BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/18

A FLEXIBLE AND INTERACTIVE EDIT AND IMPUTATION SYSTEM FOR RATIO EDITS

by

Brian Greenberg and Rita Surdi
Statistical Research Division
U.S. Bureau of the Census

Room 3587, F.O.B. #3
Washington, D.C. 20233

(301)763-7530

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, DC 20233.

Recommended by: Paul Biemer

Report completed: August 6, 1984

Report issued: August 6, 1984

A FLEXIBLE AND INTERACTIVE EDIT AND IMPUTATION SYSTEM FOR RATIO EDITS

Brian Greenberg and Rita Surdi

All survey and census programs are subject to nonresponse and erroneous reporting, whereas data users demand complete and accurate data to be used for a varietv of statistical purposes. Although the implementation of an edit and imputation svstem is highlv survey specific, coherent methodologies can be developed that integrate diverse features and needs into a structured framework. Various imputation strategies, subject-matter expertise, and auxiliary information can be incorporated within such a framework.

A widely used criterion for economic data requires that the ratio of two responses lie between prescribed bounds. The upper and lower bounds are determined by historical information, subject-matter expertise, and when feasible, bv a sample of responses. In addition to comparing two fields on the report form, ratio edits can incorporate data from an earlier time frame as well as information from an external data file. A system to edit data under ratio edits has been developed at the Bureau of the Census and a prototype model has been developed for the Annual Survev of Manufactures. A modification of this prototype system was designed and used to process two segments of the 1982 Economic Census. An interactive version of this system has been developed for use by subject-matter analysts for on-line processing of referral cases.

I. INTRODUCTION

All survey and census programs are subject to nonresponse and erroneous reporting, whereas data users demand complete and accurate data to be used for a variety of statistical purposes. It is well-recognized that the data collection agency has the optimal vantage point and attendant obligation to provide valid allocations for missing values and to adjust spurious responses. The development of statistically precise and mathematically rigorous edit and imputation systems is essential in meeting this objective and is vital in providing users with high quality data products.

Although the implementation of an edit and imputation system is highly survey-specific, coherent methodologies can be developed that integrate diverse features and needs into a structured framework. Within such a framework, various imputation strategies, subject-matter expertise, and auxiliary information can be incorporated. State-of-the-art edit systems draw upon operations research optimization techniques, mathematics, and statistical analysis to incorporate prior knowledge and concurrent information. Development and implementation of such systems require that mathematical and statistical investigators work jointly with subject-matter specialists familiar with the survey environment.

The role of the edit process is to alter erroneous responses and not to alter valid ones. In most discussions of editing the focus is usually on altering erroneous fields; however, we should beware of overzealousness and take precautions against changing correctly reported values. One should endeavor to assert that a record is acceptable, even in the face of several failed statistical edits if information can be garnered from ancillary sources or from the record itself to support its validity.

One imputes because of item nonresponse and because fields have been targeted for change based on patterns of edit failure. The role of the imputation process is not simply to create a consistent record nor to allocate values based on a random generation from a presumed underlying distribution. The ideal goal (though generally not practicable) is to create a revised record close to what a respondent would have reported were there no errors. In particular, when one imputes in a field deleted due to edit failures the imputation strategy should take into account the reported value (albeit incorrect) whenever possible, and the imputation for edit failures might be different from that for item nonresponse. For example, in some surveys, a frequent reporting (or keying) problem is that a field is in error by a multiple of one thousand. For the fields susceptible to this sort of error, one should attempt to detect it and divide the recorded response by one thousand.

The relation between editing and imputation is fundamental, and it is crucial to integrate these two features when designing an error correction system. One aspect of the relation is technical: imputed values should not fail edits except in prespecified special cases. Accordingly, an important aspect of the imputation process is the editing of imputed values—assuming that non-imputed variables all pass edit checks. An imputation procedure based on an estimation process, especially one involving a stochastic component, can yield specious imputations. For example, due to the contribution of a residual, an estimate of a missing value may be negative—usually proscribed. But more generally, interrelated data items often must conform to edit constraints, and to ensure that one does not impute a value that would be rejected if it were reported, the candidates for imputation have to be checked for feasibility. Those not feasible have to be either reimputed or adjusted. If a non-feasible or suspicious imputation occurs in a set of fields that were targeted for change due to edit failures, an alternate set of fields to adjust may be indicated. Of course, if the imputation strategy can ensure feasibility, so much the better.

Another aspect of the relation between editing and imputation is far more intimate and must run throughout a coherent system. Simply stated, the variables and criteria that contribute to the editing of reported data and are embedded in the edit constraints should play a role in determining a valid and meaningful imputation. For example, if the imputation is to be based on matching to records from other respondents (e.g., hot deck, statistical matching) the connection between the edit step and the imputation is that the matching be based on variables that enter edits for missing fields. If the imputation is based on other reported values on the same record (as in a regression procedure), once again, the variables most prominently contributing to the impute should be those in edits for that field. By utilizing variables most closely related to the field to be corrected in *both* editing and imputation, one endeavors to guarantee that imputed values pass all edits.

The seminal paper relating editing and imputation is by Fellegi and Holt, [2]. In that paper, the primary focus is on categorical data, and the imputation strategy most discussed is matching records to other respondents. For fields to be imputed, the variables driving the match are the same ones used to edit those fields. Important work utilizing this connection between editing and imputation for continuous (economic) data and linear edits has been done by Gordon Sande [5]. In this work, mathematical programming was employed to determined a feasible region, and an acceptable record was one falling into the feasible region. After the fields to delete were identified, matching was used to obtain a feasible impute. Once again, the fields involved in edits of a given variable (or set of variables) to be imputed were used for the match. Further discussion of the relation between editing and imputation is contained in [6].

There are basically three types of edits: structural, statistical, and subject-based. Structural edits are based on a logical relation between two or more fields, for example, a total must equal the sum of its parts, or, because of a skip pattern inherent in a questionnaire, two variables lying on disjoint paths cannot both be non-zero. Statistical edits are constraints based on a statistical analysis of respondent data, for example, the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters. Subject-based edits incorporate "real-world" strictures which are neither statistical nor structural, for example, the ratio of wages paid to hours worked (i.e., hourly wage) must exceed the minimal hourly wage. Of course, some edits are hybrids.

In determining the validity of a respondent record, structural edits *must* pass while statistical or subject-based edits *should* pass unless there is cogent countervailing evidence. A record revised by an automated system should pass all structural edits and imputed values should pass virtually *all* edits. That is, we may accept some respondent records even if selected statistical or subject-based edits fail because there may be countervailing information, but we should be unwilling to allow an automated system to impute an edit failing value except under very controlled circumstances.

As with edits, we can classify imputation rules into three basic types: structural, statistical, and subject-based, each based on the same principles as the corresponding edits. One employs a structural imputation when a structural relationship holds between several variables (e.g., a total must equal the sum of its parts), so that if one of these constituent variables is missing, an appropriate imputation may be inferred from the remaining. An example of a statistical imputation is the use of a regression model where the dependent variable is to be imputed, and the coefficients of the independent variables are derived from presumed valid responses. The more sophisticated E-M algorithm will also fit in this category. Subject-based imputations are contributed by subject-matter experts who are knowledgeable about the respondent population, subject-matter of the survey, and recurring sources of respondent (or keying) error. For example, subject-matter specialists may be aware that some respondents report a variable in pounds rather than tons as per instructions, and when this detected on a record an effective correction would be to divide the response by 2000.

Broadly viewing an edit and imputation system as a model to correct for misreporting and to allocate for non-response, it will have to incorporate each of the three types of edit and imputation procedures discussed above. The statistical modeling techniques for treating non-response that are currently making their way into journals are sophisticated and potentially powerful. However, from the point of view of implementation, they must be embedded in a comprehensive system for survey editing and imputation. A facile application of some statistical strategy (especially one which ignors edit constraints) will not suffice for a sensitive and meaningful broad-based system. For any survey, subject-matter specialists must be part of a team designing an edit and imputation system. A flexible and structured methodology can provide a framework for subject-matter expertise and statistical techniques and integrate them to model edit and imputation requirements.

II. AUTOMATED VERSION OF CORE EDIT SYSTEM

A. Overview

The system described in this paper, referred to as the core edit, endeavors to adhere to the strictures of an edit and imputation system as outlined above. We regard the advances made by Fellegi, Holt, and Sande as methodological progenitors, and we freely borrow ideas and constructs from each. The notions of implied edits, their generation, and their use are discussed in [2] and the principle of a feasible region for continuous data under linear edits is discussed in [4].

For the system to be described, we begin with a family of explicit ratio edits, generate the implied edits, and use all the edits to determine fields to delete for edit-failing records. After the designated fields are deleted, we have a record with some missing values and remaining fields consistent, and we use all edits (including implied) and the remaining (presumed valid) field values to obtain a feasible region for each missing field. By imputing a value that lies in the acceptance region for each missing field, we ensure that no edit failures will be introduced by the imputation process. But equally important, the feasible region, by providing a range of acceptable values, aids in the selection of a suitable imputation from a range of options.

For each field on the record, we create a brief subroutine, called an imputation module, consisting of a sequence of imputation rules. To impute for a missing field, the value generated by the first applicable rule is tested for feasibility (that is, consistency with all other fields on the record). If that value is feasible, it is accepted as the imputation, and the system proceeds to the next missing field. If not, we generate a value based on the next applicable rule, determine if it is feasible, and proceed down the rules as necessary. Should all rules generate non-feasible values, we make no imputation and the record is flagged for review.

Each imputation module is created using information furnished by subject-matter specialists who are familiar with the survey questionnaire, the target population, sources of non-random error, and the availability of auxiliary information. As noted above, some imputation rules are structural, some subject-based and others statistical. Imputation modules are easy to create and they can be easily revised to accommodate new understandings about the data being edited.

The core edit was originally designed for use on the Annual Survey of Manufactures (ASM). In developing this system, all survey specific procedures were isolated in well-defined segments so that changing only these modules would make the system usable for other surveys and censuses. This system was successfully used to process two segments of the 1982 Economic Censuses: The Enterprise Summary Report and the Auxiliary Establishment Report. ASM-specific modules were removed from the system, and subject-matter specialists in the Economic Surveys Division at the Census Bureau created imputation modules for these two surveys. As part of an edit and imputation evaluation project, imputation routines used by Business Division for editing basic data items for selected retail, wholesale and service establishments on the 1982 Economic Censuses have been incorporated into this system. Industry Division will soon conduct large-scale testing of this system on the Annual Survey of Manufactures.

B. The Edits and Feasible Region

In an earlier paper, [3], the first author discusses the nature of ratio edits, the procedure for generating implied edits, and the techniques for locating fields to delete for edit-failing records. We refer the reader to that paper for a detailed discussion of these topics. After setting the stage and introducing necessary definitions and notation, we proceed directly to a discussion of imputation strategy.

We assume that our data are continuous and non-negative, for each record there are N fields, $F_1,\ldots,F_N$, and we denote by A_i the value of field F_i. A ratio edit between field F_i and field F_h is the requirement that

$$L_{ih} \leq A_i/A_h \leq U_{ih}$$

where L_{ih} and U_{ih} are non-negative, extended real numbers (i.e., U_{ih} can be infinite), which are specified in advance. Given two ratio edits

$$L_{ih} \leq A_i/A_h \leq U_{ih}$$

$$L_{hj} \leq A_h/A_j \leq U_{hj}$$

the implied ratio edit is

$$L_{ih}L_{hj} \leq A_i/A_j \leq U_{ih}U_{hj} .$$

After all implied edits are generated and suitable reductions are made, for each pair $(i,j) \varepsilon NxN$, there is an edit

$$L_{ij} \le A_i/A_j \le U_{ij}.$$

Prior to processing data, implied edits are generated, the system detects inconsistencies in the edit set (i.e., a lower bound for some ratio will exceed its upper bound), and the implied edits are reviewed and changed if necessary by subject-matter specialists.

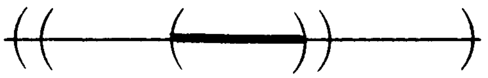
In the editing of an individual record, after erroneous fields are identified and deleted and the remaining fields on a record are verified as consistent, it is necessary to impute for missing values. Suppose K fields on a given record are to be imputed ($K< N$). By reordering, we can assume the missing fields are $F_{N-K+1},\ldots,F_N$ and the fields $F_1,\ldots,F_{N-K}$ all have valid values. Imputations will be made sequentially beginning with field F_{N-K+1} in the following manner. Consider all edits involving field F_{N-K+1} and those fields considered reliable, namely $F_1,\ldots,F_{N-K}$, to obtain an interval in which A_{N-K+1} must lie. That is, we have edits:

$$L_{N-K+1,\,j} \le \frac{A_{N-K+1}}{A_j} \le U_{N-K+1,\,j}$$

for all $j=1,\ldots,N-K$. Since the L's and U's are known real numbers, and A_j for $j=1,\ldots,N-K$ are known, we have a set of N-K overlapping closed intervals:

$$L_{N-K+1,\,j}\ A_j \le A_{N-K+1} \le U_{N-K+1,\,j}\ A_j\,.$$

The intersection of this region is represented by the shaded area below

,

and this is the interval in which A_{N-K+1} must lie to be consistent with all other fields. Denoting this interval, called the feasible region, by I_{N-K+1}, we note that I_{N-K+1} is not empty whenever the edit set is consistent and the non-blank fields conform to the appropriate edits. After selecting an imputation for field F_{N-K+1}, we proceed to derive the feasible region for F_{N-K+2} (i.e., I_{N-K+2}) using all appropriate edits and the field values A_j for $j=1,\ldots,N-K+1$.

C. An Example Based on the 1982 Economic Censuses.

The imputation rules currently used in Business Division for retail, wholesale, and service establishment respondents to the 1982 Economic Censuses are defined by a series of decision logic tables. As part of an edit and imputation evaluation project, for selected Standard Industrial Classification (SIC) groupings, these rules are incorporated into the core edit and data from establishments in these SIC's were edited using this system. For a typical establishment, there are four data records: (1) the response data, (2) 1982 Administrative Data, (3) 1981 Administrative Data, and (4) 1977 Economic Census data, although for some establishments one or more of these data records may be missing.

To impute for a missing field, for example Annual Payroll (APR), the edit system first determines the feasible region for this field as described in Section B. It then tests candidate values for feasibility in a specified sequence. In this example, the first candidate value would be the 1982 Administrative Data value for Annual Payroll. If that value lies in the feasible region for APR the system makes a direct substitution and imputes for APR the corresponding 1982 Administrative Data value. If the 1982 Administrative Data value for Annual Payroll does not yield a suitable impute, the system next derives an imputation candidate based on the 1981 Administrative Data value for Annual Payroll. If that value is in the feasible region the system accepts it, otherwise the system derives a potential imputation based on the 1977 Economic Census value for APR. If that value is not acceptable, a value is derived from other response variables on the report form, in this case, Quarterly Payroll or Number of Employees.

If the reported value of APR is very large, far exceeding any reasonable value (as detected by some edit), an imputation candidate is generated by dividing the reported value by 1000, sometimes called rounding. If this rounded value lies in the feasible region for APR it is accepted as the impute. Since respondents sometimes report in dollars rather than in 1000's of dollars as instructed, when a rounded value is feasible this adjustment to the reported value is very reasonable. The rounding option is not included in the imputation module for Number of Employees because the corresponding reporting error does not occur in that field.

The point of this example is to give the flavor of what an actual imputation module might look like. Special situations, such as part-year employers, were not discussed; however, they were incorporated into the system with ease. This example does illustrate how subject-matter expertise and auxiliary data can be incorporated into an imputation module.

D. Example using the Annual Survey of Manufactures

In creating the imputation modules for a prototype edit and imputation system for the ASM, we worked closely with subject-matter experts to develop imputation routines for each variable being treated. For most data records, the prior year report from the same respondent (establishment) was available. Thus, in addition to the field-to-field edits discussed earlier, we also had year-to-year edits to work with. These edits are of the form

$$B_i/B_j \; L'_{ij} \leq A_i/A_j \leq B_i/B_j \; U'_{ij}$$

where B_i is the prior year value in field i, i=1,...,N. (That is, the accepted prior year value of the ratio of field i to field j is modified by limit multipliers to determine an acceptable range for the current year ratio.) These edits, prior year values, and the implied edits all contributed in determining fields to delete for edit-failing records and in determining the feasible region for each field (see [3] for details).

The imputation modules incorporate a large amount of survey-specific information supplied by subject-matter specialists. For example, certain fields were the sum of other fields, for selected fields (although not others) a blank was usually an indication that the response should likely be zero, rounding was used on selected fields, and for some fields an accepted prior year value of zero was a strong indication that zero would be appropriate again. After these subject-based or structural imputation options were incorporated into the system, those of us working on the ASM system developed a sequence of regression models. Each field to be imputed became the dependent variable, and related fields became independent variables. In most cases, for each dependent variable, the independent variables consisted of fields involved in explicit edits furnished by subject-matter specialists as discussed in the introduction (see [1], for more details.)

For each of the ten selected variables whose missing value was to be imputed using a statistical regression, two variables were chosen as independent variables for a family of regression models. Given a triple of dependent variable and two associated independent variables, six models were obtained for each field to be imputed (see below). Three of the models use prior year data and three use only current year data. After all six models for a dependent variable were derived, they were ranked according to their ability to predict that variable. We will describe the criterion by which the models were ranked and the manner in which they are employed within a coherent strategy. A(X) will denote the current year value for variable X, and B(X) will denote the prior year value. In this study we used data collected on the 1981 ASM in which responses which failed edits were deleted. In discussing the models, "DEP" will denote the dependent variable, "IND1" and "IND2" will denote the corresponding independent variables, and β_k will denote estimates of regression coefficients, where k = 1,..., 8. Estimates of these coefficients were obtained for each of 27 industry groupings based on 4-digit SIC codes.

For each triple of dependent and two independent variables, we considered the following six models:

Model 1: $A(DEP) = \beta_1 * A(IND1)$

Model 2: $A(DEP) = \beta_2 * \frac{B(DEP)}{B(IND1)} * A(IND1)$

Model 3: $A(DEP) = \beta_3 * A(IND2)$

Model 4: $A(DEP) = \beta_4 * \frac{B(DEP)}{B(IND2)} * A(IND2)$

Model 5: $A(DEP) = \beta_5 * A(IND1) + \beta_6 * A(IND2)$

Model 6: $A(DEP) = \beta_7 * \frac{B(DEP)}{B(IND1)} * A(IND1) + \beta_8 * \frac{B(DEP)}{B(IND2)} * A(IND2)$.

We derived estimates of the regression coefficients for each triple listed below:

DEP	IND1	IND2
WW	SW	PW
PW	WW	MH
OW	SW	OE
OE	TE	OW
MH	WW	PW
SW	VS	TE
VS	SW	CM
SLC	SW	LE
TE	VS	SW
CM	VS	SW,

where

WW =	wages for production workers
PW =	number of production workers
OW =	number of non-production workers
OE =	wages for non-production workers
MH =	hours worked for production workers
SW =	total salary and wages
VS =	value of shipments
SLC =	supplemental labor costs
TE =	total number of employees
CM =	cost of materials
LE =	legally required supplemental labor costs.

Given a dependent variable, DEP, and an SIC, regression coefficient estimates were obtained, that is, β_k, k = 1,...., 8. The six models above were ranked using the statistic

$$D_j^2 = \sum_{i=1}^{N} (A_i(DEP) - A_{ij}(DEP))^2 / _N \qquad j=1,...6.$$

Note that A_i(DEP) is the observed value of DEP for the i^{th} case, A_{ij}(DEP) is the predicted value of the i^{th} case of variable DEP using Model j, and N is the number of in-scope records. That is, D_j^2 is a measure of cumulative difference between the observed values of DEP and the predicted values of DEP using Model j, for j=1,...,6. The models were ranked by ascending value of D_j^2, with minimum D_j^2 preferred. (Note that Model 5 will always be ranked before Models 1 and 3, and Model 6 before 2 and 4. Of course, the more familiar statistic $E_j^2 = (N/N-r_j)\ D_j^2$, where r_j is the number of independent variables in Model j, or other measures of difference between observed and predicted values, could be employed for rankings.)

The models developed for each dependent variable were incorporated into the imputation scheme for that variable with each model providing an option for imputation. To impute for a missing field, the model ranked first is tested to see if the value it predicts furnishes a valid imputation (i.e., falls in the feasible region). If it does, that value is substituted for the missing field. If the value based on the first ranked model is not suitable, we test the value based on the second ranked model and so on, testing each candidate until a feasible imputation is found. If any of the information required for a model is missing, we move down to the next ranked model. If none of these models provides a suitable imputation, alternate procedures are called upon. A necessary condition for a suitable imputation is that the candidate value lie in the feasible region, thus, by use of this strategy, we are able to guarantee that imputed values pass all relevant edits.

Note that in our regression models, we did not add a residual error term; but of course, we certainly could have done so. In some regression-type imputation procedures, the candidate for an imputation value can be less than zero because of the addition of a residual. That is, the impute would fail the non-negativity constraint, and when this occurs, that value is rejected as an acceptable impute. However, these systems rarely check as to whether an impute containing a residual conforms to other edits. The core edit system is well suited to the incorporation of a residual term since each candidate imputation is checked for feasibility. The objective of this section is not to advocate any one imputation scheme, but rather to impart a flavor as to how a statistical model can be incorporated into this system.

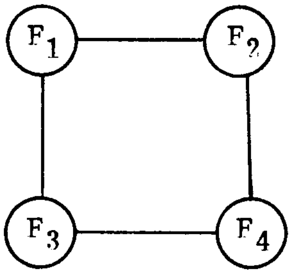
III. INTERACTIVE VERSION OF CORE EDIT

All large scale automated edit and imputation systems run data records in batch mode, and based on the actions taken by the automated system, records are selected for analyst review. The analyst then examines the overall performance of the automated system and further adjusts individual records as needed. Typical causes for analyst review are large edit changes or changes on records for large establishments. We have developed an interactive version of the core edit system for use by analysts during the review process. The interactive system allows an analyst to target one or more fields for revision, observe the feasible region, select amongst the system generated imputation options, delete alternative fields, and observe (while on-line) the impact of any changes. If a field was deleted because of edit failures in a record, this interactive version of the system can be used to generate alternative sets of fields to delete.

When using the interactive, on-line version of the core edit to review referral cases, both the original and revised versions of a record to be reviewed are displayed, and the following message is printed "Is this record acceptable?" If so, the system proceeds to the next record for review. If not, the system ask which fields the analyst wants to examine further.

For concreteness, suppose we are working with the edit for retail, wholesale and service establishments and the analyst wants to examine Annual Payroll (APR) and Sales (SLS). The user indicates these fields and processing begins with APR (conforming to the order in which fields are to be imputed). The system next displays the range of the feasible region for APR, the current value, and the values generated by each impute option embedded in the imputation module for APR. That is, it will print out the 1982 Administrative Data value, the value based on the 1981 Administrative Data, the value based on the 1977 Census data, etc. The user can then choose from these values for an alternative impute, or enter any other value for that field. For example, if the analyst detects a keying error on APR, he/she can enter the correct value from the respondent form. After completing APR, the system proceeds to SLS, displaying the feasible region and the values based on each imputation option. At this stage, the feasible region will be determined in part by the new value of APR. After completing the review of SLS, the system asks once again if the record is acceptable. If not, the analyst can repeat this process, but we expect one pass to suffice in most cases.

In addition to allowing the user to adjust the imputes, the system allows the user to delete alternative fields for edit-failing records. For example, the pattern (graph) of edit failures on some record might have looked like:



,

(where an arc between nodes indicates an edit failure between corresponding fields), and the automated system might have selected F_1 and F_4 for deletion based on pre-assigned weights, see [3] for details. If, on inspection, an analyst determined that field F_3 was in fact incorrect, then field F_3 would be targeted for deletion by the analyst, and the system would (depending on the assignment of weights) proceed to delete F_2 in order to remove remaining edit failures. Imputation will follow, and the system will ask the user if the revised record is acceptable, etc.

It is our expectation that this interactive system will prove to be an aid to analysts in the review process. By displaying the feasible region, the various system-generated options for imputation, and the source of each option, the interactive system will furnish the analyst with a range of information to bring to bear in the review of a referral case. By observing the influence of each correction on subsequent fields to be adjusted, the analyst will have a greater understanding of the impact of each revision. By providing guidelines for the analyst, this system can help reduce some of the tenuousness and subjectivity in the review process.

To date, analysts who have used this system on test decks have commented favorably and remarked that it is a system they can use to advantage. Note that once the core edit is set up to run records in batch mode, the interactive version is available with no extra effort. That is, when working with this system a user need only specify whether he/she wants to run records in batch mode through the automated version or on-line for referral cases.

IV. SUMMARY

To some extent, it was our intention to design an edit and imputation system that conforms to the guidelines set forth in the Introduction. But at the same time, the knowledge gained working with potential users in the subject-matter areas, learning their needs, and understanding the facets of their expertise, contributed to these guidelines. A edit and imputation system should blend statistical and subject-matter expertise in a coherent framework and integrate edit constraints with imputation strategy. We have described a structured system that attempts to meet these requirements and is sufficiently flexible to accommodate a variety of users. Development work continues on this system, enhancements are being made, and additional users are being identified.

Acknowledgment: The authors thank James O'Brien for his careful reading of earlier versions of this paper and for his many helpful suggestions.

This paper will be presented at the 1984 Annual Meeting of the American Statistical Association in Philadelphia, Pennsylvania and will appear in the Proceedings of the Section on Survey Research Methods.

References

1. Fagan, J. (1984). Developing a Family of Models for Selected Fields on the Annual Survey of Manufactures. Unpublished Manuscript, Census Bureau.

2. Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automated Edit and Imputation. JASA, 71, 17-35.

3. Greenberg, B. (1981). Developing an Edit System for Industry Statistics. Computer Science and Statistics: Proceedings of the 13th Symposium of the Interface, 11-16, Springer-Verlag, New York.

4. Greenberg, B. (1982). Using an Edit System to Develop Editing Criteria. Proceedings of the Section on Survey Research Methods, ASA, Cincinnati.

5. Sande, G. (1979). Numerical Edit and Imputation, International Association for Statistical Computing, 42nd Session of the International Statistics Institute.

6. Sande, I. (1982). Imputation in Surveys: Coping with Reality. The American Statistician, 36, 145-152.