

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 11
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

Item 3 of the provisional agenda

EDIT/IMPUTATION SYSTEM FOR THE U.S. DECENNIAL CENSUS

Submitted by the U.S. Bureau of the Census¹

¹ Prepared by William E. Winkler.

Abstract

This document describes a Fellegi-Holt edit/imputation system for the U.S. Census. The edit subportion of the system uses a version of the DISCRETE edit system. The imputation subportion of the system uses Gibbs sampling to impute missing household characteristics and then uses a combination of logistic and ordinary regression to impute person characteristics.

Keywords: Fellegi-Holt, error localization, optimization, imputation.

I. INTRODUCTION

1. In the paper (Fellegi and Holt 1976), the authors introduced ideas to provide a theoretical solution to the problem of statistical data editing. Their methods had the virtues that, in one pass through the data, an edit-failing record could be assured to satisfy all edits and that the logical consistency of the entire set of edits could be checked prior to the receipt of data. The implementations of the system have had additional advantages over traditional if-then-else rule edit systems because edits reside in easily modified tables and the main mathematical routines do not need to be modified.

2. Moving Fellegi-Holt principles into survey practice has been slow because of the need to develop sophisticated software algorithms for integer programming and set covering. Early implementations have shown much promise and flexibility. Very large applications have been severely hampered by the need for faster hardware and for the development of possibly faster algorithms in operations research.

3. This paper describes an application of the Fellegi-Holt principles to the U.S. Decennial Census. To improve editing, the decennial-edit-development group introduced two new ideas to editing and developed two sets of nontrivial statistical imputation methods. Robert Hemmig created an efficient structural form for household-person edits that drastically improves the ability to write the computer-code and to adapt to different situations. A new set-covering theory and computational algorithms (Winkler 1997) has been produced for DISCRETE that increased the speed of the implicit-edit generation algorithms by a factor of 100 over the previous version (Winkler 1995). Yves Thibaudeau developed a sophisticated way of using the EM algorithm and Gibbs sampling to impute household characteristics. Todd Williams developed models and software for a combination of logistic and ordinary regression for person characteristics.

4. The outline of the paper is as follows. In the second section, more background about an earlier application of Fellegi-Holt systems to discrete data and about our specific development for the Decennial Census is given. The second section gives more details about the edit subportion of the system and the third section gives more details about the imputation subportion of the system. The fourth section is discussion and the final section consists of a summary.

II. BACKGROUND

5. This section describes previous work in applying Fellegi-Holt methods to the largest applications such as a household census of persons in Canada and background on the U.S. decennial application is provided.

6. The two major components of a Fellegi-Holt system are a programme to generate the complete set of implicit edits from a set of explicit edits and a program to edit/ impute the survey data. Because the implicit-edit generation program can be run weeks or months prior to production, speed is not a crucial issue in moderate-size applications. While the edit subportion of the edit/imputation system can use integer-programming algorithms which, in some applications, are slow, both the edit applications in DAISY (Barcaroli and Venturi 1997) and DISCRETE (Winkler 1995) are quite fast. The crucial issue in the largest applications is whether implicit edits can be generated in a reasonable amount of time.

7. In the largest applications, the underlying data structure often gets too large for the current hardware and computational algorithms. With a household census, the inherent data structure is very large because there must be fixed locations for answers for each possible person that might answer the question. For the householder, possible spouse, possible first child, possible second child, possible first parent, possible second parent, possible roommate and so on, there must be a data structure that holds the reported information.

8. In 1991, the Fellegi-Holt system CANEDIT was used to edit the Canadian census. In the document describing their newly adopted NIM methodology, Bankier et al (1995) provide a summary of how data-structure-size difficulties can make implementing Fellegi-Holt systems a challenge. In the following, the 1991 application of CANEDIT is described. NIM which is not intended as a general Fellegi-Holt system is not described. To implement CANEDIT, the Canadians partitioned the set of households into 1-person households, 2-person households, 3-person households and so on. With 3-person households, the number of edits was more than 200 and with 4-person households, the number was greater than 300. CANEDIT was able to deal with households of up to 12 persons. For some of the larger households, however, the edit rules were simplified. Also Decade rather than Age was used in the edits. These two compromises were made to reduce the amount of computation which would have been too large otherwise.

9. The U.S. Decennial Census is a household survey of persons which collects age, race, sex, Hispanic-origin, and relationship to head of household. A householder must always be present and the survey also collects tenure (whether the householder owns or pays rent). Most edits involve ages and how they relate to the age of the householder or the age of the spouse of the householder. As such, most of the edits are dependent on householder information. For instance, a parent must be older than the householder and a child must be younger than the householder. Another edit is that the householder and the spouse must be of the opposite sex. Most edits are quite simple. The main set of edits consists of 333 explicit edits for 33 variables that assume 96 value-states.

III. EDIT

10. This section describes the edit subportion of a production system for the U.S. Decennial Census. There are two key features. In the first, we take the original set of input variables and produce a large number of new variables representing age relationships between persons in the household. For instance, if we have an edit that specifies a parent must be 12 years older than the householder, we create the edit $E = \{\text{person1} < \text{person2} + 12, \text{person2_relat} = \text{parent}\}$. The variable person1 always refers to householder because a householder must always be present. These variables V_E associated with the edits of form E take 2 values: 1 if the condition within the brackets holds and 2 if the condition does not. The edit E (referred to as derived or conditional) replaces an edit of the form that explicitly enumerates all the different age combinations for which person1 (householder) is 12 or less years than person2 when person2 has relationship = parent. The set of explicitly defined edits consists of all the original edits that did not involve age relationships among two persons and the set of derived edits. In the first tests of the system, edits for only 3-person households were developed. One set of callable routines uses the original variables and the age-edits for pairs specified in form E to output the new set of variables for each incoming record. Explicit edits are specified in terms of the original variables and the derived variables. Both the implicit-edit-generation programme and the main edit programme use all variables. Because error-localization will specify derived variables in the solution to the minimum number of fields to impute, another program converts the error-localization solution that includes both the original variables and derived variables to one that includes only the original variables.

11. The second part of the new application is a way of reducing the set of edits on households of arbitrary size into a set of 3-person household edits, editing, and then recombining the results. This was possible because all the edits of a person other than the householder are conditional on information in the householder subportion of the record.

12. Another feature is a much faster set-covering algorithm for generating implicit edits (Winkler 1997). The time for generation of the complete set of 1560 edits from 252 explicit edits having 96 value-states is reduced from 6 hours to two minutes. The increased speed is useful for investigating different sets of edits and may be useful in very large applications where the set of implicit edits could not be generated even in one or two hundred hours. In practice, the set of implicit edits can be generated weeks or months prior to production. After a set of records passes through the edit modules, it goes to the imputation module. Each field that must be imputed has a flag set.

IV. IMPUTATION

13. Unlike some edit/imputation systems, editing is separated from imputation to make use of certain aggregate properties in each region. For Census processing, the U.S. is divided into approximately 550 District Offices (DOs) representing contiguous geographic regions. Each DO represents between 200,000 and 700,000 individuals. Each DO is subdivided into tracts which each represent approximately 2000 individuals.

14. Imputation proceeds in two stages. In the first stage, missing housing

characteristics are filled in in an entire file representing a DO. For each housing unit, missing values are filled in for the following indicator variables: rent/own, householder_male/female, householder_nonblack/black, householder_35+/not, unit in mail response universe or not, and unit has one occupant or more than one. This yields a contingency table of 64 cells for which a loglinear model of all 3-way interactions is fitted. A straightforward EM algorithm is used for the modeling. Once a model is developed, a modified version of Gibbs sampling is used to produce imputations from the posterior distribution. The advantage of Gibbs sampling is that it yields integer values when we fill in imputed values. The EM — while much faster than Gibbs sampling — yields fractional values which must be rounded in a controlled manner so that we can fill in the set of characteristics over the entire DO.

15. In the second stage, missing person characteristics are filled in for each household based on models that make use of the housing characteristics. For instance, imputed values are quite dependent on whether tenure variable is owner or renter, whether the householder is male or female, whether the race is nonblack or black, and whether the householder is 35+ or not. We begin by filling in each person's relationship to the householder when the relationship is missing or marked for imputation. For this, two different logistic regression models are needed where the dependent variable has ten different levels of response of the relationship variable. The most significant predictor variable for a person's relationship is the difference in age between the person whose relationship is missing and the householder.

16. The first model is used when both ages are present. The predictor variables are the difference in ages and the mean number of householder children within in a housing unit within a tract. The second model is used when either of the two ages is missing and the predictor variables include 1) if there is a spouse of the householder, 2) the number of persons in the housing unit, 3) the mean number of children within a housing unit for the tract, 4) the mean householder age for the tract, 5) the sex of the householder, 6) tenure, 7) if an enumerator visited the housing unit, and 8) race (nonblack or black). Two of the predictor variables are used because they significantly increase the fit of the models. They also allow the variations in relationship and age that may occur between different tracts.

17. The key feature of the edit/system is the imputation methodology that models and preserves distributional characteristics at the tract level. The methods for creating derived edits greatly facilitate adapting the Fellegi-Holt methods and should find applications elsewhere. Speed in processing a DO is not an issue. On a DEC Alpha, the edit modules require less than 1 hour of CPU time and the imputation modules require less than 2 hours.

*This paper represents views of the author and not those of the U.S. Bureau of the Census. The author thanks Michael Bankier of Statistics Canada for correcting erroneous statements about CANEDIT in an early draft.

REFERENCES

Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1995), "Imputing Numeric and Qualitative Variables Simultaneously," Social Survey Methods Division, Technical Report.

Barcaroli, G. and Venturi, M. (1997), "DAISY (Design, Analysis and Imputation System): Structure, Methodology, and First Applications," in *Statistical Data Editing, Volume II*, United Nations Statistical Commission and Economic Commission for Europe.

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

Thibaudeau, Y., Williams, T. and Krenzke, T., (1997), "A Model Based Approach for Imputing Short Form Items in 2000," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.

Winkler, W. E. (1995), "Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 108-113.

Winkler, W. E. (1997), "Set Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.