SPEER (STRUCTURED PROGRAM FOR ECONOMIC EDITING AND REFERRALS)

Brian Greenberg and Thomas Petkunas*
Bureau of the Census, Washington, D.C. 20233

KEY WORDS: edit, imputation, expert systems, SPEER

I. INTRODUCTION

All survey data must be edited to detect improbable response combinations on a questionnaire, to make changes to keyed reported data when necessary, and to impute for missing items. The objective of editing is to detect and correct errors that may have been caused by a misunderstanding of a survey question, faulty reporting, or problems in data entry. In general, staff responsible for a particular survey designs the edit strategy and imputation methodology, develops procedures for linking the edit and imputation programs to the broader data processing system, and writes computer specifications for these activities. For a new survey, programs are typically designed from scratch and must be thoroughly tested prior to actual use -- a time consuming and labor intensive process. Programs are frequently highly complex so that detecting and correcting errors is a difficult task. For continuing surveys complexities compound over time and introduction of modifications to accommodate changes in questionnaire design can become difficult -- if not risky. The need for multipurpose edit and imputation programs has become increasingly clear.

The flow of a data record through all stages of error detection and correction is a combination of automated procedures, manual review, and an interactive combination of the two. The order of various features in the error detection and correction process varies from institution to institution. Programs for checking consistency between items on a questionnaire and imputing for missing data must be integrated with programs which assign and check Standard Industrial Classification (SIC) codes, evaluate geographic coding, and so on. We describe below the sequence of activities at the Census Bureau for data consistency checks for the typical economic establishment survey or census.

The data collection instrument is a printed survey questionnaire which is mailed to the survey universe to be returned to the Jeffersonville, Indiana facility of the Census Bureau where data is keyed. During data entry there are rudimentary checks to detect data entry errors. The data entry clerk is alerted to possible errors by edit checks and he/she can examine the form that has been keyed to determine if the value keyed was as reported or if there was a keying error. Errors can be caused by data entered into the wrong key-code, extra or not enough digits, incorrectly punched characters, or other problems of this sort. The data entry clerk is only responsible for correcting keying errors and is not responsible for corrections to respondent data. Detecting and preventing errors at the time of data entry is usually thought of as quality control of the data entry process. Keyed data are sent to headquarters in Suitland, Maryland where they are run through an automated batch edit program which detects inconsistencies, makes changes to the keyed respondent data, and imputes for missing responses, see Greenberg and Petkunas (1987).

Within the automated edit and imputation routines, selected records are targeted as referral cases and are directed for analyst review. The criteria typically are: (1) large change to reported data, (2) imputations for large establishments, and (3) unsuccessful imputation of a value that will pass tolerance checks. The analyst will review referral cases, make adjustments if needed, and send establishment records back through the automated edit and imputation routines. The automated routines may accept the analyst changes and send the record to the tabulation record file, or they may further adjust the data record. In the latter case, the system may send the revised record directly to the tabulation file or it may, once again, direct the record for analyst review.

During the review process, an analyst can accept or override actions taken by the automated system. The analyst will have the respondent's questionnaire, will be able to call respondents by telephone, and will have the use of alternative data sources to determine a reasonable number to impute for nonresponse or to adjust an assumed erroneous field value. Changes made by the analyst are often quite subjective and could be a source of subsequent edit failure. After an analyst writes the changes onto a referral document the changes are sent to Jeffersonville to be keyed onto the data records. Data records are run through batch edit processing again. A record can, once again, be targeted for changes and review, and this process can pass through several iterations before resolution.

The cycle of automated routines followed by analyst review and back again involves (1) processing records at headquarters, (2) sending referral listings to Jeffersonville, (3) hand corrections to referral documents, (4) keying of corrections, (5) sending corrections back to headquarters, and (6) a subsequent cycle of processing at headquarters. There are ample opportunities for delays and new errors (for example, in keying), and these multiply as the number of cycles grows.

The need for an on-line, interactive analyst review capability has been evident. The objective is for an analyst to key corrections directly onto the data record during the review process and have changes edited as they are entered. Such a capability would streamline the review process, make it more efficient, and reduce further errors. We have added an on-line, interactive capability into SPEER, which is described in Section III. This capability makes the review process more effective and time and cost efficient.

The SPEER system was developed to adapt the innovations in editing made at Statistics Canada by Fellegi and Holt (1976) and Sande (1976,1978, and 1981) to economic establishment surveys under ratio edits. Although SPEER was designed for automated batch processing, the system has evolved the capability to perform interactive review of referral records and interactive data entry. The system can incorporate a wide variety of user-specific, user-specified requirements and is sufficiently flexible to accommodate diverse user expertise within a coherent structure. The design of SPEER has moved into the area of expert systems in an attempt to integrate mathematical methodologies with subject-matter expertise. Survey staff are extremely knowledgeable about the survey questionnaire, the target population, and in many instances specific sources of response error and nonresponse. The SPEER structure has been designed to allow this expertise to be incorporated into the SPEER edit and imputation programs.

Survey staff have a rather proprietary feeling about their data -- and on balance that feeling is valuable. They have a great deal of expertise which they bring to processing tasks and they pride themselves on producing the best products they can. They will not willingly trust their data to a "black box" to which they cannot contribute and over which they have little control. In the end, the survey staff does bear responsibility for the data products and they must know that their special expertise is being utilized. In the design of a multipurpose edit and imputation system one must pay careful attention to user acceptance -- and acceptance is enhanced by flexibility and comprehensibility.

One of the salient benefits in a multipurpose edit and imputation system is that a wide range of survey staffs can partake of advances in edit methodology. To the extent that rigorous editing methods and processing procedures form the core of a general system, these techniques can be brought to users who otherwise would not have ready access to them. Moreover, a multipurpose edit and imputation structure which links edit and imputation functions while not locked into any single imputation method will give users the opportunity to test and evaluate diverse techniques for imputation. In this discussion, we do not concentrate on any particular imputation methodology within SPEER but rather address the edit system as a whole and regard imputation as a user defined component.

In Section II, we describe SPEER capabilities, structure, and basic methodology. We do not go into great depth and refer the reader to Greenberg (1981, 1982, 1987a and 1987b) and Greenberg and Surdi (1984) for more detail on methodology. We describe the expert system aspects of SPEER and discuss how working with users led to the evolution of the SPEER system. In Section III we describe the on-line, interactive features in SPEER for review of referral documents and as a Computer Assisted Data Entry device -- CADE, in the emerging jargon. In Section IV we describe actual uses of SPEER.

II. METHODOLOGY IN SPEER

SPEER is a multipurpose edit and imputation system for numeric data under ratio edits. For our purpose, a typical establishment record will consist of a vector with numeric data fields

$$(x_1,\ldots,x_n).$$

A ratio edit is the requirement that the quotient of two field values lies between prescribed bounds which are read into the system as parameters. A typical ratio edit is of the form

$$L_{ij} \leq x_i/x_j \leq U_{ij}$$

where L_{ij} and U_{ij} are the lower and upper allowed limits for the ratio of x_i to x_j. For example, the ratio of the annual total salaries paid to construction workers divided by annual total hours worked by construction workers must be within reasonable limits.

SPEER is divided into four main components: Edit Generation, Edit Checking, Error Localization, and Imputation.

If

$$L_{12} \leq x_1/x_2 \leq U_{12}$$

$$L_{23} \leq x_2/x_3 \leq U_{23},$$

are two ratio edits, the implied edit is

$$L_{12}L_{23} \leq x_1/x_3 \leq U_{12}U_{23}.$$

Starting with a set of user supplied explicit edits the Edit Generation subroutine first derives all implied edits. These are returned to the survey staff so they can evaluate the logical implications of the ratios and bounds they provided. At this stage inconsistencies in the user-supplied bounds can be detected and any unexpected implications of the explicit edits can be examined. Adjustments to the bounds are made and the revised limits are processed through the edit generator for subsequent analysis. This process can be repeated. After subject-matter specialists are satisfied with the explicit ratios they are entered into the edit routines as parameters. Note that since there are n fields and each edit consists of exactly two fields, there will be

$$\binom{n}{2} = n(n-1)/2$$

edits and any pair of fields will be jointly contained in some edit.

The implied edits allow for multiway comparisons between fields to aid in determining potentially erroneous values. For a general discussion of the uses of implied edits for both categorical and continuous data, see Fellegi and Holt (1976). In Greenberg (1981 and 1982) we show how edits are generated for SPEER and provide a number of examples.

Edit checking is a very simple operation; the program determines which edits pass or fail for a given record. The full set of edits -- both explicit and implied -- are used in the Edit Checking routine. If all edits pass and no data values are missing, the record is considered acceptable. If no edits fail but some data items are missing the record is sent for imputation of the missing fields. In addition to the use of current reported data for edit checking, data can also be checked against prior year data or against administrative data values. Prior-year edit checks are extensively used in the Annual Survey of Manufactures adaptation of SPEER -- see Section III and Greenberg (1981).

If one or more edits are failed by a record, the record is sent to Error Localization to determine a set of fields to delete so that the remaining fields will be mutually consistent. That is, the remaining fields will jointly fail no edits. Typically the objective is to delete as few fields as possible. Fields can be weighted to reflect their overall reliability with more reliable fields having a higher weight. The objective then becomes to delete a weighted minimal set of fields so that the remaining are mutually consistent.

Each ratio edit involves exactly two fields and the error localization routines in SPEER take advantage of this structure. We represent the pattern of failed edits by a graph in which fields correspond to nodes and arcs represent failed edits between the corresponding nodes. The graph in Figure 1 indicates that field 3 failed edits with fields 1, 2, 4, and 5; field 4 fails edits with fields 2 and 3 and so on.

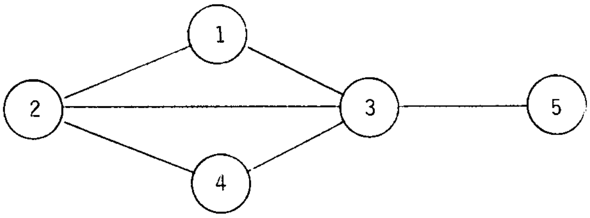


Figure 1. Failed Edit Graph

The objective is to delete a subset of the nodes in this graph so that there are no arcs remaining; thus there are no remaining failed edits. A (minimal weighted) set of nodes selected for deletion in the graph corresponds to a (minimal weighted) set of fields to target for correction. Let us assume that corresponding to the diagram above, field 2 and field 3 are targeted for deletion. As there are no arcs joining the remaining nodes the corresponding fields are mutually consistent.

One can select imputation values for the deleted fields so that all fields on a record will be mutually consistent. That is, it is possible to assign values to fields 2 and 3 in such a way as to ensure that they are consistent with each other and the remaining fields so that all fields are mutually consistent. The procedure for selecting nodes to remove from the failed edit graph is described in Greenberg (1981). New field values are assigned in the imputation subroutines of SPEER.

The underlying methodology in the Error Localization routine employees the methods introduced in Fellegi and Holt (1976). A failed edit matrix is set up in which rows of the matrix correspond to failed edits and columns correspond to fields. To find a minimal weighted set of fields to revise on an edit failing record, one solves, in principle, a Set Covering Problem. A discussion of the Set Covering Problem as applied to error localization for automated edit and imputation is contained in Garfinkel, Kunnathur, and Liepins (1986) and Liepins, Garfinkel and Kunnathur (1982).

Suppose (after re-ordering if necessary) that fields $x_1,\ldots,x_k$ for $k \leq n$ were reported and not targeted for change by the error localization subroutine. This means that they are mutually consistent. In particular, for all $i,j \leq k$, the ratios

$$L_{ij} \leq x_i/x_j \leq U_{ij},$$

are satisfied, that is, all edits involving these fields are satisfied; so these fields are mutually consistent. If k=n, then the complete record is consistent. Assume now that k<n, and let us establish an imputation range for x_{k+1} . Note that for all $j \leq k$, we have the ratio

$$L_{k+1,j} \leq x_{k+1}/x_j \leq U_{k+1,j}$$

and by multiplying through by x_j we also have the pair of inequalities

$$x_j L_{k+1,j} \leq x_{k+1} \leq x_j U_{k+1,j}$$

where x_j, $L_{k+1,j}$ $U_{k+1,j}$ are known constants for all $j \leq k$. Each $j=1,\ldots,k$ determines an interval in which x_{k+1} must reside to be consistent with x_j. Thus, if x_{k+1} lies in the intersection of the k intervals defined above, it will be consistent with each of the fields x_j for $j=1,\ldots,k$. Since the explicit edits are consistent, the intersection is not empty, see Greenberg (1981), and it is referred to as the feasible region for field x_k. This region can be represented by the shaded area below, where the parenthesis represent upper and lower limits for x_{k+1} based on the various x_i . That is, left and right paranthesis represent, respectively, $x_j L_{k+1,j}$ and $x_j U_{k+1,j}$ for each $j \leq k$. Note that the innermost bounds for the feasible region do not necessarily come from the same ratio edit.

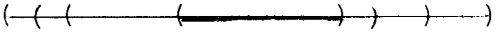


Figure 2. Feasible Region

Under each implementation to date, fields have been imputed sequentially. For each field to be imputed -- whether missing or deleted due to edit failures -- the feasible region is derived. An imputation value is selected which lies within the feasible region and thus will be consistent with every other field value on the record; either reported and accepted or imputed at an earlier stage. The imputations are based on strategies selected by subject-matter specialists and they are incorporated into the SPEER system. Since each imputation will lie within the feasible region, one can guarantee that no imputed value will fail the edits.

In each actual use of SPEER, the routines for Edit Generation, Edit Checking, and Error Localization have remained (virtually) unchanged. However the imputation procedures were different for each implementation. The imputation rules are designed by survey staff based on special considerations and appropriate statistical procedures. For example, in some well-defined cases a blank can be reasonably inferred to represent a zero and one imputes a zero for blank in these cases. At times respondents report in different units than specified by the instructions. In these cases, the imputation is the resulting conversion to requested units. Administrative data form the basis for imputation in other cases. Regression models in which the field to be imputed is the dependent variable can be employed (Greenberg and Surdi 1984). Whatever the methods used, for each field to be imputed an imputation module is created which contains a sequence of imputation rules provided by survey staff for that field.

The following schematic may represent the SPEER structure:

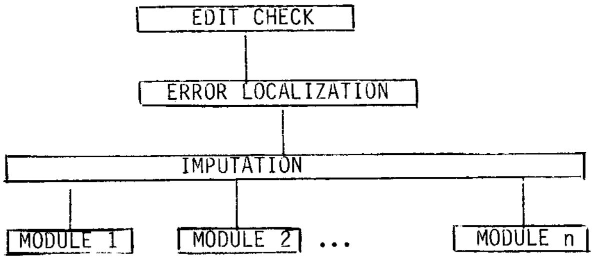


Figure 3. SPEER Structure

In the segment labeled "IMPUTATION" the system derives the feasible region for imputation as was described earlier. The sequence of imputation rules for field x_i is embedded in Module i. In addition to the highly field specific rules as described above, each module contains a simple regression model which can be used as a generic imputation in the absence of applicable expert rules and each contains a default imputation.

Suppose a given field is selected for imputation. First the feasible region for the missing field is computed. Next the program reaches into the imputation module to obtain candidates for the value to be imputed. The first applicable rule is examined and an imputation is derived based on this rule. If the derived value falls within the feasible region, it is accepted as a valid imputation. If not, the second rule is accessed and an imputation value is derived and checked against the feasible region. This continues until an acceptable value is reached. The value ultimately selected as the imputation will typically be derived from subject-matter based rules and this value will be consistent with all other fields on the record because it is forced to lie within the feasible region. If no rule supplied by subject specialists provides an acceptable imputation, a feasible default imputation is selected and the record is targeted for analyst review.

Let us provide an example of what a rule sequence might look like. Suppose one is to impute for a field such as Annual Payroll (APR) on an economic census or survey. For concreteness, let us couch our discussion in terms of the 1987 Economic Censuses. Under the first rule the system might derive an imputation based on 1987 Administrative Data value for APR. If that value does lie within the feasible region, it is accepted as the imputation for field APR. If the value derived does not lie in the feasible region the system might next derive a candidate imputation based on 1986 Administrative Data for APR. If the candidate based on 1986 Administrative Data is not suitable we pass to a third option; for example, a regression model using one or more related fields might be employed to derive a candidate imputation. This examination of candidate imputation values continues until a candidate is examined which lies within the feasible region.

Imputation rules can be extremely field-specific. For example, suppose some field is to be reported in tons. Assume that the feasible region allows valid responses to be between 500 and 1,000 tons and the value 1,800,000 was reported and deleted as an error. The first applicable option might be to divide the reported value by 2,000 based on analyst information that respondents sometimes report in pounds rather than tons. We would derive 900 tons and observing that this value is feasible accept it as the valid imputation. A common error in reporting economic data is that respondents sometimes provide answers in dollars rather than in thousands as per instructions. For fields in which this error may occur, the first rule is usually to divide the reported response by 1,000. By having the feasible region to examine, one can infer if dividing the reported value by 1000 yields an acceptable value to use for correcting the field. The feasible region serves as a screening tool in determining whether to accept any candidate imputation.

We provide a few examples of survey specific rules taken from the 1987 Census of Construction Industries. Two fields are Cost of Materials (CM) and Subout (SO) -- payments made to subcontractors. For some establishment records, CM may be fairly high and SO may be blank (i.e., not reported). The subject-staff inference is that the reporting establishment does the actual construction activities and does not subout to contractors. Thus, SO will be set to (imputed as) zero. Conversely, if SO is high and CM is blank, the reasonable inference is that the reporting unit is a general contractor and CM is to be imputed as zero. If they are both blank, the arguments above do not pertain.

Another example from this census concerns the fields Supplementary Labor Costs (SLC), Voluntary Payments (VP) and Legally Required Payments (LE). The field LE consist of government required additional labor cost such as social security payments, unemployment taxes, and a few others. The field VP covers expenses such as health plans, retirement fund contributions and so on. The total

$$SLC = VP + LE$$

is supplemental labor costs. On some records, due to misunderstanding of the question, respondents put all of VP into LE. Hence, VP = 0 and LE is detected as high, but the value for SLC (which now equals LE) is a reasonable value. When such a case is detected, a portion of the LE value is moved into VP and LE is adjusted as well. This type of action would generally have been taken by an analyst when reviewing a record. This is an example of another type of rule we have incorporated it into the SPEER code. Examples of the use of these subject-matter decision rules built into the imputation protocols abound for each implementation of SPEER.

In addition to the very survey - specific rules noted here, there are also rules of a more general nature which can be used across surveys. Under the adaptation of SPEER for the Annual Survey of Manufactures, a family of regression models is employed. The variable to be imputed is viewed as the dependent variable with one or more correlated fields serving as independent variables. If the independent variables are present, the model can derive a candidate imputation value. When the prior year value of the missing field is available, this variable is also used in the model.

The imputation modules (shown in Figure 3.) contain the survey and field specific expert information as provided by subject-matter specialists. In the terminology of expert systems, they form the knowledge base. Although SPEER was not initially designed to be an expert system, the need for an expert system philosophy in SPEER became apparent very quickly when working with prospective users. Systems currently in use do incorporate subject-staff expertise through a sequence of rules. Subject staff are typically very reluctant to replace such systems with programs that do not have the capabilities to take advantage of their experience and knowledge.

The mathematical procedures embedded in the other SPEER system components form the driver routines which access the knowledge base and assist in selecting from among the decision rules. We have described the ideas taken from an expert system structure which allow us to blend subject-matter expertise with mathematical procedures. In SPEER the mathematical procedure and the subject-matter rules can be treated as separate. One can extend the mathematical methods and revise the flow of the system as a whole, unencumbered by survey-specific considerations. The survey-specific rules can be examined in their own right; updated and revised as needed, independently from the programs through which they are applied. On the other hand, the mathematical procedures and decision rules are integrated. The mathematical procedures provide a framework to assist in choosing the most appropriate decision rule and to ensure that the value imputed will pass all applicable edits. As an expert system for edit and imputation SPEER does more than provide a vehicle for accessing expert rules; it also provides a mathematical framework to help decide from among the rules, choosing only rules which are valid for the record under consideration.

The SPEER programs can handle a large number of variables. The variables are typically divided into basic and secondary. The basic variables are those fundamental to the operations of an establishment and these are edited jointly in a core program as described above. The secondary items are grouped into satellites consisting of related items and these items are edited against one another in each satellite. First the basic items are edited against each other and then the satellite items are made to conform to the basic items as well as each other in the same satellite.

In the actual implementations of SPEER, building imputation modules can be a fairly time-consuming process. SPEER staff works closely with subject-matter specialists to elicit their expertise to design a hierachey of imputation rules. Attempting to convert the experience of subject analysts into a sequence of rules that cover a wide range of imputation scenarios is a difficult process. When the process is complete, the subject-matter staff does understand the new edit and imputation system. They have a system that is relatively easy to change, update, or revise when necessary. The edit and imputation program resides with the operating division as their product.

III. ON-LINE, INTERACTIVE SPEER

The Enterprise Statistics program consists of a series of publications based on data collected in the censuses of Wholesale Trade, Retail Trade, Service Industries, Manufacturers, Mineral Industries, Construction Industries, and selected Transportation industries. Two of the reports include: the Large Companies publication and the Auxiliary Establishment publication.

The Large Companies publication is based on responses to questionnaires sent to companies with 500 or more employees in the industries above. The published tables show selected financial statistics of large companies. The Auxiliary Establishments publication presents data on auxiliary units of multi-establishment firms. The primary functions of auxiliary establishments are to manage, administer, service, or support the activities of the other establishments of a company. Examples of auxiliary establishments are research and development centers, warehouses, and administrative offices. Published tables furnish detailed financial statistics of auxiliaries by industry classification of management or supporting service functions they provide, their employment size, and their geographic location.

The first implementations of SPEER were for the 1982 Enterprise Summary Report (ES-9100) and 1982 Auxiliary Establishment Report (ES-9200). The Enterprise Summary Report provides data for the Large Companies publication and the

Auxiliary Establishment Report provides the data for the Auxiliary Establishments publication.

The decision was made to employ SPEER once again to edit these two report forms for the 1987 Economic Censuses. We employed the most current version of the SPEER programs in order to take advantage of the newly developed on-line, interactive capabilities for review of referral cases. Two versions of SPEER were designed; one for each of these programs. As subject matter staff were familiar with SPEER requirements due to work on the 1982 censuses, the development process went fairly quickly.

When the SPEER interactive routines are used for processing of referral record, the system converses with the analyst using it. The analyst can override the decision rules residing in the batch version of the system and replace them based on his/her expertise and auxiliary information about the case under review. Using this system, the analyst accesses a specific record and reviews the processing done by the automated system. The analyst typically has the original respondent questionnaire, can call the respondent by telephone, or can access other related information not on the establishment data record. Based on this additional information and his/her own experience, an analyst may overrule the decision rules built into the automated system.

The batch version of SPEER and the on-line, interactive versions have the very same underlying program. The difference is that the interactive version of the program pauses at selected points in the code to wait for input from the keyboard as it is in conversational mode with the user. In the remainder of this section, we will describe features in the interactive version and discuss how it is used. The system is menu-driven and allows a user to interact with all SPEER subroutines. The exact interactive capabilities as well as screen display were dictated by subject specialists in the Enterprise Statistics Branch, Economic Surveys Division, with whom we worked closely throughout.

When using the interactive version of SPEER to review a referral record, the analyst will indicate which field he/she wishes to examine. The program can display the current residing field value, the reported value (if any), candidate values derived from each imputation option, and the range of the feasible region. Provided with these system guidelines, the analyst has information at his/her disposal to assist in the decision making process and will select an imputation value for the field under review. If there is reason to believe that the most appropriate imputation value lies outside the feasible region (for example, because of explanatory notes on the form or through a call-back to the respondent), the analyst has the option of entering an imputed value outside the range.

If there is a second field to be reviewed on this record, the program can display on the monitor the feasible region, currently residing value, the reported value (if any), and candidate values for imputes derived according to each option as it did for the first field. Note, however, that each of these values is based, in part, on the new value of the revised first field. As above, the analyst will determine an appropriate imputation value to enter and move on to the next field, if any. After all fields have been examined and adjusted if needed, the review is complete. The revised record will be consistent and no further batch processing will be required.

In Figure 4 we show a very simplified version of the interactive imputation display which can be seen by an analyst. Using the 1987 Economic Censuses as example, consider the review of Annual Payroll (APR). The display shows an acceptable range (the feasible region) for APR from 250 to 750. The current value as was derived by the automated system is 375. The next value is the actual reported value of 82 followed by the value derived from 1987 Administrative Data and the candidate imputations based on 1986 Administrative Data, 1982 Economic Census data, etc. The blank next to 1982 Economic Census position indicates that the 1982 Economic Census value was not available to derive a candidate imputation. The values of 225 and 180 are those derived from the appropriate regression models. The "average value impute" is based on the average value of the ratio of APR and some related field. The ordering above reflects the order in which the rule options are applied by the batch version of the system. Note the "current value" corresponds to the first acceptable candidate imputation option. Observe also that the two regression models yield values outside the feasible region.

By having the range in which the imputed value must fall to be consistent with all fields, plus a variety of options, the analyst then has a significant amount of information at his/her disposal to assist in the review of a referred case. The analyst can supply an imputed value other than one of those shown below through the use of "Option 8". He/she can also impute a value outside the feasible region and have the system accept the value through the use of a multiplier. A multiplier is used to extend the range of the feasible region by extending the limits of the ratio edits. For example, the analyst may discover through a call-back that the reported value is actually correct and reinstate the value 82.

```
                 IMPUTATION OPTIONS FOR APR

                 A. RANGE OF APR:                (250,750)
                 B. CURRENT VALUE:               375
                    OPTIONS
  1. REPORTED VALUE:                             82
  2. 1987 ADMINISTRATIVE DATA:                   375
  3. 1986 ADMINISTRATIVE DATA BASED:             430
  4. 1982 CENSUS DATA BASED:
  5. REGRESSION MODEL 1:                         225
  6. REGRESSION MODEL 2:                         180
  7. AVERAGE VALUE IMPUTE:                       403
  8. ANALYST SUPPLIED VALUE:
```

Figure 4. Interactive Imputation Display

Through whatever means, the analyst may determine a revised imputation for field APR and enter it on the data record. This value is accepted by the program and field APR is considered to be completed.

If there is a second field to be reviewed on this record, for example, Number of Employees (EMP), the program once again can display on the terminal screen the feasible region for EMP, currently residing value, and candidate values for imputes derived according to each option; as it did for APR. Note that the imputation rules for field EMP will be different from those used for APR. The feasible region and each of the candidate imputation values is based, in part, on the new value of APR just entered by the analyst. As above, the analyst will determine an appropriate value for EMP, enter this value, and move on to the next field to be reviewed, if any. After all fields needing review have been examined and adjusted if needed, the review is complete. The revised record will be consistent, and no further batch processing will be required. The analyst will return the completed record to the data base and select the next record for review.

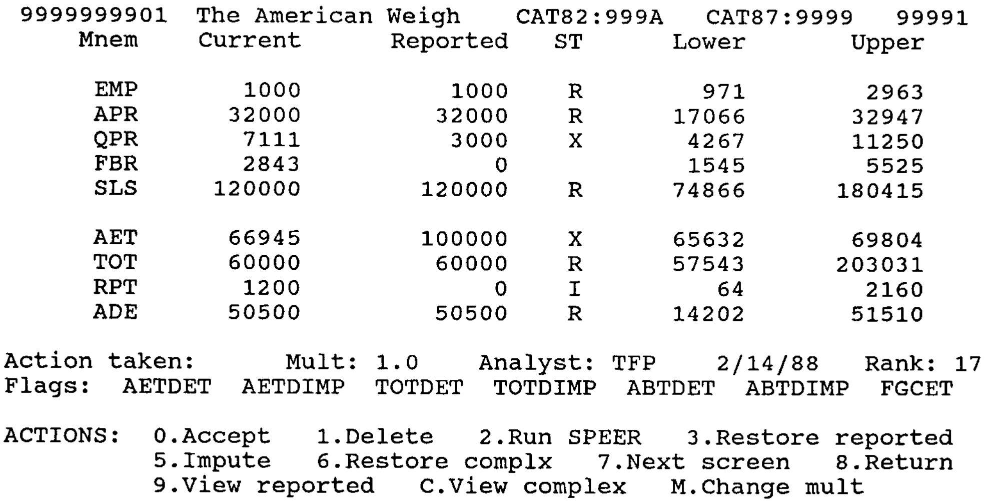
One of the most valuable features for the survey staff was the display of the feasible region. This information served as a guide in the selection of imputation options. If the feasible region was, for example, the interval (100,350) for some field, the analyst could enter a value of, for example, 500. The system would alert the analyst that the value was outside the range and ask if that value was in fact desired. If the response was "yes", the range would be increased to (approximately) the new interval (68,506). This is done through the use of a multiplier, and the upper limit would be multiplied by 10/7 and the lower by 7/10 and a little "margin of error" would be added. For the record under review every ratio containing that field will also have its upper and lower bounds expanded so that the new, previously "out-of-bounds" value will no longer be out of bounds, and hence fail no edits. The ratios are reset to initial values for the next record. Thus, a non-typical record can be made acceptable to prevent unnecessarily forcing conformity to prior assumptions.

The important observation from the perspective of an expert system is that a true expert (the analyst) converses with the automated expert program in order to augment the system expertise and override decision rules as needed. Analysts have found this system easy to use, and it makes their decisions in the review of establishment records less tenuous than has previously been the case. The design of the screen and many of the variations in the system were based on requests from survey staff. The display shown above is an early version of the imputation screen.

The interactive SPEER does much more than allow analysts to alter imputation decisions made by the batch programs. The program is menu driven with a large number of options, which we describe below. In Figure 5. we display Screen One of the interactive SPEER used for referral cases on the Enterprise Summary Report, ES-9100. Screen One covers the basic items on the ES-9100 report form. This is where a major portion of an analyst's time will be focused because the basic items contain much of the important information about an establishment. The satellite items are treated on subsequent screens. There are a total of 3 screens, 9 basic items, and 46 satellite items in the ES-9100 edit programs. The top line contains specific information identifying the company. Included in this header line are the Census File Number, the name of the establishment, the 1982 category code, the 1987 category code, and the Microfilm Reference number.

The first column on the left displays the mnemonics for each basic item: Number of Employees (EMP), Annual Payroll (APR), First Quarter Payroll (QPR), Fringe Benefits Required (FBR), Sales (SLS), Total Ending Assets (AET), Total Assets (TOT), Total Rental Payments (RPT), and Accumulated Depreciation for the End of Year (ADE).

The next two columns display data values for each of the basic items. All dollar values are displayed in thousands of dollars. Column two shows the data values as they appear after they are run through the complex edit batch



```
  9999999901  The American Weigh   CAT82:999A  CAT87:9999   99991
     Mnem     Current       Reported   ST       Lower        Upper

      EMP        1000           1000    R         971         2963
      APR       32000          32000    R       17066        32947
      QPR        7111           3000    X        4267        11250
      FBR        2843              0             1545         5525
      SLS      120000         120000    R       74866       180415

      AET       66945         100000    X       65632        69804
      TOT       60000          60000    R       57543       203031
      RPT        1200              0    I          64         2160
      ADE       50500          50500    R       14202        51510

Action taken:        Mult: 1.0    Analyst: TFP     2/14/88    Rank: 17
Flags:  AETDET  AETDIMP  TOTDET  TOTDIMP  ABTDET  ABTDIMP  FGCET

ACTIONS:  0.Accept    1.Delete    2.Run SPEER    3.Restore reported
          5.Impute    6.Restore complx    7.Next screen    8.Return
          9.View reported    C.View complex    M.Change mult
```

FIGURE 5. Display for ES-9100 Screen One

programs. Values in this column are considered to be consistent with each other. Next are the originally reported values for each basic item for this company. For those items whose reported value differs from the current value, the current value is highlighted.

The fourth column displays the current status flag for each basic item. There are five different status flags: reported greater than zero and passed edits (R), reported greater than zero and changed by SPEER edit (X), reported greater than zero and set equal to zero by SPEER edit (Z), imputed to a positive value from zero (I), and a nonresponse set of zero (N). The final two columns display the lower and upper limits of each basic item's feasible region. A value must lie in this region to be considered consistent with all other basic items.

The first line following the data displays the record multiplier, the analyst's identification, today's date, and the company's rank. The company's rank shows how large this particular company is in relation to the entire universe. Typically, companies with a rank of 10 or higher will be given more attention from the analyst.

Flags displayed on the next line describe changes to the entire record -- not just the changes for the basic items on screen one. This allows the analyst to see a snapshot of the entire record without scrolling through all the screens. For example, the flag FGCET tells the analyst the field Capital Expenditures (CET) has been changed by a substantial amount. The definition of a "substantial amount" being decided beforehand by subject-matter specialists.

The menu contains 11 actions designed by the subject-matter specialists. We describe each of them below.

0. Accept:
This option is used to indicate that the present status of the record is acceptable. This may be the status directly after the batch run with no analyst action or it may be after analyst changes have been made. When this option is entered, the record is sent back into the database where it remains until tabulation.

1. Delete:
This option is designed to remove a record from the database. When this option is entered, a flag is set and remains with the record. The record is then returned to the database where it will remain until a batch program deletes all records with this flag. Since the Delete option does not actually remove the record from the database instantaneously, it is still possible for the analyst to access this record if needed.

2. Run SPEER:
Invoking the SPEER edit allows the analyst to immediately see how the changes he has made will effect the rest of the record. This option also allows the analyst to perform a number of "What if's". That is, the analyst can try a number of alternatives to see how each one will affect this record.

3. Restore reported:
This option reinstates the originally reported data with one key stroke. This option is useful for records that may not conform to the edits but whose reported data are determined to be correct. This eliminates entering reported data for every field.

5. Impute:
This option blanks out all values to allow the analyst to impute an entire record from just one or two specified data values. Typically an analyst will use this option to impute the entire record from fields EMP and QPR using data from administrative records.

6. Restore complex:
This option reinstates the data values as they were originally displayed at the start of this session.

7. Next Screen:
This option displays the next screen which contains other data items, typically satellite and detail items. The subsequent screens also have menus and enable the analyst to revise data.

8. Return:
This option returns the record to the database to be reviewed again at a later time. This is helpful if an analyst needs more information to review a referral but that information is not immediately available. The analyst can go on to another referral and come back to this one when that information is available.

9. View reported:
This option displays all the originally reported values on one screen. This gives the analyst a picture of the entire establishment without paging through all screens.

C. View complex:
This option displays all the current edit values on one screen. Again, this gives the analyst a picture of the entire establishment without paging through all screens.

M. Change mult:
This allows the analyst to manually change the multiplier for this record. This will override the multiplier that is currently used, whether is was calculated by the SPEER edit or calculated manually by an analyst. The analyst can also set the multiplier equal to "infinity" which would allow the entire record to pass edits. This can be done when restoring the reported data.

Actions 1, 3, 5, and 6 have safeguards incorporated into them. It takes two keystrokes to invoke these actions. After selecting one of these actions, a bell sounds, the menu disappears, and a message is displayed. This guards against an analyst overwriting the current data by mistake or deleting a record from the universe by accident.

The SPEER programs can generate a large quantity of diagnostic information on a record-by-record basis. The choice of diagnostics to be displayed on the screen is one of the options given to survey staff.

In addition to a large amount of diagnostic information available to the analyst at time of record review, information is also available to managers to monitor the review process. Information is available on the performance of individual analysts and between analysts. For example, one can monitor how often each analyst employed a multiplier, accepted the automated system actions, over-ruled the system, made a telephone call to the respondent, and so on. In addition to monitoring performance of the individual analyst to evaluate his/her work, one can observe similarities across analysts. One use of this capability is to detect the frequency with which the automated system has been over-ridden by the analysts to determine if changes should be made in the automated system. As far as we know, these capabilities are the first to monitor the activities of individual analysts, evaluate their performance, and use this information to understand and perhaps improve this highly subjective and important process.

SPEER is written in fairly simple FORTRAN and is easy to transfer from one operating system to another and the programs were adapted to microcomputers with no difficulty. The batch version of SPEER for the ES-9100 and ES-9200 questionnaires was run on the UNISYS operating system primarily due to communication lines established between headquarters in Suitland and the Jeffersonville processing center. After records were run in batch mode on the UNISYS mainframe, referral cases were down-loaded to a local area network. Analysts performed their review of referral cases using IBM microcomputers connected through the local area network sharing a single database.

As noted above, when the interactive system is used, new data values are edited at the time they are entered onto a record. This capability led to the development of an on-line data entry and edit program. This SPEER data entry system has been used by Industry Division for the 1989 Annual Survey of Manufactures for late adds.

The Annual Survey of Manufactures (ASM) provides for intercensal year estimates of key measures of manufacturing activity for industry groups and important industries. These key measures, as well as other detailed statistics for manufacturing, are collected in the censuses of manufactures. An annual survey has been taken each of the years between censuses starting with 1949. During intercensal periods, these annual surveys provide a continuous series of basic statistics for industries and they furnish benchmarks for current business indicators and for measures of industrial production and productivity.

After a certain time in the processing of any survey at the Census Bureau, data capture activities for that particular questionnaire are closed down in Jeffersonville. Records received after data entry facilities are closed at Jeffersonville are referred to as late adds and must be entered onto the database by the analysts at headquarters. This is a time-consuming and costly process.

Staff responsible for the Annual Survey of Manufactures requested an interactive version of SPEER for data entry for late adds. Using this system on microcomputers, data are edited as they are being entered, hence there is no need for further batch editing. The system is currently being expanded and it will be transferred to the VAX. The programs are menu driven and follow the basic SPEER structure with specialized screens and options designed according to the needs of the Annual Survey of Manufactures staff for this specific purpose.

IV. IMPLEMENTATION EXPERIENCE

Work started in 1980 on what evolved into the SPEER system. The original objective was to design programs for the Annual Survey of Manufactures which incorporated the advances in methodology made by Fellegi and Holt and by Gordon Sande at Statistics Canada. We worked very closely with the staff in Industry Division to design an Annual Survey of Manufactures prototype. We initially had no intention of developing a multipurpose, multi-user system.

We were approached by Enterprise Statistics staff in Economic Surveys Division to see if we could adapt these programs to edit the 1982 Auxiliary Establishment Report and the 1982 Enterprise Summary Report. The system was adaptable and was successfully used for this purpose. Shortly thereafter the programs were again modified and used to edit the Manufacturing, Wholesale, Retail and Service Censuses for the 1982 Economic Censuses of Puerto Rico. Each time we used this system enhancements were made to the programs and about this time the name SPEER was adopted and we began to focus more on the multi-user aspects of the programs.

The next major activity was to modify SPEER for the 1987 Census of Construction Industries in a project spread over two years. Construction Surveys Division programmers were assigned to work on the project along with survey staff to ensure that the SPEER expertise resided in Construction Surveys Division after the project was completed. At the same time staff from Industry Division revisited the work done earlier and they designed an edit system for the 1986 Annual Survey of Manufactures and the 1987 Census of Manufactures along the lines of SPEER. The edit programs based on SPEER methods have subsequently been used on the 1988 and 1989 Annual Survey of Manufactures. As discussed earlier, we are currently working with Industry Division Staff to develop an interactive data entry system for late adds.

We next revisited the work with Enterprise Statistics and employed SPEER for the 1987

Summary Enterprise Report and the 1987 Auxiliary Establishment Report. These applications saw the first use of the interactive edit review capabilities for referral cases.

For each application, the programs became the "property" of the operating divisions. Each division is responsible for maintaining, updating, and using the system in subsequent surveys and censuses. We did not wait until we had a full-blown system with all desirable features before we ventured to use it. In a sense the system has been under continual development. The direction for change has been dictated by the needs and requests from users. It is in this respect that we view the SPEER programs as having evolved into the system described in this report.

Acknowledgment

The authors wish to thank David D. Chapman, Lisa Draper and Magdalena Ramos for their valuable comments on an earlier version of this paper.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

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

GARFINKEL, R. S., KUNNATHUR, A. S., and LIEPINS, G. E. (1986), "Optimal Imputation of Erroneous Data: Categorical Data, General Edits," Operations Research, **34**, 744-751.

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

GREENBERG, B. (1982), "Using an Edit System to Develop Editing Specifications," Proceedings of the Section on Survey Research Methods, American Statistical Association, 366-371.

GREENBERG, B. (1987a), "Edit and Imputation as an Expert System," in Statistical Policy Working Paper Number 14: Statistical Uses of Microcomputers in Federal Agencies, Statistical Policy Office, Office of Information and Regulatory Affairs, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C., 85-92.

GREENBERG, B. (1987b), Discussion, "Session on Designing Automated Data Editing Systems," Proceedings of the Third Annual Research Conference, Bureau of the Census, Washington, D.C., 204-212.

GREENBERG, B., and PETKUNAS, T. (1987), "An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division," in 1982 Economic Censuses and Census of Governments Evaluation Studies, U. S. Department of Commerce, Bureau of the Census, 85-98.

GREENBERG, B., and SURDI, R. (1984), "A Flexible and Interactive Edit and Imputation System for Ratio Edits," Proceedings of the Section on Survey Research Methods, American Statistical Association, 421-426.

LIEPINS, G. E., GARFINKEL, R. S. and KUNNATHUR, A. S. (1982), "Error Localization for Erroneous Data: A Survey," TIMS Studies in Management Sciences, **19**, 205-219.

SANDE, G. (1976), "Numerical Edit and Imputation," invited paper presented at the International Association for Statistical Computing, 42nd session of the International Statistical Institute, Manila, Phillipines.

SANDE, G. (1978), "An Algorithm for the Fields to Impute Problems of Numerical and Coded Data." Statistics Canada Report.

SANDE, G. (1981), "Descriptive Statistics Used in Monitoring Edit and Imputation Processes," presented at Workshop on Automated Edit and Imputation, Computer Science and Statistics; Proceedings of the 13th Symposium on the Interface, Springer-Verlag, New York.