programmes are general purpose, in the sense that they do not have to be rewritten to be applied to different surveys. The PV module that processes ratio tests produces edited data that passes all ratio tests. Similarly, the PV module that processes balance tests produces data that passes all balance tests. PV is unable to determine imputations that simultaneously satisfy ratio tests and balance tests. For example, if one uses the ratio module and then the balance module, items involved in both type of edits may not satisfy all ratio edits at the conclusion of PV editing. In the 1992 Economic Censuses, however, of the 515 items involved in balance tests, only 163 of these items were also involved in ratio tests (see Table2). Most, but not all, of these 163 items were total items (as opposed to detail items) that were first ratio edited and then in subsequent balance editing they were held fixed. For these items, non-simultaneous balance editing following ratio editing would not result in some ratio edits not being satisfied. Draper and

Winkler (1997) report on their recent research into determining imputations that simultaneously satisfy both ratio and balance edits.

10. To incorporate PV into a survey-specific processing system, one performs the following steps:

- STEP 1: Prepare a survey-specific script file for the script processor program which creates survey-specific FORTRAN code.
- STEP 2: Compile the FORTRAN code for the PV edit modules and the FORTRAN code from STEP 1.
- STEP 3: Link the resulting object code into the survey-specific processing system.
- STEP 4: Create parameter files and generate implicit ratio-edits.

IV. COMPONENTS OF PV

1. Verification Module

11. The verification module verifies survey codes such as those for geographic area (states, counties, places, ZIP, and Metropolitan Statistical Areas) and industries (Standard Industrial Classification and North American Industrial Classification System). It does this by comparing provided codes to a master reference list. In some cases, the verification module returns another code associated with the matching code on the reference list.

2. Ratio Module

12. The ratio module performs ratio edits, followed by imputation of the minimum number of edit-failing items. The PV programmers used the system code described by Winkler and Draper (1996) and added to it code for 23 different imputation formulas. These formulas include regression models and functions of historic data, administrative data, or industry average ratios. These imputation formulas are general, instead of being census-specific code as they were for the 1992 Censuses. For the 1997 Censuses (and the 1996 ASM), subject-matter experts will use the PV script file to define census-specific edits and imputations.

3. Balancing Module

13. The balancing module consists of submodules for performing the following three types of balancing:

Simple-one dimensional (1d) balancing: $y = x_1 + x_2 + \dots + x_n$,

Nested-one dimensional (1d) balancing:

$$\begin{aligned} y_1 &= x_{11} + x_{12} + \dots + x_{1n(1)} \\ y_2 &= x_{21} + x_{22} + \dots + x_{2n(2)} \\ &\vdots \\ y_m &= x_{m1} + x_{m2} + \dots + x_{mn(m)} \\ z &= y_1 + y_2 + \dots + y_m \end{aligned}$$

Two-dimensional (2d) balancing:

X_{11}	X_{12}	...	X_{1n}	r_1
X_{21}	X_{22}	...	X_{2n}	r_2
...
X_{m1}	X_{m2}	...	X_{mn}	r_m
c_1	c_2	...	c_n	z

where r_i = sum of row i , c_j = sum of column j , and $z = \sum r_i = \sum c_j$ is fixed.

Each submodule first checks for additivity. If additivity is present or if all items are missing, no additional editing is performed.

14. The submodule for simple-1d edits can perform various adjustments commonly used by subject-matter experts, such as checking for rounding errors and raking totals to the details' reported relative proportions, to historic relative proportions, or to industry-average relative proportions. Other options include placing the residual into a Not-Specified-by-Kind (NSK) category or setting data to missing when weight adjustment is used to handle item nonresponse. In addition, we developed a default 1d-balancing procedure, called "trim and adjust", that adjusts to additivity based on specified item-quality weights and upper and lower bounds for each item.

15. For nested-1d balancing and 2d balancing, the balancing equations were represented as a network with arcs corresponding to items and residuals. This permitted us to decompose the general nested-1d balancing problem into 16 special cases and to solve the 2d balancing problem by using controlled rounding (Cox and Ernst, 1982) following iterative proportional fitting. Sigman and Wagner (1997) provides additional details about the PV balancing module.

4. Script files

16. The script file describes how PV processes a particular census or survey. In some of our testing of PV, members of our centralized survey-planning staff prepared script files, with minimal assistance required from PV programmers. In other testing, a PV programmer prepared a script file after meeting with subject-matter experts. In any case, the preparation of script files for PV has partially replaced the two-step procedure of a subject-matter expert first writing a specification and then a programmer writing or modifying a program based on that specification. We have found that the script file serves as an excellent communication device that quickly describes for analysts and managers how their data are being edited. Also, the ease of modifying the script has significantly speeded up our testing of alternative editing and imputation options.

5. Parameter files

17. There are two types of PV editing parameters: ratio-type parameters and imputation coefficients. Ratio-type parameters include ratio-test tolerances, industry-average ratios, and

the trim-and-adjust upper and lower bounds. Imputation coefficients are used in regression-model imputation formulas.

18. Software was developed that determines ratio-tst tolerances by applying exploratory-data-analysis (EDA) techniques to ratios of historical data. The software determines symmetrizing transformations using the procedure described by Hoaglin *et. al.*, (1983) and then calculates the tolerances from the following “resistant fences” formulas:

$$\text{lower tolerance} = q_{25} - k H, \quad \text{upper tolerance} = q_{75} + k H,$$

where q_{25} is the lower quartile, q_{75} is the upper quartile, H is the interquartile range, and k is a constant. Thompson and Sigman (1996a, 1996b) discuss choosing an appropriate value of k and also compare the resistant-fences approach to alternative approaches.

V. PRELIMINARY EXPERIENCES

19. The PV Development Team finished its work in early 1997, and the first operational use of PV occurred in the summer of 1997 for processing the 1996 ASM. This section describes some of our experiences in developing and implementing PV.

1. PV-script benefits

20. The script file describes how PV processes a particular census or survey. In some of our testing of PV, members of our centralized survey-planning staff prepared script files, with minimal assistance required from PV programmers. In othertesting, a PV programmer prepared a script file after meeting with subject-matter experts. In any case, the preparation of script files for PV has partially replaced the two-step procedure of a subject-matter expert first writing a specification and then a programmer writing or modifying a program based on that specification. It was found that the script file serves as an excellent communication device that quickly describes for analysts and managers how their data are being edited. Also, the ease of modifying the script has significantly speeded up our testing of alternative editing and imputation options.

2. Size of editing and imputation systems

21. PV has been successful in reducing the sizes of editing and imputation systems for the economic censuses and ASM. Table 3 compares the number of different modules in these systems before and after PV. Not only do the re-engineered systems have fewer modules, but the sizes of the modules are also smaller.

3. 1996 ASM

22. The first operational use of PV was the editing and imputation of the 1996 ASM. Prior to this, PV was tested with 1995 ASM data. This testing revealed that some of the very complex procedures for ASM could not be implemented in PV. Some of these procedures were implemented in custom code, but others were re-evaluated and eliminated. The 1996 ASM used the PV ratio module to edit key items, followed by the use of the PV balancing module. The key items included a number of items that were also totals of simple-1d balance complexes. Following ratio editing, these items were held fixed in the subsequent balance editing. The use of the PV balancing module eliminated much of the custom code in the ASM satellite edits.

23. At the time of writing this paper, production running of the 1996 ASM was only about half completed. Based on a subjective evaluation, however, subject-matter analysts are pleased with the edit-and-imputation results.

Table 3. Number of modules in editing-and imputation systems before and after PV

Census Bureau organizational unit	Survey	Number of editing-and-imputation modules		Programming language
		Before PV	After PV (includes PV modules)	
Industry Division	1992 Census of Mineral Industries	9	4	FORTRAN
	1992 Census of Manufactures			
	1995 Annual Survey of Manufactures	8	4	
Services Division	1992 Census of Utilities	80	35	COBOL (3 PV modules in Fortran)
	1992 Census of Wholesale Trade			
	1992 Census of Retail Trade			
	1992 Census of Finance and Real Estate			
	1992 Census of Services			

REFERENCES

- Cox, L. and Ernst, L. (1982). "Controlled Rounding," *Infor*, 20, pp. 423-432.
- Draper, L. and Winkler, W. (1997). "Balancing and Ratio Editing with the New SPEER System," to appear in *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Fellegi, I.P. and Holt, D. (1976), " Systematic Approach to Automatic Editing and Imputation," *Journal of the American Statistical Association*, 71, pp. 17-35.
- Greenberg, B. (1984). "The use of Implied Edits and Set Covering Approach in Automated Data Editing," Technical Report Census/SRD/RR-86/02, Washington, DC: U.S Bureau of the Census.
- Greenberg, B. and Surdi, R (1984). " Flexible and Interactive Edit and Imputation Stem for Ratio Edits," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 421-426.
- Greenberg, B.; Draper, L.; and Petkunas, T (1990). "On-Line Capabilities of SPEER," *Proceedings of the Statistics Canada Symposium*, Statistics Canada, pp. 235-243.
- Hoaglin, D; Mosteller, F; and Tukey, J (Eds) (1983). *Understanding Robust and Exploratory Data Analysis*. NY:Wiley.
- Sigman, R and Wagner, D (1987). "Algorithms for Adjusting Survey Data that Fail Balance Edits," to appear in *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Thompson, K. and Sigman, R (1996a). "Statistical Methods for Developing Ratio Edit Tolerances for Economic Censuses," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- _____ (1996b) *Evaluation of Statistical Methods for Developing Ratio Edit Parameters*. Technical Report #ESM-9601, Washington, DC: Bureau of the Census.
- Winkler, W. and Draper, L. (1996). "The New SPEER Edit System," in *Statistical Policy Working Paper 25: Data Editing Workshop and Exposition*, Washington, DC: Office of Management and Budget, pp. 50-58.