United States
Department of
Agriculture

National
Agricultural
Statistics
Service

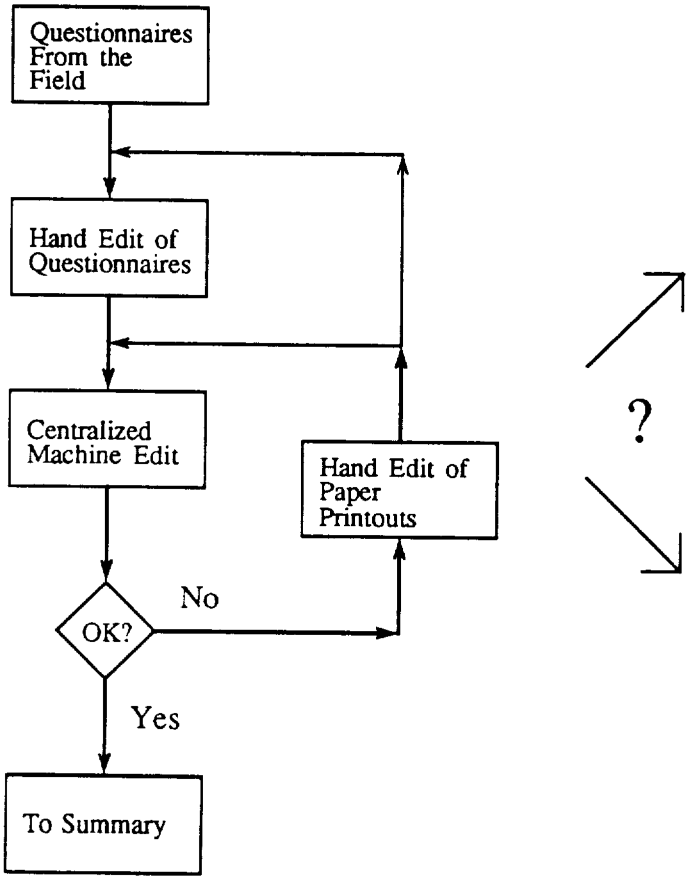
Research and
Applications
Division

NASS Staff Report
No.SRB-88-10

September 1988

# A Review of the State of the Art in Automated Data Editing and Imputation

Mark Pierzchala


Questionnaires From the Field
Hand Edit of Questionnaires
Centralized Machine Edit
Hand Edit of Paper Printouts
OK?
No
Yes
To Summary
?

CATI
CAPI
Paper Questionnaires
Heads Up Data Entry
Integrated Survey Processing
Specifications Generator
Database Environment
Distributed Processing
Centralized Processing
Automated Correction
Computerized Editing Tools
Interactive Processing
Statistical Edits
Macro-edits
Expert Systems
Objectivity
Repeatability
Defensibility
Greater Productivity
Increased Timeliness
Data Captured as Reported
Preservation of Statistical Distributions
(Univariate and Multivariate)

**UNE REVUE DE L'ÉTAT ACTUEL DE LA TECHNIQUE DE LA VÉRIFICATION ET DE L'IMPUTATION AUTOMATISÉES DE DONNÉES,** Mark Pierzchala, la Division de la recherche et des applications, le Service National des Statistiques Agricoles, le Ministère de l'Agriculture des États-Unis, Washington, D.C. 20250-2000, septembre 1988. Rapport No. SRB-TRS-88-10.

## RÉSUMÉ

Ce papier étudie des approches générales de l'automatisation de la vérification et de l'imputation de données et résume le progrès actuel de chacune des approches. L'état actuel de la technique de quatre organisations, Statistique Canada, le Bureau du Recensement des États-Unis, le Bureau Central des Statistiques des Pays-Bas, et le Service National des Statistiques Agricoles est revu.

## SOFTWARE AND HARDWARE

The body of the paper was composed and formatted with the UNIX$^{TM}$ troff text formatter on a SUN Microsystems$^{TM}$ 3/50 workstation The cover and this page were composed and formatted with the Frame Maker$^{TM}$ publishing and authoring package. A Texas Instruments OmniLaser$^{TM}$ 2115 laser printer using the PostScript$^{TM}$ markup language was used to print the report.

Credits and Trademarks

SAS is a registered trademark of SAS Institute Inc.
Sun Microsystems is a trademark of Sun Microsystems, Inc.
UNIX is a trademark of AT&T Bell Laboratories.
Frame Maker is a trademark of Frame Technology Corporation.
PostScript is a registered trademark of Adobe Systems Incorporated.
OmniLaser is a trademark of Texas Instruments Incorporated.

**A REVIEW OF THE STATE OF THE ART IN AUTOMATED DATA EDITING AND IMPUTATION**, by Mark Pierzchala, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250-2000, September 1988. Staff Report No. SRB-TRS-88-10.

## ABSTRACT

This paper explores some general approaches to the automation of data editing and imputation and summarizes the progress made to date in each of the approaches. The state of the art in four institutions, Statistics Canada, U.S. Bureau of the Census, Netherlands Central Bureau of Statistics, and National Agricultural Statistics Service (NASS) is reviewed.

## KEYWORDS

Editing, automated editing, statistical editing, macro-editing, productivity in survey processing, and imputation.

This paper was prepared for limited distribution to the research community outside the U.S. Department of Agriculture.

## ACKNOWLEDGEMENTS

The following people contributed to this report in the following ways:

A. Resource people from outside the National Agricultural Statistics Service (NASS).

Dr. Jelke Bethlehem of the Netherlands Central Bureau of Statistics, Dr. Brian Greenberg of the United States Bureau of the Census, John G. Kovar of Statistics Canada, and Leopold Granquist of Statistics Sweden.

B. Resource people from NASS.

Mike Steiner, Duffy Barr, Pat Miles, Dania Ferguson, Dale Atkinson, John Lange, Jim Brewster, Robert Tortora, Brian Carney, Jerry Fisher, Fred DeRoo, Jim Williams, George Hill, Tom Gregory, Toni Tremblay, Cathy Mazur, and Jim Stepanich.

C. NASS reviewers of the drafts of this report.

Brian Carney, Mary Ann Ciufini, Mark Harris, Marty Holko, Lesley Cross, Dania Ferguson, Duffy Barr, George Hanuschak, and Dale Atkinson.

D. Library work and photocopying of articles.

Carol Harris and Jennifer Kotch.

# Table of Contents

## Summary

The state of the art in automated data editing and imputation is explored in four statistical institutions: Statistics Canada, U.S. Bureau of the Census, Netherlands Central Bureau of Statistics, and the National Agricultural Statistics Service (NASS). Automation is proceeding in two major ways. On one hand, the computer is being developed to provide more powerful assistance to the subject matter specialist who is still at the center of the process. On the other hand, the computer is being developed as a tool to largely supplant the specialist as a data editor. The theory that might make the latter possible is not fully developed, however. These differing approaches are not mutually exclusive. It is possible to program the computer to perform corrections and imputations in certain well defined situations and to refer other, more difficult, situations to the specialist who would then handle the problems with the more powerful editing tools. Systems are being developed or are in use at Statistics Canada, U.S. Bureau of the Census, and the Netherlands Central Bureau of Statistics that either greatly modify the role of the specialist or provide more powerful tools with which to perform the editing function. The four institutions have varying approaches reflecting fundamentally different priorities of the organizations, such as efficiency, user satisfaction, statistical rationale, and defensibility, as well as the need to expand and modify a system to cope with more demanding survey requirements.

Each organization surveyed has its own list of editing system features that it would like to see developed. The choice of features must reflect not only the milieu in which the organization functions but also the way in which it would like to operate in the future. For example, the implementation of Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) threaten to render parts of some editing systems redundant if the organization should choose to collect its data by these means. Any one of several approaches to the problem of automation of editing and imputation systems is feasible.

Research in this area should be directed along the lines set by management, taking into consideration the underlying statistical rationale, the computers to be used, the mix of data collection technologies to be implemented, and the roles that the subject matter specialist, the clerk, the data entry personnel, and the computer are to play in the future as regards the editing function.

## I. Introduction.

### A. Background.

The editing process in NASS is cyclic, batch oriented, and involves professional review of computer edit printouts. The process involves hand coding and correction of items in an initial edit and entry process followed by redundant hand correction and data entry of fields that violate edits. The system requires a certain amount of paper shuffling as questionnaires are filed, pulled, and refiled as many times as the corresponding records fail the edits. In some surveys, such as the June Agricultural Survey, imputation is done by the commodity specialist not only for item nonresponse but also for total nonresponse and in either case with unknown effects on the distributions of the data. In data editing the emphasis is on within record (within-questionnaire) data validation leaving between-record analysis to post-edit analysis packages. Relatively few corrections are carried out by the computer; almost all must be made by specialists again with unknown effects on univariate and multivariate distributions. Edit failures are listed on printouts with no indication as to which of the fields is most likely to need correction. The effect of the edits, taken either individually or in combination, on the quality of the data is unknown. From a human perspective the process can be tedious as the individual must work through hundreds of pages of computer printouts, not knowing the success or failure of corrections and imputations until the next editing cycle. The desire to eliminate or reduce these problems and also to broaden the agency's perspective on the editing process is the impetus for this study.

### B. The study.

The comparative study concerns four survey institutions, NASS, Statistics Canada, the Netherlands Central Bureau of Statistics, and the U.S. Bureau of the Census. These institutions operate in different environments and thus have taken different approaches to reducing editing problems. The environment of each organization includes the survey environment and a historical environment. The former includes the types of surveys conducted, the tightness of survey deadlines, and the ways in which the populations are multivariately distributed. The latter concerns the history of how things have been done in the organization in the past.

## II. Terms of Reference.

As this is a comparative study of the implementation of generalized or multipurpose editing systems involving four organizations in three different countries, terms of reference are needed.

### A. Definition of the term editing.

The definition of the term editing varies. For example, editing may be considered either as a validating procedure or as a statistical procedure. Both procedures aim to reduce errors in data sets but each has its strengths and weaknesses. Additionally, editing can be done at the record level or at some level of aggregation of individual records.

#### 1. Editing as a validating procedure.

As a validating procedure, editing is a within-record action with the emphasis on detecting inconsistencies, impossibilities, and suspicious situations and correcting them. Examples of validation include: checking to see if the sum of parts adds up to the total, checking that the number of harvested acres is less than or equal to that of planted acres, and checking if a ratio falls within certain bounds as set by a subject matter specialist based on expert knowledge in the field. Validation procedures may also be thought of as being established without reference to collected data.

#### 2. Statistical editing.

As a statistical procedure, checks are based on a statistical analysis of respondent data (Greenberg and Surdi, 1984). A statistical edit usually follows validation in an editing system. It may refer to a between-record checking of current survey data or to a time series procedure using historical data of one firm. As a between-record check the emphasis is on detecting outliers of either univariate or multivariate distributions. One manifestation of between-record checking would be edit limits generated from distributions of a subset of current records. The most usual subset would be the first n records that enter the system. The error limits then would be applied to all records, the n records being run through the system again.

As a time series check the aim is to customize edit limits for each firm, based on that firm's historical data as fitted to time series models. This approach has been used in the Energy Information Administration of the U.S. Department of Energy using spectral analysis (Dinh, 1987). Cathy Mazur of NASS is also investigating the use of a time series approach to edit daily livestock slaughter data. The use of historical time series to check data may be one way in which to detect inliers, that is, data which should greatly deviate from the mean but do not.

#### 3. Macro-editing.

These are edits which are run on aggregations of data, perhaps at some summary level or in some economic cell. Leopold Granquist of Statistics Sweden is developing some of these ideas (Granquist, 1987a, 1987b). They have also been mentioned in Statistics Canada as a future research topic (Sande, 1988). The U.S. Bureau of the Census does macro-editing though the process is not at a high level of automation. The aim of macro editing is to edit data to find inconsistencies at the publishing level. It should be possible to trace the inconsistencies at the aggregate level to the individual records involved. Macro-editing focuses on those records in which corrections will have an impact at the particular aggregate level.

### B. Priorities of development efforts.

Examples of priorities in improving the editing process are: rationalizing the process, streamlining the process, and expanding the system's capabilities in handling expanded survey requirements. Rationalization refers to, among other things, statistical defensibility, maintaining univariate and multivariate distributions, and the consistent handling of errors and missing data. Streamlining focuses on performing tasks more efficiently with more powerful tools. Expansion of capabilities means larger edit programs, more flexibility at the local level, a larger number of variables, retention of the data as it is first keyed, more kinds of edit functions, and a system which does not preclude the addition of any features.

Organizations with differing priorities will allocate research and development resources differently. For example, rationalization of the edit process requires the development of theory and algorithms, whereas streamlining requires the acquisition of new hardware and new systems development.

C. Manner of making corrections and imputations; the roles of people and machines.

The manner in which corrections are made can be a very contentious issue. Implementation of new technology will at least have the effect of modifying the way in which editing tasks are done. This includes tasks performed by the subject matter specialist (in NASS the agricultural statistician), the clerk, and the data entry operator. In the most extreme manifestation of automation, certain parts of the jobs of these people could be eliminated. The resolution of this issue may be as much a personnel management question as a statistical one. There are several ways in which corrections could be made. Some examples:

- The subject matter specialist takes action on edit failures in a cyclic process using paper printouts in a batch processing system.
- The subject matter specialist takes action on edit failures in an interactive computer session.
- The computer takes action on some edit failures without review. The more difficult cases are referred to the subject matter specialist to contact the respondent or otherwise deal with the matter.
- The data entry operator corrects some types of errors at time of data entry leaving other errors for the computer or the specialist to handle.

The size of the survey, the time frame of the survey, and the resources that the survey organization is willing to commit to editing will all determine which mix of computer actions and personal actions is possible (see G, further). Also it is unlikely in economic surveys that any generalized system will be able to handle all records.

The choices made concerning the roles of people and machines will affect the acquisition of hardware and software. A specialist correcting errors interactively will require some type of terminal or personal computer to do so. If only the specialist is to make corrections then research should focus on giving the specialist more powerful tools. If the editing function of the computer is to be increased (for whatever reason), or if greater defensibility is desired, then research should be directed towards algorithms and theory.

D. The future role of CATI and CAPI.

Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) are technologies which can perform record validation at the time of data collection. If an organization will be collecting data primarily through these new technologies then it may be redundant to commit large resources to the validation part of an edit program. If on the other hand the organization must collect data through the mail, (as must the U.S. Bureau of the Census), or otherwise continue to use paper questionnaires, then further development of the validation programs is probably justified. One consideration is the timing of the implementation of CATI and CAPI. If implementation is 10 years away then it would be more justifiable to develop a new validation system than if the implementation of CATI and CAPI were only two years away. Another consideration is how much of the validation program will be transferred to CATI and CAPI. In NASS, CATI data are run through an editing program as not all edits are contained within CATI.

E. Location and dispersion of the work.

One of the organizations surveyed, the Netherlands Central Bureau of Statistics, is located in one building and therefore dissemination of new technology is aided by the proximity of resource personnel. The other organizations have many locations. The issue here is whether the same tasks are carried out in many locations and if different tasks are carried out in different locations. If the former is true, there is a problem of support, training, commitment, and consistency. In this case the organization may need simple-to-use systems as expertise is shared by telephone or word of mouth. If different tasks are carried out in different places then there is a problem of coordination between parts of the system. For example, the U. S. Bureau of the Census's data are key-entered in Jeffersonville, Indiana while editing is carried out in Suitland, Maryland. In this case the separation of functions enforces a division of labor that might preclude the implementation of some systems. In other words greater resources may have to be committed to making the interface between many users and the system more understandable. Hardware and training costs may be more expensive in organizations with many locations.

F. Specialization of survey processing tasks.

Specialization impacts the processing of survey data in that all of the specialized editing tasks, both before and during the survey, must be coordinated in an overall editing system. An example of specialization is one person creating an editing program and another using it to process survey data. In a modern survey organization, survey processing may be divided into tasks performed by tens or hundreds of people. The greater the amount of specialization the more the system will have to be constructed in modules embedded in an overall system that will coordinate the work of many different people. One effect of automation may be to improve productivity enough to allow fewer people to handle more tasks. For example, survey processing from questionnaire design to writing and testing of edits may be handled by one small group of people.

G. Time frames, sample sizes, and editing resources of surveys.

The size of the survey, the time frame of the survey, and the resources that the survey organization is willing to commit to editing will all determine which mix of computer actions and personal actions is possible. For example, a great number of survey schedules to be processed in a limited time may preclude personal action on each record. As another example, an organization with tight deadlines may not be able to let specialists enter data and correct it at the same time, as the speed that comes with specialization is required. On the other hand, an organization with declining staff numbers and tightening deadlines may be forced to adopt heretofore unneeded technologies. It may have to improve the productivity of the editing personnel in their handling of each record, or it may have to allow the computer to handle more of the routine errors without review, referring only difficult cases to the specialist.

H. Statistics to be derived and analyses to be conducted vis-a-vis the effect of editing and imputation on distributions of the data.

Different types of imputations have different effects on the marginal and joint distributions of the data. For example, in item nonresponse one possibility is to impute the average of the item from good records into records with that item missing. Another possibility is to impute values of the item from good records into the incomplete records (hot-deck imputation). In the former case the distribution of the item will change, becoming more peaked at the average value. In the latter case the distribution will not be changed as much. Both methods will give the same estimated average (at least in the limit) but the first method will understate the magnitude of the standard error. This is an issue of whether or not distributions must be maintained. Some statistics are not sensitive to altered distributions, for example, averages, totals and proportions (although their standard errors are). Other statistics, such as measures of dispersion or multivariate analyses, are sensitive to altered distributions. If distributions are to be maintained then it may be better to leave the bulk of the editing, correction and imputation to the computer. That is, some imputation procedures, including hand imputation, may not be suitable for some statistics and analyses.

Any imputation scheme rests on (sometimes implied) assumptions about the distributions of data for the nonrespondents compared to that of respondents. These assumptions should be stated and critically examined as to their validity.

I. The degree of variability of population attributes affects imputation.

As observed by Duffy Barr, (1988, personal communication), it may be easier to impute for missing values in a gas station survey than for agricultural items because gas station prices may be much less variable than items in an agricultural survey.

J. Planned uses of the data and the availability of the data.

This point derives from point H. If record level data must be released to other organizations then the collecting organization is obliged to leave the multivariate distributions as intact as possible as not all future uses of the data are known in advance. Once the data are outside of the organization there is no telling how they will be utilized, that is, whether multivariate analyses will be carried out or statistics will be generated that require distributions to remain intact. For example, the Economic Research Service of the U. S. Department of Agriculture obtains record level data through NASS from the Farm Costs and Returns Survey (FCRS) survey. As NASS does not know every future use of the data the editing procedure should maintain the multivariate distributions. (NASS does not currently impute in the FCRS.) At a minimum, if a large amount of imputation is done, imputations should be flagged and a description of imputation procedures included with the data.

K. Types of surveys.

The types of surveys being processed, such as economic, social, or production surveys, will reflect on the complexity of the editing programs as regards such items as routing, type of data (categorical or continuous), the degree of inter-relation between the fields as expressed through edits, the reliability of the edits, and the type and complexity of the edits themselves. These attributes affect the degree to which the editing process can be automated.

L. Previous survey experience.

Organizations have different experiences regarding the degree of noncooperation, item nonresponse, partial nonresponse, and the frequency of errors in collected data. The relative amounts of resources spent on each survey step will be different. As a result, different organizations will have different perspectives and priorities in the development of new systems. Systems which may be justified in some organizations on the basis of the tradeoff between cost and data quality may not be justified in others. Also at stake is the validity of the editing and imputation procedures and their defensibility. An organization with high rates of nonresponse may have greater difficulty in establishing a workable and defensible system in which the computer makes the bulk of corrections. For example, it would be harder to implement hot-deck imputation in a survey with 30% nonresponse than in one with 5% nonresponse because donor records may not be available for all recipient records in the former case.

M. Hardware.

Computer intensive procedures, or interactive handling of the data, may be too expensive on leased mainframes if the organization is charged according to resources used. This may result in greater reliance on hand processing or a situation in which some features are not even considered due to the cost. On the other hand, microcomputers may not have the capacity to handle some editing functions or they may not have access to historical information.

N. Software environment and programming support.

The term software environment refers to whether or not the editing system will reside in a data base environment. If editing is to be carried out in a data base environment the question is whether the data base will be shared out between locations or be centralized. If the latter case is true, then at least part of the edit will have to be carried out on the computer carrying the data base. Programming support refers to the amount of support available to customize editing programs for each survey, to modify a generalized program for each survey, or to support a program in different environments (editing on microcomputers as opposed to a mainframe for example) as well as maintaining existing programs.

O. Purposes and costs of editing.

See Granquist (1988a), and Pullum, Harpham, and Ozsever (1986), for good discussions on the purposes of editing systems. These papers address the tradeoffs between improvements in data quality and costs of editing. In the former paper, Granquist estimates that editing takes from 20 to 40 percent of survey budgets in periodic surveys in Statistics Sweden and wonders if the benefits are worth the expenditures. In the latter paper, which discusses the editing of the World Fertility Survey, it is reported that estimates derived from raw data tapes in 6 countries were essentially the same as those derived from edited data tapes. In other words, the machine editing had no appreciable effect on the analysis other than to delay the production of statistics by one year. The authors of these two papers do not question the basic necessity of editing, but consider that some editing resources could be allocated to other areas to improve data quality or that editing could be done in better ways.

Pullum, et. al., cite 5 reasons why the World Fertility Survey did implement stringent editing policies. These are cited as general beliefs as to why editing is done.

1) To produce a gain in the yield of the fieldwork, that is, to minimize the number of responses excluded from analysis.

2) To improve the validity of the findings, that is, to remove systematic errors that may lead to bias.

3) To improve the correspondence between the structure of the questionnaire and that of the responses, the net effect being the easing of further tabulation and analysis.

4) Users have more confidence in data which are internally consistent because such consistency reflects on the entire process of data collection and preparation.

5) The perception that editing is a hallmark of professional survey research.

In his review of Pullum's et. al. World Fertility Survey paper, Granquist (1988a) maintains that only reasons 3 and 4 are real benefits of editing in the way it was carried out here, that is, through a Generalized Edit system. Granquist (1984a) describes the following purposes of editing:

1) To give detailed information about the quality of the survey.

2) To provide basic data for the improvement of the survey.

3) To tidy up the data.

Granquist further believes that Generalized Edit systems usually apply too many checks, that editing systems do not essentially improve data quality, and that editing systems can give a false impression of data quality.

P. Productivity and costs of editing.

Another way in which to consider the effects of editing costs on the manner in which automation is affected is to plot the rapidly declining costs of computing against labor costs that are either constant or climbing. Kinds of automation considered too expensive 5 to 10 years ago, (for example computationally intensive programs or interactive handling of corrections), may be less expensive now, or in the future, than remaining with a labor intensive status quo.

## III. The Fellegi and Holt school of edit automation and imputation.

The literature emanating from this school of thought is concerned primarily with the stage of editing known as data validation. This school is characterized by its foundation in set theory, borrows heavily from techniques in Operations Research, Statistics, and Computer Science (Sande, 1979), and is guided by certain principles: that each record satisfy all edits, that correction be accomplished by as few changes as possible, that editing and imputation both be part of the same process, and that any imputation procedure retain the structure of the data. Automation of editing and imputation are required because some of the above desired principles are beyond the ability of the human editors. Automation may not be any cheaper than the more labor intensive methods, but the computer can apply all edits quickly and consistently (Fellegi and Holt, 1976). Emphasis is placed on the rationalization and the defensibility of the editing process. Statistics Canada (where Fellegi is Chief Statistician of Canada) and the U.S. Bureau of the Census are implementing this approach.

### A. Changing as few fields as possible in correction of data.

In their 1976 paper Fellegi and Holt outline a set theoretic approach which, if applied to categorical data or to linear edits of continuous data, would lead to the identification of a *minimal set* of fields that need to be corrected in order to clean the record. The corrections, if made according to the editing rules, would guarantee that the whole record would pass all edits. This result can be extended somewhat since some nonlinear edits can be rendered into a linear form (e.g., one can render a ratio edit into two linear inequalities), (Giles and Patrick, 1986). This approach requires that a *complete set* of edits be generated from the *explicit edits* written by the subject matter specialist. The idea is that there are *implied edits* which can be generated by logical implication from the explicit edits. For example, if $1 < a/b < 2$ and $2 < b/c < 4$, are explicit edits, then $2 < a/c < 8$ is an implied edit obtained algebraically from the explicit edits. The complete set of edits is the union of the explicit edits and the implicit edits. Once the complete set of edits is determined, a minimal set of fields can be determined for every possible set of edit failures. The determination of a minimal set of fields is called *error localization*. There are still some cases involving nonlinear edits in which it is impossible in general to find minimal sets because the complete set of implied edits cannot be found. The minimal set does exist however (Greenberg, personal communication).

### B. Editing and imputation as the same process.

In the ideal Fellegi and Holt automated editing process, imputation constraints when taken together are called a *feasible region* and are derived from the set of complete edits. Corrections or imputations falling within this feasible region are guaranteed to pass the edits. Fellegi and Holt show that for categorical data or for continuous data under linear edits, either there is a feasible region or some edits are in conflict. In practice there are some types of nonlinear edits which are not amenable to the determination of a feasible region. In such cases, the imputations can be run through the edits again to ensure that all imputations conform to the edits. In any case, it is a precept of this school of thought that all corrections and imputations will pass all edits, although this may not be strictly adhered to in practice.

### C. Retaining the structure of the data.

One of the major objectives of the Fellegi and Holt school is to retain the structure of the data. This means that univariate and multivariate distributions of survey data reflect as nearly as possible the distributions in the population. Statistics Canada is doing this already by the use of hot-deck imputation. The U.S. Bureau of the Census uses hot-decking for agricultural surveys, for some demographic surveys, and the decennial censuses. Hot-deck imputation seeks to find a record similar to that of the incomplete record in the current set of survey records and to impute the missing variables from the complete record to the incomplete record. This can be done one variable at a time, the aim being to preserve the univariate distributions, or all variables at once, the aim then being to preserve the multivariate distributions. Retaining structure is important if there is to be multivariate analysis, if not all uses of the data are known in advance (e.g., it is not known who will have access to it), or if statistics which depend on the distribution (e.g., quantiles) are to be calculated.

D. Implementation.

Implementation of the approach of Fellegi and Holt has proved to be a challenge for nonlinear edits and continuous data. Checking the consistency of explicit edits, the generation of implied edits, and the determination of an acceptance region require Operations Research (OR) methods (Sande, 1979). In hot-deck imputation, procedures from OR are needed to minimize the search for donor records. For a minimal set of fields, a best corresponding set of matching variables must be determined. An exact match between a candidate and donor record may not be possible in the continuous case, thus a *distance function* is used to define similarity. Some numerical imputations are not guaranteed to pass edits as are categorical imputations, thus redonation may be necessary, (Giles and Patrick, 1986). A donor record may have characteristics similar to those in the candidate record but the operation may have a different size, thus scaling is required. Continuous edit checks that are linear are amenable to known Operations Research procedures, whereas non-linear edits (such as conditional checks) are not. In the words of Brian Greenberg, U.S. Bureau of Census, "To the extent that the methods developed by Fellegi and Holt for categorical data and by Sande for continuous data under linear constraints are employed in these (editing and imputation) routines, a high level of rigor will be introduced into this system. Any success in developing procedures to systematically address the comparable criterion for conditional numerical, conditional categorical, or mixed edits will be a fine methodological advance.", (Greenberg, 1987). In some records more than one minimal set of fields may exist. If so, some procedure is needed to determine which set should be corrected. One method is to assign weights to reflect the relative reliability (in the opinion of the subject matter expert) of each field. Thus if multiple minimal fields are found, the least reliable set of fields is updated.

E. Manifestations.

Both Statistics Canada and the U.S. Bureau of the Census have implemented this editing philosophy to a certain degree. Neither system fully automates the editing process. Since the systems are not fully automated some records are reviewed by the specialist. These records are either too difficult to be dealt with by the machine, or are referred to the specialist according to certain pre-determined criteria such as size of firm.

1. United States Bureau of the Census

In the U.S. Bureau of the Census, present implementation of the philosophy of Fellegi and Holt resides in a subsystem called the SPEER system (Structured Program for Economic Editing and Referrals). SPEER handles continuous data under ratio edits, and has six main components: Edit Generation, Edit Analysis, Edit Checking, Error Localization, Imputation, and Diagnostics (Greenberg, 1987). From survey to survey, it is the Imputation module which requires great change. In the Census experience, the Edit Generation, Edit Checking, and the Error Localization modules remain virtually unchanged (Greenberg, 1987). SPEER resides within a larger editing system. This reflects the fact that there are a number of tasks (such as SIC code assignment, GEO assignment) that SPEER is not designed to perform. Additivity checks are also handled in SPEER. Other types of checks can be handled before or after SPEER is invoked or in special satellite routines within SPEER itself. Changes made outside SPEER at times cause violations of edits within SPEER. Census has adapted the approach of Fellegi and Holt as far as possible to increase the degree of automation. Greenberg has extended the approach of Fellegi and Holt into the realm of ratio edits. However, this is done by considering the ratio edits as a set unto itself, doing what is possible within that set and sending the result to the broader system for further processing.

Imputation modules are applied one field at a time. These imputation modules consist of a series of rules that are utilized in a sequence until one of the rules generates a value that will satisfy the edits.

These modules are easy to create and can easily be revised to accommodate new understandings about the data (Greenberg and Surdi, 1984). When the imputation modules fail, the record is output to the specialist. In the interactive process the statistician is presented with a list of fields in error and with ranges within which the value of each field must fall. The specialist enters a value for one field at a time, and each time the computer recalculates the ranges for the remaining fields to be changed. The result of the determination of a minimal set of fields and of the calculation of feasible regions is that the cyclic process of error printouts, error correction, and more error printouts is diminished or eliminated.

Brian Greenberg, Principal Researcher in the Statistical Research Division, views the editing process in two stages: (1) automated batch runs for all records, and (2) manual review for specially targeted records. It is not desired to remove the analyst review component of the process. The aim is to provide the analyst with more information on the review document coming out of the batch run to assist in review tasks. The analyst review tasks should be done in an interactive mode working with the computer. The objectives of the analysts' job would not fundamentally change though the mechanics and logistics might.

The Bureau of the Census has processed one large survey, the Census of Construction Industries, with their system which included the SPEER subsystem. This was done on a mainframe because of the number of records involved (several hundred thousand). For two surveys in the Economic Surveys Division, the 1987 Enterprise Summary Report, and the 1987 Auxiliary Establishment Report, the Bureau of the Census is considering employing a combination of mainframe and microcomputers. The mainframe would be used for batch processing and automated imputation and would refer difficult cases to the specialist to handle on a microcomputer. The number of cases handled on the microcomputer would depend on the referral criteria which in turn would depend on how much the editing and imputation algorithms on the mainframe were trusted. Referral criteria can include the magnitude of changes made by SPEER or the size of the firm involved. In addition, the Industry Division has developed computer programs based on the SPEER methodology, and they have been used for the 1986 Annual Survey of Manufactures and the 1987 Census of Manufactures. The Agricultural Division of the Bureau of the Census is considering using the system for Agriculture Economic and Land Ownership Survey for which data collection will start in 1989.

## 2. Statistics Canada

In the Statistics Canada survey processing system for economic surveys, two modules of this system will handle distinct parts of the editing process. The first module is the Data Collection and Capture (DC2) module, the second is the Generalized Edit and Imputation System (GEIS). DC2 is in the prototype stage whereas the GEIS has recently been completed and documented. Different modules are being created for different parts of the edit process because in Canada the response unit may be different from the statistical unit. For example, a firm might provide data for two factories on one questionnaire. In this case the responding unit would be the firm and the statistical units would be the factories. DC2 would customize the forms to the respondent, do some basic editing at that level, and flag questionnaires for follow-up. All document control (status codes, etc.), a substantial amount of correction and all necessary follow-up is done in this preliminary edit. GEIS is meant to handle data at the statistical unit level, that is after the data have been processed by DC2. Only unresolved cases or cases of minor impact are passed to the Generalized Edit and Imputation System as a last resort, at which point an effort is made to resolve all problems by imputation (Kovar, 1988).

For now, GEIS will handle data which have not been processed by DC2. In this instance it is expected that the amount of hand editing will be held to a minimum. Hand checking will be confined primarily to making sure that numeric data are entered in numeric fields and the like, and that control data on the first page is correct. GEIS has not yet been used in a production mode as the developers are still looking for clients. It is in the GEIS system that the philosophy and techniques of the Fellegi and Holt school of editing are currently in place.

Currently, GEIS handles only those edits that are linear and data that is positive but within these constraints most edits and data can be handled. Many nonlinear edits can be recast in a linear form and negative data values can be given as a difference of two positive numbers (e.g., profits = income - outgo). These constraints are to be relaxed in the future. GEIS is embedded in the relational data base management system ORACLE, which facilitates the organization and the handling of data (Kovar, 1988). This aids in monitoring the edit and imputation process.

GEIS as an editing program consists of four main parts: specification of edits, analysis of edits, application of edits, and outlier detection (Kovar, 1988). The specification of edits is done by a subject matter specialist working together with a methodologist. Specification is typically done on a microcomputer. The system performs a syntax check and also checks that variables employed in the edits have been specified in the questionnaire.

Further checking of the edits occurs in the analysis of the edits. This can be accomplished because the edits are linear and the data is positive. The edit analysis checks the consistency of the edits. The analysis also checks for redundant edits that do not further restrict the feasible region of the data values. The system then generates the acceptable ranges for all variables, the extreme points of the feasible region, and the set of implied edits (Kovar, 1988 and Sande, 1979). This part of the system aids the analyst in determining if the edits are meaningful. It also helps verify whether all edits were entered correctly.

In the application of the edits, an error localization procedure is invoked to determine the minimal number of fields to be corrected. Alternatively the same procedure can be used to find the minimally weighted set of fields to be corrected. This latter alternative utilizes additional information on the reliability of the fields as judged by the specialist. If an edit failure can be cleared up by the imputation of only one value for a variable then that value is imputed, that is, the error localization procedure handles deterministic cases. Uncorrected or unimputed records are passed onto the imputation procedure. In the imputation procedure, two general methods are available, donor imputation and other imputation procedures. Donor imputation is implemented by hot-deck imputation. This is meant to be the primary method of imputation. Hot-deck imputation is preferred because it retains the structure of the data. Other imputation procedures include imputation of historic values (which can be trend adjusted), imputation of means, and ratio and regression estimators. These methods are backup methods used when the hot-deck procedure fails. They will not preserve the structure of the data as effectively as the hot-deck method. GEIS also has a facility which allows a choice of imputation methods by field.

Outlier detection is in the form of a statistical edit that operates on all records at once and cannot be applied at the same time as the other edits. The module can serve two distinct purposes: to determine the edit bounds, or to identify outlying values which can be flagged for imputation or for other considerations in subsequent modules (Kovar, 1988).

The mathematical procedures needed for optimization and search are being written in C. GEIS is being produced in several releases, with new features available in each release. Methodologists do not feel that they have all the answers yet and would like to provide a wide selection of alternatives for planning the edit and imputation. However, offering maximum flexibility and maximum satisfaction results in a system which does not hang together very well. A more unifying theoretical basis is needed (Sande, 1988). The GEIS system is not as easy to use as desired and a great deal of intervention is still necessary. Statistics Canada expects system implementation to require a substantial amount of time.

### F. Future Research

The following list of topics must be researched in order to fully implement the goals of the Fellegi and Holt school. This list was compiled from the literature and from personal communication with people from the U.S. Bureau of the Census and Statistics Canada.

- Non-linear edits (including conditional edits) in order to generate the set of implied edits and hence the complete set of edits and to generate a minimal set for non-linear edits.
- Negative values of variables.
- Implicitly defined constants.
- What to do with multiple solutions (multiple minimal sets) in error localization.
- Variance estimation in the presence of imputed data (Sande, 1988).
- Zero values versus missing values, that is, does a blank on a questionnaire represent a zero or was the item skipped.
- More intelligent or expert systems.
- Automated macro-edits carried out on tabulations of statistics rather than micro data with comparison between cells, between variables with historic data, and with other data sources, in order to avoid embarrassing data discontinuities, identify design and estimation problems, and lead to the formulation of improved micro-edits (Sande, 1988).
- Determination of which imputation option to use in which context.
- How to edit and impute when the data from a reporting unit includes data at the location level, establishment level, and all Canada level (Sande, 1988).
- What should be done when this year's imputation must be based on last year's imputation.
- Mixed edits (Giles, 1987).
- The blend in a multipurpose system between general routines and survey-specific procedures (Greenberg, pc).

G. Problems with this approach.

Following is a list of problems gleaned from the literature and from personal communication.

- The theory is not worked out for all cases. It can be implemented for categorical data, continuous data under linear edits, and ratio edits but not for some kinds of nonlinear edits such as conditional or mixed edits. (Though it may be possible to render some nonlinear edits into a linear form.)
- Programming complexity (Fellegi and Holt, 1976).
- Distinguishing valid zeroes from missing data (Kovar). This is a problem for Statistics Canada because some of their survey data are obtained from computerized tax files with no recourse to the tax form in determining what a blank means.
- The choice of the appropriate imputation method in various contexts.
- The determination of which minimal set to correct when multiple minimal sets are found.
- The subject matter specialists may not be happy with going from a system in which specialist action is required to a system in which the computer takes most of the actions.

H. The desirable features of the system.

Following is a collection of features from several papers which were either explicitly stated or implied. Features are not prioritized. These features may or may not be implemented at this time.

1. Methodological features

- There should be an orderly framework and philosophy for the development of edit and imputation procedures.

- Each record should satisfy edits by changing the fewest possible fields. ("Keeping the maximum amount of data unchanged."), (Fellegi and Holt, 1976).
- Imputation should be based as much as possible on the good fields of any record to be imputed (Fellegi and Holt, 1976).
- The frequency structure of the data file should be maintained for joint frequencies as well as for marginal frequencies.
- Imputation rules should be derived from corresponding edit rules without explicit specification.
- Imputations should not violate edits.
- The system should supply methodologically sound modules to be assembled by the user (Kovar).
- Defensibility of methods should be a priority (may constrain the user somewhat), (Kovar).
- When imputing for a deletion due to edit failures one should endeavor to utilize the reported value although it is incorrect. (Example, the respondent answers in pounds when tons were requested.) (Greenberg, 1982b).
- Feedback should be provided on the impact of the system on estimates.
- The system should detect univariate outliers (Kovar).

2. System oriented features.

- It should not be necessary to specify imputation rules (Fellegi and Holt, 1976).
- The edit program should determine a minimal set of fields that needs to be changed in order to satisfy the edit.
- The edit program should logically deduce a set of implied edits.
- Each record should be edited only once.
- Edit procedures should be implemented in a generalized editing system as part of the automation of the editing process. Thus systems development need not be delayed by specific edit specifications. (However, it still may be desirable in some cases to introduce survey specific procedures. See the last item under letter F., Further Research).
- Program rigidity should be reduced. (Fellegi and Holt, 1976). That is, changes to the program should be easy to accomplish without making errors.
- The generalized editing system should be embedded in a relational data base management system (Sande, 1988). The data base environment should allow:
  a. easy monitoring of imputation process (Kovar).
  b. measurement of impact (of editing process) on estimates (Kovar).
  c. measurement of the effect of imputation on particular stages of the process (Kovar).
- The software should be portable between computers of varying types (Sande, 1988).
- The system should be modular, allowing the development of distinct stages of the editing and imputation process. It should also allow ease of updating and the ability to add new modules.
- Records as captured by the system should be kept in a file separate from those being edited and corrected. This allows evaluation of the technique and also allows one to go back to the respondent with data as the respondent gave it.

3. Subject matter specialist oriented features.

- The subject matter specialists should be an integral part of a team implementing automated editing and imputation procedures.

- Only edits should have to be specified in advance as imputation rules would be generated by the computer (Fellegi and Holt, 1976).
- The subject specialist should be able to inspect the set of implied edits derived from the explicitly specified edits in order to evaluate the explicitly specified edits.
- In any one survey, different sets of edit specifications (corresponding to different parts of the questionnaire) should be able to be created concurrently by two or more specialists. (The system would take care of any reconciliation.), (Fellegi and Holt, 1976).
- Respecification of edits and imputation rules should be done easily (Greenberg, 1984 and others).
- The specialist should be able to experiment with the explicit edits either before the survey is conducted or after a small subset of the records are in, thus gaining information on the explicit edits.
- The generalized edit and data correction system should allow respecification of edits without reprogramming (Fellegi and Holt, 1976).
- The edit program should be in modules so that the subject expert can easily enter or change explicit edits, and new versions of the system can easily be installed.
- The edit program should provide feedback on how the edits are affecting data quality.
- The system should be comprehensible to and readily modifiable by the users of the subsystem.
- The system should be flexible and satisfying to the user, (give the users what they want), (Kovar).
- The system should have a menu approach in which the user is presented with a choice of functions to apply to the data or subsets of the data (Sande, 1988).
- A sample of edit failures should be checked. Statistics should be collected on the frequency of edit failure, the frequency with which each field is involved in an edit failure, the frequency with which each edit is failed and the frequency with which each field is identified for change (Sande, 1988).

## IV. Streamlining and Integrating the Survey Process; the Blaise System from the Netherlands.

In this approach, as implemented by the Netherlands Central Bureau of Statistics, the automation of the editing process is but one part, albeit an important part, of the automation of the overall survey process. In this approach, no new theoretical tools are implemented to rationalize the editing process. The idea is to take the current cyclic batch process performed on mainframes and to put it on microcomputers, thus creating an interactive process. Because the survey process is now handled with essentially one integrated system of modules, the data need be specified only once. That is, the specialist does not have to write an editing program, as it is derived automatically from the specification of the questionnaire on the computer. In addition CATI and CAPI modules are generated automatically from the Blaise questionnaire. The emphasis is on streamlining current methods and on integrating most computerized survey functions.

The subject matter specialist plays two important parts, the specification of the questionnaire and the resolution of the data. In the resolution of the data, the specialist is either a data checker/corrector or a data typist/checker/corrector.

### A. Subject matter specialist as error checker/corrector.

The data are entered as normal, by very fast data entry personnel. The file is passed to the subject matter specialist who uses the microcomputer to correct errors one questionnaire at a time. After corrections are made, the record is re-edited on the spot and redisplayed with new error messages, if any. Thus the batch cycle is changed to an interactive micro cycle (micro having two meanings here, microcomputer, and each record cycling by itself through the editing program until it is correct). The questionnaires are available to the specialist for reference. The specialist sees the errors on the microcomputer screen along with error messages. The questions as they appeared on the questionnaire are not presented on the screen, rather mnemonic variables are displayed. The questions are available from a help screen if needed. The record can be routed to a holding file of difficult questionnaires, if necessary. The system does not generate a minimal set of fields to be corrected as in the Fellegi and Holt school. It does display the number of times each field is involved in an edit failure. The premise is that the fields flagged the most often should be the first corrected as they are the ones most likely to be causing the problems.

### B. Subject matter specialist as data typist/checker/corrector.

The data are entered by the specialist, who is not as fast as the regular data entry personnel. However, the record is edited as it is entered. The specialist entering the data is also qualified to correct errors. Thus data entry and reconciliation are combined. The extra time in data entry is offset by the one time handling of the record. Codes can be entered interactively.

### C. The Blaise survey processing system.

The Netherlands Central Bureau of Statistics is currently writing an automated survey processing system called the Blaise system. This system is written in a structured language called Blaise which in turn is written in Turbo Pascal. (Blaise is the first name of the famous French mathematician Pascal.) The key element of the Blaise system is the Blaise questionnaire. This questionnaire is not a survey collection instrument but rather a specification of questions, routing instructions, and edit checks which is entered into a computer. Once the Blaise questionnaire is correctly specified, all other software instruments used to process a survey are automatically generated from it. This includes an intelligent data entry module, an interactive error correction module, survey instruments such as CATI, CAPI, and PAPI (Paper And Pencil Interviewing), and an interface with standard statistical packages such as SPSS. This approach has at least two major advantages: data need to be specified only once (eliminating redundant specification for data entry, edit program, and the like), and if there is a change in the questionnaire, the changes in the other modules follow automatically. The subject matter expert's task of correcting the data is not changed, only the tools used are, enabling the expert to do the job while engaged in an intelligent and interactive session with the computer. The work is concentrated in the department where the knowledge is, that is, the subject matter department. The computer does not present options but it can be programmed to carry out some kinds of imputations without human intervention.

D. Implementation.

In the Netherlands the system is currently being implemented in its first few versions. The first modules installed were the data entry and editing modules, and an interface with SPSS. Other modules are in the prototype stage (testing was being carried out at the beginning of 1988). These include the CATI and CAPI modules. The PAPI module for the generation of the paper questionnaires has not yet been fully developed. Blaise can produce an editable list of questions which may then be developed into a paper questionnaire.

In the Dutch implementation, the first step in the survey process is to design a Blaise questionnaire. This acts as a knowledge base in an artificial intelligence system. The Blaise system is the reasoning mechanism which utilizes the knowledge base in the Blaise questionnaire to produce other software products such as CATI or CAPI. Progressing from the Blaise questionnaire to a software product is not just a matter of straightforward translation. The questionnaire must be interpreted in various ways according to the product desired (Denteneer, et al., 1987).

The questionnaire is constructed of blocks which usually correspond to topics. The blocks are composed of functional paragraphs. A question paragraph, a route paragraph, and a check paragraph are almost always required although other kinds of paragraphs such as a variable paragraph and a type paragraph are available to ease the construction of complicated questionnaires. The block construction allows for easy updating of questionnaires from one survey to another or for easy standardization between different surveys. The paragraphs hold the information upon which all other products are based. As all other software products are generated automatically, the survey managers are spared the task of respecifying the data.

In the Dutch experience, benefits in the first implementation of the data editing and data entry programs included a 10% reduction of the workload in the subject matter department, a reduction of the total processing time of 3 to 4 weeks (it does not say from how long) and improved quality of data (it does not say how this was measured). It was reported that survey personnel liked working with the new system and that they liked seeing the results of their work immediately (Bethlehem, 1987a).

E. Future work.

The Netherlands Central Bureau of Statistics is working to implement the CATI and CAPI modules. It is their intention to make much greater use of the computer gathering devices in order to verify the data as it is being collected. Their CAPI machines are small Toshiba portable computers weighing about 5 or 6 pounds each. They were chosen for their weight and power usage. Dr. Bethlehem describes the use of CAPI and CATI in terms of the Deming philosophy of quality control - to make sure that quality is built into the product in the first place rather than relying on post collection software to edit in quality after the fact.

F. The desirable features of the system.

Following is a collection of features which were explicitly stated or implied in several papers. The features are not prioritized. These features may or may not be implemented at the present time.

- Forms should not need to be prepared for entry (Bethlehem, 1987a).

- The cyclic nature of editing should be removed (Bethlehem, 1987a).
- The work should be concentrated as much as possible in same department (Bethlehem, 1987a).
- The work should be done as much as possible on the same computer (Bethlehem, 1987a).
- There should be a reduction in time needed for editing (Bethlehem, 1987a).
- The system should be applicable to different surveys.
- Superfluous activities should be eliminated.
- Error checking should be an intelligent and interactive process carried out between the subject matter specialist and the computer.
- The structure of the questionnaire and the properties of the resulting data should be specified only once (Bethlehem, 1987a).
- Editing automation should be part of a larger automation process. (This is really the crux of the matter, that a Blaise questionnaire that contains total information about the questionnaire be constructed, and from which all products fall out, including edit programs, CATI and CAPI instruments, etc. The Blaise questionnaire is considered a knowledge base in an artificial intelligence context.)
- Data entry should have an interactive capability.
- An interface with statistical packages should be possible without the necessity of respecification of the data.
- The system should be user friendly (Bethlehem, 1987a).
- The system should be based on microcomputers such as IBM AT/XTs and compatibles (Bethlehem, 1987a).
- The updating of questionnaires from one survey to another should be easy (Bethlehem, 1987b).

G. Impetus of the project.

The Netherlands Central Bureau of Statistics implemented a data editing research project. The objective was to assemble management data on the editing process through the analysis of the processing of four selected surveys of various types and characteristics. For example, in the initial hand editing stage of survey processing, the steps in the editing process were listed and evaluated for their contributions to improved data quality. Editing activities were classified into three types: real improvements, preparation for data entry, and superfluous activities (such as writing a minus sign for missing data). In one survey, the relative share of the time spent on these activities was 23%, 18%, and 59% respectively. These measurements were made by inspection of the questionnaires after the surveys were completed. Other quantitative and qualitative aspects of survey processing were measured such as rates of edit failure, time needed to clean a file, as well as the ways in which personnel interacted with the computer systems. Some of the findings of the project (Bethlehem, 1987a):

- Different people from different departments were involved.
- Different computer systems were involved.
- Not all activities were aimed at quality improvement.
- Manual check of complex routing structures was difficult and time consuming.
- The editing process was cyclic
- Repeated specifications of the data were necessary. The term repeated specification refers to the practice of specifying variables, valid values for the variables, relationships between variables, and routes to be followed depending upon values of the variables in the survey stream. These items were specified for the questionnaire (on paper or as a CATI or CAPI instrument), and again in data entry software, in an editing and analysis program, and in a summary program. This problem was compounded if the various tasks were carried out on different computers using different software. In those cases significant resources were spent just transferring data from one location to the next. This last point led to the design of the Blaise system (Denteneer, et al., 1987).

## V. NASS Development Effort.

The theme underlying NASS's Survey Processing System (SPS) is one of integration. The term integration impacts the SPS in two major ways. Firstly, this term refers to one of the impetuses of the development of the SPS, that is, the integration of NASS surveys into one coherent sequence of surveys. This was originally done under the name of the Integrated Survey Program and is now known as the Quarterly Agricultural Survey Program. Secondly, the term refers to the integration of the distinct steps of the survey process under a unifying system. As such, the SPS has modules for data capture, data validation, and statistical editing although this latter module is not fully developed or utilized. In the future the SPS will also encompass modules for imputation, analysis, summary, and reports with further connections to a public use data base and a secure agency data base. A *specifications generator* is to be developed and will serve as the unifying feature of the SPS. It is envisioned that the specifications generator will output files for further processing into paper questionnaires, CATI and CAPI instruments, and an editing system. Integration will also serve to ensure the consistency of editing procedures across all surveys.

The implementation of the Integrated Survey Program served as an impetus to the development of the SPS because the previous edit and summary system could not handle the requirements of the new program. For example, there was a need to be able to process each section of the questionnaire differently as regards completion codes and refusals in order to summarize the sections independently. In addition, the old system could not handle the large number of variables demanded by the new survey system and it would not allow data to be compared between records.

Beyond the system limitations mentioned above, a number of new capabilities are desired for statistical purposes. The term editing was expanded to include statistical edits which involve cross-record comparisons at the time of the more traditional data validation (Vogel, et al., 1985). It was also desired that the effect of imputations and non-response be known at various levels of aggregation, that procedures be consistent across all surveys, and that NASS procedures be statistically defensible. Editing and the imputation of missing data are not considered as part of the same process in the sense of Fellegi and Holt. That is, the edits are not used to define a feasible region for imputation. However, nothing in the system prevents records with imputed values from being run through the edits once again.

NASS is probably one of the few agencies in the world to have programmed its editing software in the statistical language SAS®. The use of SAS for editing has sparked some interest in an international editing research group because of its portability, its wide use as an analysis tool, its flexibility, and its amenability to developments in statistical editing and in micro-macro combination edits (Atkinson, 1988b). These advantages also apply to NASS, that is, nothing is precluded, therefore options are maintained.

### A. Limitations of the old system (Generalized Edit System).

Following is a list of limitations gleaned from various NASS reports, conversations with NASS personnel, or noted from personal experience.

- An artificial limit to the number of parameter cards is often exceeded.
- Parameters are difficult to write.
- Error printouts are difficult to understand.
- Types and numbers of edit functions are limited.
- The system is not updated systematically.
- The system is not supported with training.
- The manual is written in undefined jargon and contains few examples.

- Comparisons between records are not allowed, that is, the statistical distributions of the data were not reviewed.
- The system only points out errors it does not correct them (Vogel, et al., 1985).
- The manual resolution of errors can vary from state to state (Vogel, et al., 1985).
- There is no built in method to evaluate the effect of editing (Vogel, et al., 1985).
- Cross record processing is not allowed.
- The system cannot handle the thousands of variables required by the new system of surveys.

B. Current and desired features of the Survey Processing System.

1. Broad aims.

- Procedures should be objective, repeatable, and statistically defensible (Vogel, et al., 1985).
- Procedures should be consistent across all surveys.

2. Impact of editing, contribution of nonresponse.

- The edit system should allow a review of the number of edit actions by type and by their individual effect on the final indication (Vogel, et al., 1985).
- The edit system should allow the contribution from nonresponse to be known (Vogel, et al., 1985).
- The Board should be able to monitor how data editing, imputation for nonresponse, and adjustment for outliers affect the estimates (Vogel, et al., 1985).
- The edit system should allow the commodity statistician to monitor the relationship between raw and edited data (Vogel, et al., 1985).
- The system should retain survey data as it is reported (Vogel, et al., 1985). This would allow some measurement of how the editing process affects the estimates.
- Statistician edits should appear in separate fields from the reported data (Vogel, et al., 1985).
- Reported data should be compared with edited data (Vogel, et al., 1985).
- Statistician editing should occur only after data are in computer media (Vogel, et al., 1985).

3. Statistical edit versus data validation.

- Data validation should be distinguished from Statistical editing. Statistical editing would follow data validation (at least in the computer program), (Vogel, et al., 1985).
- Development of statistical edits at the time of edit should be carried out, (Vogel, et al., 1985 & Barr, 1984).
- Data errors identified during the statistical edit and analysis process should be resolved using statistical procedures to ensure consistency across surveys across States (Vogel, et al., 1985).
- Data analysis routines should be incorporated into the statistical edit stage of the edit (Vogel, et al., 1985).
- The system should have the ability to detect outliers and inliers (Vogel, et al., 1985).
- There should be an audit trail for adjustments from outliers (Vogel, et al., 1985 & Barr, 1984).
- Interface with analysis tools should be allowed (Ferguson, 1987).

4. Imputation.

- Imputation procedures for both list and area frame surveys should be incorporated (Barr, 1984).

5. Added capacity and flexibility.

- The system should allow item code validation of questionnaires by state and version.
- The system should have the ability to process states individually with their own error limits.

6. Ease of use.

- The system should have menu driven systems that are easy to use (Vogel, et al., 1985 & Barr, 1984).
- The elimination of the hand edit is a goal (Vogel, et al., 1985).
- Data verification should be on an interactive basis (Vogel, et al., 1985).
- Customized data listings should be available (Ferguson, 1987).
- The system should provide easy and timely access to data at all levels, i.e., reporter, county, district (Barr, 1984).
- Edit checks should be easy to specify (Ferguson, 1987).
- An error description should be provided in the error printout (Ferguson, 1987).

7. System attributes.

- The system should provide the capability to interface with the following (Barr, 1984):
  a. data entry systems
  b. microcomputers/minicomputers
  c. the NASS data management system
  d. statistical packages
  e. LSF
  f. CATI and hand held data recording devices
  g. report generators.
- An interface with the data base should be allowed (Ferguson, 1987).
- The system should start with the capabilities of the NASS Generalized Edit System and build from there (Barr, 1984).
- There should not be an artificial limit to the number of edits (Barr, 1984).
- The system should be capable of editing and maintaining several levels of data within or between records including the ability to use previously reported data (Barr, 1984).
- There should be flexibility in data entry formats including full screen, on and off line procedures, entry without item codes, etc., (Barr, 1984).
- The system should have survey management capabilities (Barr, 1984).
- The system should meet all agency security needs (Barr, 1984).
- A specifications generator should be developed from which files for paper questionnaires, CATI and CAPI instruments, and an editing system can be generated from one specification of the survey variables.

C. Implementation.

The new Survey Processing System is being written and documented. It is being used to process the Quarterly Agricultural Surveys, the Farm Costs and Return Survey, and the Prices Paid by Farmers Survey.

The emphasis so far is on handling expanded survey requirements. These include an increase in the numbers of edits and variables and the use of cross-record checks to improve data validation as it is currently handled in the agency. The system can access previously reported data and since it is in SAS it has the capability of comparing data between records. Though data correction is still by printout, English messages instead of error codes are printed out. It is far easier to write edits than previously and there are no limits to the number of edits that can be written. Research on some features listed above has yet to begin. This includes work on statistical edits, the automation of all or most of the corrections, and the elimination of the hand edit before data entry.

1. Broad aims.

If the objectives of editing are objectivity, repeatability, and statistical defensibility then they have not been fully attained in most of NASS's surveys. The current NASS editing systems primarily use within record computer checks and a subject matter specialist. In that the SPS is written in SAS, it is capable of accommodating procedures which would accomplish these goals. The attainment of these objectives is firstly a matter of definition and theory and secondly a matter of systems development.

2. Impact of editing, contribution of nonresponse.

The Survey Processing System allows a review of the number of edit actions by type but does not allow a review of their effects on the final indications. The contribution from nonresponse for crops is made available to the Agricultural Statistics Board before estimates are set. The corresponding statistics for livestock will be generated starting in September 1988. There is no provision for monitoring how data editing and adjustment for outliers affect the estimates. The system has the capability of allowing the commodity statistician to monitor the relationship between raw and edited data but as this capability has not been used the programming code has been commented out. (That is, it is still there if anyone wants to use it.) The system does not retain survey data as it is reported, nor do statistician edits appear in separate fields from the raw data. The issue of comparing reported data with edited data is problematic because CATI reports and paper reports are mixed in the same file and CATI reports are, in effect, edited at the time of data collection. Statistician edits occur both before and after the data are in computer media, not solely afterwards.

3. Statistical edit versus data validation.

Data validation is distinguished from statistical editing in the Survey Processing System. That is, a place is reserved for a statistical editing module if the research on how best to implement a statistical edit is carried out for each survey. A statistical edit is distinguished from statistical analysis in that a statistical edit is carried out at the time of the data validation. NASS's Prices Paid by Farmers survey employs a statistical edit to search for outliers. The June Agricultural Survey edit program also does some cross-record checking at the field and tract levels but does not check for outliers. An audit trail for adjustments from outliers exists for the Farm Costs and Returns Survey but not for the other surveys. An interface with analysis tools is in place.

The concept of statistical editing in NASS remains undeveloped. NASS's analysis packages serve some of the same purposes as a statistical edit, in that the analysis packages serve to detect outliers. A major difference between the analysis packages and the statistical edit as listed above is that the analysis packages are run after data validation and imputation and not at the time of data validation. Another difference is that edit limits are not generated from the analysis packages. Statistical editing may refer to cross-record comparisons of current survey records for one or a set of variables. The concept may also refer to using historical data within a record in order to set individual error limits for each firm. See the discussion on statistical edits in section II. A. above for elaboration.

4. Imputation.

The capability to implement imputation procedures for both list and area frame surveys does not yet exist within the Survey Processing System. Imputations are being carried out in the Quarterly Agricultural Surveys but these imputations are not done within the SPS. They are carried out between data validation (in the SPS) and analysis. Though imputed values are not rerun through the validation edit it is possible to see the effects of some imputations in the analysis package. If at that point there are any glaring imputation mistakes they can be corrected.

5. Added capacity and flexibility.

The Survey Processing System allows as an option the validation of questionnaires by state and version. The capability of states to process data with their own error limits is also available.

6. Ease of use.

The Survey Processing System has some menu driven systems in place with which to generate parameters for editing and analysis. The elimination of the hand edit is an unrealized goal for NASS as a whole, although there are states that have eliminated it. Data validation is not on an interactive basis, but since the system is in SAS, this could be done either on the mainframe or on microcomputers. In order to install interactive editing, an editing interface would have to be written so that the records could be immediately re-edited when changes are made. Customized data listings are available and this capability is being updated. The system allows easy and timely access to data at all levels because it is written in SAS. The specification of edit checks is much easier than before although not as easy as desired. However, a better specifications generator will be created in the future. Error descriptions instead of error codes are provided in the printout.

7. System attributes.

The Survey Processing System, utilizing the many SAS capabilities, has the capability to interface with all present or anticipated NASS data handling systems. The SPS has all of the capabilities of the old system and many more. There is no artificial limit to the number of edits. The system can edit and maintain several levels of data within or between records and can access previously reported data. Data can be entered in a variety of ways including by item code, full screen entry, and on and off-line procedures. The system has some survey management capabilities but these are to be improved. The system fully meets all of NASS's security needs. The specifications generator has yet to be developed beyond its rudimentary beginnings.

D. Discussion.

Not all of the desired features listed above have been implemented or even researched. However, considerable progress has been made especially in the realm of system development. Much effort has been put forth in order to preserve maximum flexibility in the Survey Processing System. Thus the system has the potential to accommodate almost any mathematical or statistical procedure, to be used on many kinds of computers, and to allow more powerful tools to be put in the hands of the editing personnel. The limiting factors are the development of theory, research, money, and the vision of management.

1. The role of the data editors in the field.

Aside from the listing of some desired features of the system, no explicit discussion has been offered in NASS documents as to the job content of the data editors, that is the data entry personnel, the clerks, and the statisticians, as regards editing. Several scenarios have already been presented in section II C. Though some changes have taken place from the standpoint of the field editors, the SPS has not yet impacted their job in a major way. However, future implementations of the SPS have the potential to change their job content dramatically, from reducing the amount of editing the personnel are expected to do to supplying more powerful tools to do the same job. Keeping in mind the constraints of resources and goals of NASS, such as objectivity, repeatability, and defensibility, the editors themselves should be brought into the process as regards how they are to act in the system. For example, if interactive capability is to be introduced, the people who are to produce with the system should have a hand in the design of the interfaces. This includes clerks as well as statisticians.

2. Computers to be used.

For national surveys, NASS utilizes an IBM mainframe under contract with Martin Marietta Data Systems (MMDS). The decentralized organization of NASS requires that the mainframe be used for at least some survey processing tasks. The role of the microcomputer and Local Area Networks (LAN) in the state offices in the realm of data editing is not clear. Several data editing systems across the world are starting to use microcomputers for at least part of their editing. Some tasks which could be done on the smaller machines include data entry, heads up data entry (where the data are corrected as entered), data verification, imputation, and statistical editing of some types. After having taken care of these tasks the result could then be uploaded to a mainframe for further processing. It is a NASS goal to have a Local Area Network in every state office with the first systems being installed in early 1989. The use of these LANs should be considered a potential productivity tool in this realm. The Survey Processing System does not preclude such possibilities.

3. Integration of CATI and CAPI with the SPS.

Even with the introduction of CATI and CAPI some sort of editing system will still be needed, as not all editing functions are carried out with these data collection technologies. Some consideration will have to be given as to how the editing functions will be divided between the SPS and CATI and CAPI. For example, it is not likely that cross-record checks could be successfully carried out in a CAPI environment if those checks were to involve records collected on another computer. On the other hand, CATI records collected on a LAN could be checked with a statistical edit conducted on a LAN.

4. Research and implementation.

Two research projects are being conducted in the Research and Applications Division. The first, by Antoinette Tremblay and Ralph V. Matthews, is entitled "A Track of Wheat Objective Yield Raw Data to Final Summary". This study tracks data from its reception in the office to final summary. Estimates are calculated at each stage of the editing and summary process to track the effect of the editing process on the level of estimates. The aim of this report falls within the realm of collecting management information. That is, it attempts to measure the effects of current editing procedures. This report, and others like it, will impact the SPS indirectly by pointing out areas in which research should be conducted.

The second project by Cathy Mazur is entitled "Statistical Edit System for Weekly Slaughter Data". The aim of this research is to determine if it is possible to use each firm's historical data in a time series model to edit current data. It is possible that this research will be incorporated directly into the SPS for some specialized surveys, especially for those surveys dealing with agricultural business firms.

A statistical edit has been implemented for the Prices Paid by Farmers Survey. It remains to be seen how far this approach can be extended to other surveys as there are some differences between this survey and the larger NASS surveys. In the Prices Paid survey all reports are run through the system in batch. Records in each of ten regions in the country are compared on an item by item basis and regional distributions are generated for each item. A second run is used to correct data which are outliers based on current reports from the first run. A second type of statistical edit based on historical data is used in this survey. Data from the previous collection period are used to generate error limits for the next period. These error limits are manually reviewed before being used, and can be changed if necessary.

Further research remains to be done in the areas of statistical editing for NASS's large surveys; automated imputation, inspection and analysis of edits in current NASS surveys, macro-edits, statistical edits, and automation of current editing procedures, that is, interactive capability. In these areas four groups of people must, at times, interact: systems developers, statistical researchers, users in headquarters who plan surveys and write editing programs, and data editors in the field. (See further, IV. Future Research.)

5. Statistical defensibility, objectivity, and repeatability.

It is possible that the definition of the term statistically defensible will change with different types of surveys. For example, in some commodity surveys, the sole intent may be to publish averages, totals, or rates. In these cases, the only access to the data would reside within NASS, and in theory at least, all direct uses of the data are known in advance. In the FCRS, there is a fair amount of multivariate data analysis done by ERS, that is, the data goes to a client and all potential uses of the data are not known in advance. In the former case, imputations of averages for item nonresponse may be statistically defensible. In the latter case this procedure probably would not be defensible, as it would have the effect of changing the structure of the data, that is, changing the marginal and joint distributions of the data. (NASS does not impute for the FCRS survey.) As another example, in the Fellegi and Holt vision of data editing and imputation, imputations must agree with edits. In the present NASS processing of the QAS, imputations are not run through the editing system again although implausible imputations may be detected in the analysis packages. Is this defensible? In order to attain the goal of statistical defensibility, the term will have to be defined at some point. The same will have to be done for the terms objectivity and repeatability.

6. Imputation.

Imputation should not be performed to such an extent that major problems in data collection are masked. That is, no imputation method can overcome high nonresponse and error rates. Further, if record level data are to be released to others, then imputations should be flagged, and the manner in which they were made should be documented and available to the user. To the extent that imputations are justifiable, whether they arise from correction of errors or through nonresponse, they should be made in an efficient and defensible manner.

According to Dale Atkinson (1988a), imputations for total nonresponse in the QAS are possible if one makes use of ancillary data such as list frame control data or previous survey data. Imputations in the QAS are based upon extensive modeling of any previous survey or control data which are available for a particular nonresponse record. The current methods do not preserve distribution in which the primary benefit (at least for the QAS) would be in better variance estimation. (This reflects how the data are used, i.e., expansions or averages reported, only NASS uses the data, etc.). The benefits do not outweigh the costs of trying to maintain a distributional structure for each item. Methods now are based primarily upon the logic used when manually imputing data. Atkinson suggests research to: 1) compare the NASS exact imputation procedures against alternative approaches used outside of NASS and widely discussed in the statistical literature, and 2) investigate ways to compensate for variance understatement resulting from imputation. He also states that statistical defensibility needs to be addressed.

## VI. Further Research.

### A. Describe the problem.

Collect management data concerning the editing process. What are the resources being used and on which survey steps are they being used? How much cycling is there in the edit process for each survey? How tight are deadlines to be in the future? How well do enumerators follow skip patterns? What percent of the data is imputed and how does this vary by topic or item? How are estimates changed as a function of the current editing process? The answers to these questions are necessary in order to compare the magnitude of possible improvements to the costs of implementing new systems and to the costs of remaining in the current system.

Current NASS operating procedure involves performing a comprehensive hand edit on questionnaires before data entry. Thus there is no way to measure the quality or necessity of the editing process by referring to computer files. Also it is not possible to measure changes in data between the time they are collected and after they have been processed. The only way to collect this kind of data would be to conduct a survey of questionnaires used in NASS surveys such as the the Farm Costs and Returns Survey, the Objective Yield Surveys, and the Quarterly Agricultural Surveys. The ultimate sampling unit would be the questionnaire. The respondents would be NASS statisticians in the state offices. A stratified random sample of the questionnaires would be collected and each questionnaire would be rated in various ways. Data items would include measures of enumerator performance such as whether the correct skip pattern was followed. Other data items would measure office performance in terms of data quality and productivity including rating the percent of editing activity as it appears on the questionnaire as to whether it was harmful, superfluous, or good (contributed to data quality).

Other management data could be collected from headquarters by tracking certain computer files. The data to be collected here would include the extent of cycling being done for the various surveys as well as the distribution of questionnaire arrivals into the system. For a random sample of offices, every error printout file could be saved for the particular survey. After the edits for that office are clean, a matching process between files would be executed, the matching being done on the ID. It would then be possible to determine a distribution of occurrences of IDs in terms of the number of times each questionnaire passed through the computer edit.

A third type of management data that could be collected would be the time spent on various survey tasks. This would have nothing to do with time sheet recordings as used in the Administrative Records System. This measurement would be done by selected employees keeping a log of how their time was spent on each survey task. It would attempt to detect superfluous activity (e.g., filing and refiling of questionnaires) that would not appear through an inspection of questionnaires. That is, it would help point out where new technologies could streamline the process. It would also provide baseline data for measuring productivity gains.

The purpose of collecting this management information would be to determine where research resources should be spent regarding at least two technologies. These are new editing procedures and the introduction of CAPI. For example, the rate of occurrence of enumerators following a wrong route may be less than 1 % or as high as 10%. In the former case CAPI might not be justified based on this phenomenon alone, whereas in the latter case it might be. Other vital information might be gathered in the realm of nonsampling errors, the effect of imputations whether done by hand or by machine, and the success in training enumerators. In this latter realm one could imagine a whole different emphasis on training according to the feedback which could be gained by this agency self inspection. For example, it may be that data quality is compromised more by enumerators following incorrect paths through questionnaires than by their misunderstanding of individual questions. If so, future training schools for the survey would emphasize following the correct path through the questionnaire.

B. Classification of errors and edits.

Construct a classification of edits and of errors occurring on NASS questionnaires. Determine the success of detecting the errors and the success of correcting them. List the kinds of edits being used by NASS in its surveys according to the edit classification. Determine if the edits in use by NASS would be amenable to the determination of a minimal set and feasible regions. Perhaps a subset of the edits would be amenable to such an approach and could thus be utilized to reduce the need for review by survey specialists. One other aspect of this research is to see if NASS over-edits its data. Statistics Sweden has gained a 50% reduction of error signals by inspecting and analyzing edits in one survey and eliminating redundant edits in some cases and broadening bounds in others (Granquist, 1988b). This reduction in noise was accomplished without a reduction in data quality and according to Granquist, a slow increase in data quality, as specialists had more time to concentrate on true problems. After every survey, NASS reviews the number of times each edit is invoked and will adjust edits accordingly. This recommendation would go beyond this routine analysis by analyzing patterns of edit failures. That is, are there edits which always or usually fail at the same time because they concern the same field?

C. Macro-edits.

Macro-edits should be able to do two things. Edit data at the aggregate level, and trace inconsistencies at the aggregate level to individual questionnaires. Macro-edits focus the analysis on those errors which have impact on published data. These tasks are already performed to a large extent by hand referral to analysis package output. Research should be pursued along the lines of automating macro edits.

D. Statistical edits.

Research on automated statistical editing, both in batch and interactive modes should be conducted. The best way to detect outliers and the best way to resolve the status of the suspicious data should be determined. The use of high resolution work stations in conjunction with interactive data analysis packages should also be explored.

E. Imputation.

Conduct research on imputation, its impact on the edits and the maintaining of distributions, (see Atkinson, 1988a). Some questions which should be answered: What is the extent of item nonresponse, partial nonresponse and total nonresponse? What are the proper imputation methods for each kind of nonresponse? How do these vary by survey topic? For example can item nonresponse in crops and livestock be handled in the same way? By sampling frame? By survey? Is it defensible to hand impute for total nonresponse in JES tracts? Why, in the QAS, are imputed livestock records not used in the summary while imputed crops and grain stocks are fed into the summary? How do agricultural populations differ in structure from other populations and how are imputation procedures used by other organizations applicable to agency needs?

F. Statistical defensibility, objectivity, and repeatability.

Conduct research on a definition of statistical defensibility, objectivity, and repeatability as they apply to the editing and imputation process.

G. The Bureau of Census SPEER software.

Inspect the Bureau of Census software when it becomes available. The Bureau of Census has offered to allow NASS to inspect the software they are developing for the IBM AT.

H. The Netherlands Central Bureau of Statistics Blaise software.

Continue to inspect the Blaise software as it is sent from the Central Bureau of Statistics in the Netherlands. This research should be carried out regardless of the editing research, as it might be applicable to the CATI and CAPI work done in this agency. It could also stimulate research into the concept of integrating all aspects of the survey, from questionnaire design to summary.

I. Microcomputers.

Determine the feasibility of using microcomputers, either solely or in LANs, to perform various editing tasks. NASS already has some microcomputer based editing and summary programs in place for special purposes, including the Peanut Stocks survey, and a system in use in Pakistan. The next logical step is to see if the Survey Processing System on the microcomputer can handle NASS's questionnaires, especially the Farm Costs and Returns Survey and the Quarterly Agricultural Surveys. Possible productivity gains could be estimated if the research in point A is carried out.

References

Atkinson, Dale (1988a). "The Scope and Effect of Imputation in Quarterly Agricultural Surveys." NASS Staff Report Number SSB8804, National Agricultural Statistics Service, U.S. Department of Agriculture.

Atkinson, Dale (1988b). "Travel Notes - Budapest, Hungary." Internal Memorandum, National Agricultural Statistics Service, concerning Atkinson's participation in the second meeting of the Data Editing Joint Group of the United Nations Statistical Computing Project, Phase 2, April 18 to April 22, 1988.

Barr, Jewel (1984). "Edit Specifications Team Report." Memorandum to Deputy Administrator Raymond R. Hancock, Statistical Reporting Service, U.S. Department of Agriculture.

Bethlehem, J. G. (1987a). "The Data Editing Research Project of the Netherlands Central Bureau of Statistics." Staff Report, Netherlands Central Bureau of Statistics.

Bethlehem, J. G., D. Denteneer, A. J. Hundepool, and W. J. Keller (1987b). "The Blaise System for Computer-Assisted Survey Processing." Staff Report, Netherlands Central Bureau of Statistics.

Bethlehem, J.G., D. Denteneer, A. J. Hundepool, and M. H. Schuerhoff (1987c). "Automating the Data Editing Process With the Blaise System." Staff Report, Netherlands Central Bureau of Statistics.

Denteneer, D., J. G. Bethlehem, A. J. Hundepool, and M. S. Schuerhoff (1987). "Blaise, A New Approach to Computer-Assisted Survey Processing." Staff Report, Netherlands Central Bureau of Statistics.

Dinh, Khoan Tan, (1987). "Application of Spectral Analysis to Editing a Large Data Base", *Journal of Official Statistics*, Volume 3, No. 4, 431-438.

Fellegi, I. P. and D. Holt (1976). "A Systematic Approach to Automatic Edit and Imputation", *Journal of the American Statistical Association*, Volume 71, Number 353, Applications Section, 17-35.

Fellegi, I.P. (1975). "Automatic Editing and Imputation of Quantitative Data", *Proceedings of the 40th Session, Bulletin of the International Statistical Institute*, Volume 3, 249-253.

Ferguson, Dania (1987). "Why a New Edit System." Internal Memorandum, National Agricultural Statistics Service, U.S. Department of Agriculture. Presented at the 1987 June Enumerative Survey school.

Ford, Barry L. (1983). "An Overview of Hot-Deck Procedures," *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies*. Ed. William G. Madow, Ingram Olkin, and Donald B. Rubin. New York, N.Y.: Academic Press, 1983.

Giles, P. and C. Patrick (1986). "Imputation Options in a Generalized Edit and Imputation System", *Survey Methodology*, Volume 12, No. 1, 49-60.

Giles, Philip (1987). "Towards the Development of a Generalized Edit and Imputation System", *Proceedings of the Third Annual Research Conference of the U.S. Bureau of the Census*, 185-193.

Granquist, Leopold (1988a). "On the Need for Generalized Numeric and Imputation Systems, Report by Statistics Sweden." Given at the seminar on Statistical Methodology, Geneva, February 1-4, 1988.

Granquist, Leopold (1988b). "A Report on an Evaluation of a Macro-editing Idea Applied on the

Monthly Survey on Employment and Wages in Mining, Quarrying and Manufacturing." Report presented at the Data Editing Joint Group Meeting in Budapest, April 18-22, 1988. Statistics Sweden.

Granquist, Leopold (1987a). "A Report of the Main Features of a Macro-editing Procedure which is used in Statistics Sweden for Detecting Errors in Individual Observations." Report presented at the Data Editing Joint Group Meeting in Madrid, April 22-24, 1987. Statistics Sweden.

Granquist, Leopold (1987b). "The Short Term Developing Program for Computer Supported Editing at Statistics Sweden." Report presented at the Data Editing Joint Group Meeting in Madrid, April 22-24, 1987. Statistics Sweden.

Granquist, Leopold (1984a). "On the Role of Editing", *Statistisk tidskrift*, 1984:2, 106-118.

Granquist, Leopold (1984b). "Data Editing and its Impact on the Further Processing of Statistical Data." Invited paper for the Workshop on Statistical Computing, Budapest, November 12-17, 1984.

Greenberg, Brian and Rita Surdi (1984). "A Flexible and Interactive Edit and Imputation System for Ratio Edits", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 421-426.

Greenberg, Brian (1984). "Using an Edit System to Develop Editing Specifications", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 366-371.

Greenberg, Brian (1981). "Developing an Edit System for Industry Statistics", *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Pittsburgh, Pennsylvania.

Greenberg, Brian (1985a). "Edit and Imputation as an Expert system." Workshop on Statistical Uses of Microcomputers in Federal Agencies, Session on Expert Systems.

Greenberg, Brian (1985b). "Example Illustrating the Need for Implied Edits for Categorical Data." Internal Working Paper, U.S. Bureau of the Census.

Greenberg, Brian (1982a). "Examples Illustrating the Need for Implied Edits for Continuous Data." Internal Working Paper, U.S. Bureau of the Census.

Greenberg, Brian (1982b). "Discussion of 'Imputation in Surveys'", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 32-33.

Greenberg, Brian (1987). "Discussion of the papers 'Towards the Development of a Generalized Edit and Imputation System' by Philip Giles and 'The Data Editing Research Project of the Netherlands Central Bureau of Statistics' by J. G. Bethlehem", *Proceedings of the U.S. Bureau of the Census Third Annual Research Conference*, 204-210.

Hidiroglou, M.A. and J. M. Berthelot (1986). "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology*, Volume 12, No. 1, 73-83.

House, Carol C. (1985). "Questionnaire Design With Computer Assisted Telephone Interviewing", *Journal of Official Statistics*, Volume 1, No. 2, 209-219.

Kalton, Graham and Daniel Kasprzyk (1986). "The Treatment of Missing Survey Data", *Survey Methodology*, Volume 12, No. 1, 1-16.

Kovar, John G. "Generalized Edit and Imputation System, An Overview.' Internal Slides, Research and Development Section, Business Survey Methods Division, Statistics Canada.

Kovar, J. G., J. H. MacMillan, and P. Whitridge (1988). "Overview and Strategy for the Generalized Edit and Imputation System." Working Paper No. BSMD-88-007E, Methodology Branch, Statistics Canada.

Liepins, G.E., R. S. Garfinkel, and A. S. Kunnathur (1982). "Error Localization for Erroneous Data: A Survey", *TIMS/Studies in the Management Sciences 19* (1982) 205-219. North-Holland Publishing Company.

Liepins, Gunar (1983). "Can Automatic Data Editing Be Justified? One Person's Opinion", *Statistical Methods and the Improvement of Data Quality* 205-213: Academic Press, 1983.

Mazur, Cathy (1988). "Statistical Edit System for Weekly Slaughter Data." Upcoming NASS Staff Report, National Agricultural Statistics Service, U. S. Department of Agriculture.

Naus, Joseph I. (1982). "Editing Statistical Data", *Encyclopedia of Statistical Sciences*, Volume 2, 445-461.

Pullum, Thomas W., Trudy Harpham and Nuri Ozsever (1986). "The Machine Editing of Large-sample Surveys: The Experience of the World Fertility Survey", *International Statistical Review*, Volume 54, 311-326.

Sande, G. (1979). "Numerical Edit and Imputation." Invited Paper to the International Association for Statistical Computing, 42nd Session of the International Statistical Institute, Manila, Phillipines.

Sande, I.G. (1988). "A Statistics Canada Perspective on Numerical Edit and Imputation in Business Surveys." Presented at the Conference of European Statisticians, Geneva.

Tortora, Robert D., Frederic A. Vogel, and J. Merrill Shanks (1986). "Computer-Aided Survey Methods", *Computer Science and Statistics, Proceedings of the 18th Symposium on the Interface*, Volume 18, pages 411-413.

Tremblay, Antoinette and Ralph V. Matthews (1988). "A Track of Wheat Objective Yield Raw Data to Final Summary." Upcoming NASS Staff Report, National Agricultural Statistics Service, U. S. Department of Agriculture.

Vogel, Fred, H. Bynum, G. Hanuschak, R. Murphy, W. Dowdy, C. Hudson, and J. Steinberg (1985). "Crop Reporting Board Standards." Report of the Crop Reporting Board Policy and Procedures Working Group, Statistical Reporting Service, U.S. Department of Agriculture.

Glossary of terms found in the editing literature.

Acceptance region

The set of acceptable values defined by the edits for each record. For categorical data the acceptance region can be represented as a set of lattice points in N-space. For numerical data it is a set of convex regions in N-space.

Anticipated Misunderstanding error

A Misunderstanding error that has been anticipated at the time the edit program was created, therefore a Misunderstanding error that the edit system can be expected to catch. See Errors, types of, and also Misunderstanding errors.

Automated Correction

The correction of data errors by computer without human intervention. One aspect of automated data editing.

Balance edit

An edit which checks that a total equals the sum of its parts. Also called an accounting edit. Example: closing inventory = opening inventory + purchases - sales. (Example from Kovar, 1988.)

Between-record edit

Edits carried out on fields involving more than one record in the survey. Statistical edits are an example of between-record edits because distributions are generated on sets of fields over all the records in the survey.

Blaise

The first name of famous French mathematician, Blaise Pascal. (Blaise rhymes with fez.)

Blaise language

A survey processing language based on the Pascal language, used to specify a Blaise Questionnaire and which generates various other survey modules automatically such as CADI, CATI, CAPI, and PAPI as well as a data editing module.

Blaise questionnaire

A specification of a questionnaire in the Blaise language. Not a data collection instrument as such, but in effect a knowledge base from which data collection instruments and editing and summary modules are derived. The aim in producing a Blaise questionnaire is to specify the data only once in the processing of the survey data. If used properly, it defines and documents the survey.

Blaise system

A computer system for survey processing which has as its key component a Blaise questionnaire, and which seeks to reduce mechanical human processing of surveys through computerization. Thus, data are specified only once in the design of the Blaise questionnaire, the program respecifies them automatically for the various survey instruments and editing and summary modules.

CADI

Computer Assisted Data Input.

Categorical data

Data which arise from a categorical scale. Their importance lies in their amenability to theory in determining minimal sets and feasibility regions. Also, in imputation, there are no considerations of scale and it is easier to find donor records based on matching variables than in the continuous case.

CATI Computer Assisted Telephone Interviewing.

CAPI Computer Assisted Personal Interviewing.

CASP Computer Assisted Survey Processing.

Censored estimate An indication computed after adjustments have been made for valid outliers.

Centralized processing The processing of data on one machine or one set of machines under the control of one group of people even if the data is collected and hand edited in widely dispersed areas.

Clean record Record which has no missing values and which passes all edits.

Complete set of edits The union of explicit edits and implied edits. Necessary for the generation of feasible regions for imputation (that is if one wants imputations to satisfy edits).

Conditional edits An edit where the value of one field determines the editing relationship between other fields. Further categorized into conditional numerical and conditional categorical edits. Example, suppose there are three fields A, B, and C. Conditional edits would exist if the relationship between fields B and C as expressed through the edits depended on the value in A.

Consistency edits Checks for determinant relationships, such as parts adding to a total or harvested acres always less than planted acres.

Consistency error An edit failure concerning a consistency edit.

Consistent edits A set of edits which do not contradict each other is considered to be consistent. If edits are not consistent then no record can pass the edits.

Customization In the context of generalized editing and imputation programs, the idea that it is easier to customize the system for a particular survey rather than writing a specialized program from scratch for each survey.

Cyclic processing The phenomenon where a record must pass through two or more computer edit runs in order to pass all edits. Usually involves an error printout, hand correction on the printout, data entry of the corrected data, and resubmission to the computer. This is an undesirable attribute of many systems and one that many organizations would like to eliminate. The Netherlands Central Bureau of Statistics would do this with their Blaise system by doing edits on microcomputers one record at a time until each is resolved. Fellegi and Holt and their following would do it by generating a minimal set of fields that need correction and a corresponding feasible region that encompasses all valid responses.

Database edit checks Edit checks based on outside data that require access to a data base.

Data verification

Checks for: correct units, correct or complete coding, enumerator comments, and entry errors.

Degenerate solution

In the process of finding a minimal set, (i.e., number of fields to be changed in order to satisfy the edits), the finding of more than one minimal set. (Also called multiple solutions.)

Deterministic edit

An edit which if violated points to an error in the data with a probability of one. Example: Age = 5 and Status = mother. Contrast with stochastic edit.

Deterministic imputation

The situation when only one value of a field will cause the record to satisfy all of the edits. Occurs in some situations (such as the parts of a total not adding to the total). The first solution to be checked for in the automated editing and imputation of survey data.

Discernible Negligence error

A Negligence error that can be determined to be an error with Probability one. For example, harvested acres of maize greater than planted acres of maize. See Errors, types of, and also Negligence errors.

Distance function

For numeric data, a function defined on the matching variables of both the candidate and donor records and used to quantify the concept of similarity. Used to find matching records in hot-deck imputation.

Distributed Processing

In an organization with many dispersed offices, the computer processing of the data at each site.

Donation failure

The failure of a recipient record to pass edits after having received donated values from a donor record.

Donor imputation

A method that pairs each record requiring imputation, the candidate record, with one record from a defined donor population as, for example, in hot-deck imputation.

Edit (definition 1)

Logical constraints on the values that each variable can assume.

Edit (definition 2)

Rules that detect prohibited response combinations

Errors, types of

A taxonomy of error types as provided by Granquist, 1984b. Errors are first classified according to the source of the error and the number of cases. The two main types of errors in this classification are Negligence errors and Misunderstanding errors. Negligence errors are further categorized into Discernible errors and Suspicious errors. Discernible errors are categorized into Identified errors and as a consequence, Unidentified errors. Misunderstanding errors are categorized into Anticipated errors and Unknown errors.

In a second scheme, errors are classified on the dimension of the editing problems of a particular survey. Here the classification is into Quality errors and Process Trouble errors. Granquist holds that Process Trouble errors derive mainly from Negligence errors and Anticipated Misunderstanding errors and that Quality errors are likely to be found among the Unknown

Misunderstanding errors. Consult the Glossary for specific definitions of these errors.

Error localization
The identification of the fields to impute. That is, the determination of the minimal set of fields to impute for.

Expert system
A system which contains a knowledge base and a reasoning mechanism.

Explicit edits
Those edits explicitly written by a subject matter specialist. (Contrast explicit edits with implied edits.)

External consistency
This term pertains to relations among the sample units for a given set of variables. It refers to the distributional properties of each variable.

Failed edit graph
As used by the U.S. Bureau of the Census, a graph containing nodes (corresponding to fields) which are connected by arcs (an arc between two nodes indicates that the two fields are involved in an edit failure). Deleting a node is equivalent to choosing that field to be imputed. A minimal set of deleted nodes is equivalent to a minimal set as defined by Fellegi and Holt.

Feasible region
Same as Acceptance Region

GEIS
Generalized Edit and Imputation System at Statistics Canada. This name reflects their belief that editing and imputation are part of the same process.

Generality
A concept that an editing system be able to handle all kinds of surveys by customizing the base system for each survey. (Contrast with Multi-purpose.)

Hand edit
An edit performed by people before data are entered into the computer.

Heads up data entry
A style of data entry where the data entry machine detects errors in the data as they are entered allowing the operator to immediately correct the errors.

Historical edit
An edit which compares historical data to current data at the record level.

Hot deck imputation
A method of imputation in which donor records are taken from the current deck of sample data. (Cold deck refers to the method of imputation where the donor record comes from past survey data). See Ford, 1983.

ID
Identification number.

Identification
In editing, the identification of fields to impute in order to satisfy edits.

Identified Discernible errors
A Discernible error which can be corrected without consulting the source of the information. See Errors, types of, and also Discernible errors.

Implied edit
An unstated edit derived logically from explicit edits that were written by a subject matter specialist.

Imputation

The assignment of a value to a field either for non-response or to replace a recorded value determined to be inconsistent with a set of edits.

Imputation module

A sequence of imputation rules, determined for each field. The imputation rules are developed with the aid of the subject-matter specialist. The rules are prioritized. These imputation rules include structural imputation, statistical imputation, hot-deck imputation, and subject matter imputation.

Injected errors

Errors in which their presence is known but not the details of the magnitude as in key-punching errors or wrong transcription of data.

Inlier

A value which does not greatly deviate from the mean but which should. This determination is made by reference to the sampling frame or other extra-survey knowledge such as an historical edit. In some sense the opposite of an outlier. Concept presented (but not called an inlier) in Vogel, et al., 1985.

Integrated Survey Processing

The concept that all parts of the survey process be integrated in a coherent manner, the results of one part of the process automatically giving information to the next part of the process. The Blaise system is an example of integrated software in which the specification of the Blaise Questionnaire gives rise to a data entry module as well as CATI and CAPI instruments. The goals of Integrated Survey Processing include the one-time specification of the data, which in turn would reduce duplication of effort and reduce the numbers of errors introduced into the system due to multiple specifications.

Internal consistency

This term pertains to relations among the variables for a given sample unit and is the reason for the edits in most survey procedures, (Ford, 1983).

Item nonresponse

A situation in which a great deal of additional information is available for the missing items, not only the information from the sampling frame, but information from other survey items.

LAN

Local Area Network.

Linear edits

Edits arising from linear constraints. For example:
a. a <= F <= b.
b. a + b = c + d.

Local area network

A group of microcomputers hooked together and which share memory and processing resources. Important to editing in that a LAN may be able to handle some editing tasks that might overwhelm one microcomputer at the same time avoiding expensive processing on a mainframe.

Logical edits

In Fellegi and Holt (1976), edits involving only qualitative (coded) data. Also seen in other articles but not adequately defined there.

Macro-edit

Detection of individual errors by: 1) checks on aggregated data, or 2) checks applied to the whole body of records. The checks are based on the impact on the estimates, (Granquist, 1987b).

Magnitude relationships

Relationships of the form X <= aY where a is a constant not equal to 0. Example: salaries <= 0.9*profits. (Term and example from Kovar, 1988.)

Matching

In the hot-deck imputation procedure, the act of matching a donor record to a recipient record.

Matching variables

Those variables used to find a match between a recipient (candidate) record and a donor record.

Micro-edit

Traditional edits performed on record level data. The logical antonym to macro edit.

Micro-macro edit

An editing procedure whereby detailed micro edits are replaced with a combination micro edit and a macro/statistical edit. The micro edits in the latter procedure are less detailed than in the first. The idea is to "develop survey edits based upon an 'impact on the estimates' philosophy rather than a 'catch all data inconsistencies' philosophy". (Leopold Granquist of Statistics Sweden as quoted by Atkinson, 1988b).

Minimal set

The smallest set of fields requiring imputation that will guarantee that all edits are passed. See also "Weighted Minimal Set".

Minimal weighted set

See Weighted Minimal Set.

Misunderstanding errors

Errors that arise due to ignorance or misapprehension of questions, concepts or definitions, also tactical errors as when respondent deliberately gives false information, (Granquist, 1984b). Further categorized into Anticipated errors and Unknown errors. Misunderstanding errors can be systematic and can thus affect the quality of the results if not anticipated or detected. See Anticipated Misunderstanding errors, Unknown Misunderstanding errors, and also, Errors, types of. Contrast with Negligence errors.

Modularity

A concept that an editing and imputation program be written in modules. This would ease the updating of the system and would also ease the implementation to different surveys by isolating those parts which need to be changed into their own modules. Thus some modules will be the same for all surveys, others will change for each survey.

Multipurpose system

A concept that if an editing system cannot be generalized to handle all kinds of surveys, it should at least have the capability to handle the multiple surveys that the organization conducts.

Multivariate edit

A type of statistical edit where multivariate distributions are used to evaluate the data and to find outliers.

Negligence errors

A result of carelessness either by the respondent or by the survey process up to the editing phase, (Granquist, 1984b). These errors are usually random. Further categorized into Discernible Negligence errors and Suspicious Negligence errors. See Errors, types of. Contrast with Misunderstanding errors.

Nonlinear edits

Edits from nonlinear constraints. For example:
a. Ratio edits
b. Conditional edits
(Conditional numerical and conditional categorical)
c. Mixed edits.
The importance of nonlinear edits is that they occur often but are not amenable to theory in the determination of a minimal set. Some nonlinear edits, such as ratio edits, can be cast in a linear form.

Nonresponse

An incomplete questionnaire or a missing questionnaire. See: Item nonresponse, Unit (Total) nonresponse, and Partial nonresponse.

Outlier

Values of items which lie outside of some bound, according to some determination of the bound.

PAPI

Paper And Pencil Interviewing.

Partial nonresponse

A situation in which some data are collected from the respondent but substantial amount of data are missing, as when a respondent abruptly terminates an interview or refuses to answer a particular section (or when an enumerator follows the wrong path in completing a questionnaire).

Process Trouble errors

Errors which if uncorrected will cause problems in the further processing of the data. Compare to Quality errors. See Errors, types of.

Quality errors

Errors which may distort the quality of the data, for example, systematic errors which lead to bias (Granquist, 1984b). Compare to Process Trouble errors. An error may be a Quality error and a Process Trouble error. See Errors, types of.

Quantitative edits

Edits applied to fields measured on a continuous scale.

Ratio edit

An edit in which the value of a ratio of two fields lies between specified bounds. Ratio checks come in two or more varieties, for example, a field-field ratio check or a year-year ratio check. The U.S. Bureau of the Census has implemented an automated editing and imputation system in the special case where all edits are ratio checks.

Record

A magnetically stored, computer readable representation of survey data. Usually there is one record for each questionnaire although it is possible that one questionnaire's data be split up into several records (Wisconsin 1985 Pesticide Survey is an example).

Relational edits

These are checks for relationships that are known to exist even if not strictly determinant. For example: the presence of one item requiring the presence of another.

Reliability of edits

A concept that some edits are more reliable then others, that is fewer false alarms or more true alarms. The measurement of the reliability of edits can in theory be a bookkeeping exercise in the edit system.

Reliability of fields

A concept that some fields tend to be more reliably filled in then others. This measurement of reliability is based upon subject matter specialist expertise. This measurement can be reflected in weights given to each field and used in the determination of a minimal weighted set.

Repeatability

The concept that survey procedures should be repeatable from survey to survey and from location to location. Also that the same data processed twice should yield the same results.

Rigidity

A program that is difficult to change without making errors and creating inconsistent specifications is considered rigid. This is an attribute of many older edit systems. Rigidity is mostly a product of specifying edits "in the form of logically complex and interrelated chains or networks". This is overcome in newer systems by requiring that these edits be specified (or broken down) into a series of simple and unrelated edit rules of a common form. (Fellegi and Holt, 1976.)

Searching

In the hot-deck imputation procedure, the act of searching for a donor record.

Similarity

In numeric data, a concept of closeness of two records based on prescribed matching variables. A distance function is used to quantify this concept according to some criteria.

Specifications generator

A module in an editing system from which files for paper questionnaires, data entry modules, editing software, CATI, CAPI, and summary software are generated. The specifications generator is the unifying feature in Integrated Survey Processing software. In the Blaise system, the Blaise Questionnaire can be considered to be a specifications generator. The specifications generator contains information relating to the data to be collected as well as to the edits to be applied to the data.

Statistical edit

A set of checks 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 (Greenberg and Surdi, 1984). A statistical edit may incorporate cross-record checks, for example, the comparison of the value of an item in one record against a frequency distribution for that item for all records. A statistical edit may use historical data on a firm by firm basis in a time series modeling procedure. A statistical edit can be used to:

a. flag outliers for manual inspection
b. flag outliers for imputation
c. to exclude outlying fields/records from the donor population
d. to flag outliers to the estimation modules.

Statistical imputation

(Example of): 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.

Statistical matching (in hot-deck)

The act of matching a donor record with a receiving record according to some statistical criteria in order to transfer some data from the donor to the recipient.

Stochastic edit
An edit which if violated points to an error in the data with probability less than one. Example: 80 < yield < 120. Contrast with deterministic edit.

Structural edit
Checks based on a logical relationship between two or more edited 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. A check that the structure of the questionnaire is maintained in the data record.

Structural imputation
Structural imputation is used when a structural relationship holds between several variables (example): a total must equal the sum of its parts.

Subject-based edit
Checks incorporating real-world strictures which are neither statistical nor structural, for example, the ratio of wages paid to hours worked must exceed the minimal wage.

Subject-based imputation
(Example of): when a respondent reports in pounds rather than in tons, the correct imputation is that of the number divided by 2000. This is based on the subject matter specialist's knowledge of how some respondents report.

Superfluous activities
Any editing activity which does not add to the quality of data or to the preparation of data to be processed, such as entering a -1 in a box to indicate a blank or computation of totals or balances. Also the reviewing of every computer action.

Suspicious Negligence errors
Negligence errors which cannot be detected with certainty. For example, a crop yield lower than some predetermined bound may either be an error or an unusual valid answer. Contrast with Discernible Negligence errors. See Errors, types of.

Tightness of an edit
A concept relating to the frequency that an edit is invoked. An edit is tight if it is invoked relatively more often than other edits.

Total nonresponse
A situation in which the only information available about the nonrespondent is that which is available from the sampling frame. Synonymous with Unit nonresponse.

Transformations of donor data
The act of scaling (down or up) the donor data to correspond to the size of the recipient firm.

Unidentified Discernible errors
Discernible errors in which correction requires referring back to the source of information. Contrast with Identified Discernible errors. See Errors, types of.

Unit nonresponse
See Total nonresponse.

Unknown Misunderstanding errors
Misunderstanding errors which occur without anyone realizing they are occurring. Granquist (1984b) holds that these types of errors can only be hinted at

by current Generalized Editing systems. He further believes that Macro-editing is a method of detecting the existence of these errors.

Validation edits

Edits checks which are made between fields in a particular record. This includes the checking of every field of every record to ascertain whether it contains a valid entry and the checking of entries in a certain predetermined combination of fields to ascertain whether the entries are consistent with each other. A concept discussed by Vogel, et al., 1985.

Weights

In the Fellegi and Holt school of edit and imputation, weights are assigned to fields based on reliability. The higher the weight the more likely a field will be imputed for (all other things being equal). Weights can also be assigned to edits.

Weighted minimal set

A minimal set (see above) in which fields are weighted according to reliability in generating imputations. All other things being equal, a choice of two or more minimal sets with the same number of elements is made by choosing the minimal set with the higher weight. The weighted minimal set can also be determined by weights assigned to edits.

Within record edit

Another name for a validation edit.