Journal of Official Statistics
Vol. 6. No. 4, 1990, pp. 355–377

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

Mark Pierzchala[1]

Abstract: Two broad approaches to the automation of data editing and imputation are described. One approach maintains the subject matter specialist as the primary reviewer and corrector of errors but seeks to provide the editor with more powerful tools. A major goal in this approach is to integrate editing with other survey functions such as data collection or data analysis. In the other approach, the computer is being developed as a tool largely to supplant the specialist as a data editor. Software is being developed to analyze edits, choose fields to be corrected, and impute acceptable values. The state of the art in four institutions, Statistics Canada, the U.S. Bureau of the Census, the Netherlands Central Bureau of Statistics, and the National Agricultural Statistics Service is reviewed. Purposes of editing, kinds of editing, its role in addressing nonsampling error, cost, and major issues affecting automation are briefly discussed.

Key words: Statistical editing; macro-editing; interactive editing; productivity in survey processing; survey management; survey integration; generalized editing systems.

1. Introduction

Two general approaches to the automation of editing and imputation are being developed. The first keeps the individual at the heart of the process but seeks to place the editor in a more productive and integrated environment. This is the approach followed by the Netherlands Central Bureau of Statistics and the National Agricultural Statistics Service (NASS) in the United States. In the second, the computer largely supplants the human data editor in the detection and correction of errors at the record level. This is the approach followed by Statistics Canada and the U.S. Bureau of the Census. These two broad approaches are not necessarily mutually exclusive as some tasks may be better performed by people and others by computer.

Theoretical and technical advances have made possible the further automation of the editing function. Theoretical work building upon the ideas of Fellegi and Holt (1976) has allowed the development of algorithms that can analyze edits and can choose fields for correction based upon those edits. The

[1] Mathematical Statistician, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C. 20250-2000, U.S.A.

Acknowledgement: The author wishes to thank the following people for supplying information on editing system development within their respective organizations: John Kovar of Statistics Canada, Brian Greenberg of the U.S. Bureau of the Census, Jelke Bethlehem of the Netherlands Central Bureau of Statistics, and Dania Ferguson and many other individuals in the National Agricultural Statistics Service. Any errors of fact, however, are the responsibility of the author. Thanks also to three reviewers of this article as well as to the editor for making many suggestions for improvement.

corrections (imputations) are automatically applied with the goal of satisfying the edits. Rapid advances in microcomputer processing speed, communications, and size of memory have brought great power to the statistical worker. Once freed from the constraints of mainframe operation, the specialist can choose the best processing flow for the survey. The microcomputer also makes possible the expanded use of computer assisted data collection.

Concern with statistical defensibility, economic forces, and increasing processing demands have all served as incentives for these developments. Concern with defensibility is at least partially rooted in the view that the editing and correction of data not only resolve problems, but can also distort the data. Lower computer processing costs in conjunction with steady or rising personnel costs provide an economic incentive for greater reliance on computers and for increasing the productivity of human editors. Often quality can be improved by the same technology that lowers processing costs. Where processing demands increase substantially, previously satisfactory systems may become overwhelmed. These increased demands may include the need to handle larger numbers of variables and records and also more complex survey designs.

Software development is but one way to solve editing problems. Some researchers question whether editing and imputation can actually improve data quality. If not, one must ask what purposes editing and imputation serve and if there are better ways to accomplish them. For example, some kinds of errors might be better detected by statistical methods or by editing at some level of aggregation. Also, the editing process may be able to give feedback for improvements in repetitive surveys.

The editing systems reviewed in this paper are generalized systems. They are designed to handle all of an organization's surveys or at least an important class of them, such as economic surveys. Generalization is important since such systems greatly reduce or eliminate the need to design editing programs for every survey. Within the broad structure provided by a generalized system, the subject matter specialist can customize the system for the specific survey. Note that in the context of this paper the word imputation refers to any action that is taken to change data, whether this be by hand or by machine, and whether it is done before or after data entry. The treatment of imputation is limited to the extent that it is related to editing.

2. Terms of Reference

As this is a comparative study of the implementation of generalized editing systems involving four organizations in three countries, terms of reference are needed. Development of automated editing systems will differ between survey organizations according to how these terms vary in each organization.

2.1. Definition of the term editing

The editing of survey data encompasses many activities. Editing may be considered either as a validating procedure or as a statistical procedure. Additionally, editing can be done at the record level or at some level of aggregation of individual records. It can take place at many stages of a survey, from collection through the summarization of data.

As a validating procedure, editing is a within-record action with the emphasis on detecting inconsistencies, routing errors, and suspicious cases, and correcting them. Examples of validation include: checking that the sum of parts adds up to the total,

that the number of harvested acres is less than or equal to that of planted acres, and that a ratio of values falls within certain bounds.

As a statistical procedure, checks are based on a statistical analysis of respondent data (Greenberg and Surdi 1984). Statistical editing may refer to a between-record (between-firm) 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. 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. Dinh (1987), Hidiroglou and Berthelot (1986), Mazur (1989), Biloq (1989), Pierce and Bauer (1989), and others have developed statistical editing methods. Macro-edits are applied to aggregations of data, perhaps at a summary level or in an economic cell. Statistics Sweden is developing some of these ideas (Granquist 1987a, 1987b). The aim of macro-editing is to find problems at an aggregate level, for example, at the publishing level. It should be possible to trace the inconsistencies to the individual records involved.

2.2. *Goals of development efforts*

Goals cited in the development of editing systems include the statistical defensibility and rationality of the editing process, increased productivity, greater integration of survey functions, and expanded capacity and capability. Statistical defensibility is a concern because editing and imputation have the potential to distort data and because the lack of a proper edit may leave distortions in place. Distortions may involve changes in level of estimates or changes in univariate and multivariate distributions. Definition of the term "statistical defensibility" is a subject of controversy. At the very least, arbitrary procedures should be removed and bias should be recognized and diminished. Changes made during the editing and imputation process should be repeatable, that is, if raw data are run through the process twice, the final result of the two trials should be (nearly) the same. Editing procedures should be consistent across individuals, locations, and time. Not only should an editing process satisfy these criteria, the organization should be able to show that it does. An automated audit trail showing changes in values and reasons for the changes will better enable the organization to defend itself and also to monitor the effects of changes on the estimates.

Increased productivity is a goal because of tightening budgets, lower personnel ceilings, and increased workloads. Though actual measurements of costs associated with editing are rarely done, the editing process consumes substantial resources. Additionally, many resources are expended on activities that are inefficiently carried out or are superfluous. In a study carried out in the Netherlands Central Bureau of Statistics, for example, Bethlehem (1987) found that many editing tasks did not contribute to data quality. There is also a desire to shift resources away from the processing of data to other activities that may contribute more to improved data quality.

Integration of survey functions will result not only in increased productivity, but also in greater data quality. In computer assisted data collection, for example, interviewing, data entry, and editing are all brought together. Edit failures can be resolved with the help of the respondent and there are fewer chances for transcription errors. Also, there is less need for in-office editing and data entry. Integration may also mean shifting some editing to the data analysis and summary portion of a survey.

Integration is made possible by software that can handle two or more related tasks.

Expanded capacity and capability are goals whenever new processing demands exceed the capabilities of old systems. For example, there may be limits to the number of variables or edits that can be handled in a system. Complex survey designs and questionnaire routing may require new data structures. Additionally, old systems may not be able to accommodate the statistical inspection of data or macro-editing in a timely manner.

2.3. Data processing environment, administrative structures, staffing, resources, and survey constraints

By data processing environment is meant the numbers, types, and placement of computers, their interconnections, and the database environment in which they operate. Many statistical organizations are undergoing a transition from centralized mainframe processing to decentralized microcomputer processing. The attributes of these two environments differ in important respects in the context of editing. In mainframe processing, the editor often uses the system indirectly. Other people must key in corrections, schedule and execute jobs, and retrieve the output. Editing jobs must share processing time with other jobs, at times at a lower priority. If the mainframe is not owned then machine time must be paid for. This may force jobs to be run in batch during off-hours and at specified intervals and data to be edited on paper printouts. Cost considerations may also preclude computationally intensive editing procedures. On the other hand, it is easier to maintain and update editing software if it resides on one computer than if it resides on many. It is perhaps easier to enforce consistency between survey locations and individuals and to monitor the progress of many locations if all of the machine editing is done on one machine. Microcomputers are attractive to the extent that they address the negative aspects of mainframe processing. As they proliferate, the data editor will have full control of the machine and will be able to prioritize its use. Once purchased, use of the microcomputer is essentially free. Concerns about its capacity and processing power have diminished. These factors taken together mean that the editor can choose between batch or interactive processing as appropriate, not as the system dictates. A questionnaire that would be treated on two or more occasions in the batch environment can now be resolved at one time. The microcomputer is also proving its worth as a data collection device both in the office for telephone interviewing and in the field for personal interviewing. Thus much editing is done at the time of data collection. On the other hand, microcomputers may present problems of control and consistency because of their large number and dispersion. Errors in the editing software must be corrected on all machines instead of on just one. Special efforts may be needed to monitor survey progress in several locations. These drawbacks of microcomputer are diminished through the use of communications hardware and software. For example, microcomputers in each location can be hooked together in a Local Area Network (LAN). Each LAN can in turn be connected to a central mainframe. Central control is now possible though it is a management as well as a technical issue. The microcomputer is considered a liberating technology. Its presence in large numbers will compel changes in the way editing is done.

A properly managed database environment increases the flexibility of an editing system because data from various sources

can be easily accessed and integrated. It also becomes easier to access historical and control data for a firm and then compare these data to survey responses. Furthermore, databases provide tools for monitoring the effect of editing on the data and for generating reports. Statistics Canada is a leader in this field as the Generalized Edit and Imputation System is embedded in the ORACLE database system (see Section 3.1.2). Shanks (1989) discusses the role of database software in survey processing, including the integration of survey tasks.

The manner in which editing tasks are allocated between administrative units will also have an effect on automation. These tasks include specification of edits, programming of edits, data entry, manual review of forms, resolution of machine generated error codes, coding of questionnaires, and systems coordination. Particular tasks may be performed in many locations. In NASS, for example, data entry, manual review of forms, and resolution of machine generated error signals takes place simultaneously in 44 field offices. In this situation, consistency between data editors has to be reinforced by the processing system itself. In organizations where editing is done in one location, consistency is encouraged through personal communication. Another type of allocation occurs when different tasks are split up between locations. Here the problems are of communication and sharing of information. In the United States Bureau of the Census for example, data entry and data editing are carried out in two widely separate locations for some surveys. Data editors can request copies of the forms but there may be a considerable delay in delivery. In another example involving NASS, edit programming and systems coordination are carried out at the headquarters in Washington, DC while editing is done in the field offices. The people at headquarters have different perspectives from those in the field. Control of the process and consistency will be a priority among those working at headquarters, whereas in the field offices ease of operation is a greater priority. Current machine editing is done on one computer, which makes it easier for headquarters to control the process. On the other hand, the field offices are forced to operate in a batch printout mode due to costs.

The expertise of staff may vary widely between organizations. Some editors are poorly paid entry-level people with no special knowledge of the subject area. In this situation, placing the editor in a more powerful environment will not be sufficient if the editor does not know how to solve a given problem. Magnas (1989) has developed an expert system for one survey that will guide the inexperienced editor through difficult situations. In this way the expertise of a few is made available to many. Where the editors are experts, the organization may well be satisfied by concentrating on improvements in the productivity of its editors.

The manner in which automated editing systems are developed may be a contentious issue. Implementation of new technology will modify the way some editing tasks are done and eliminate others. For example, in a batch system where data editors make corrections on computer printouts, a hand edit before data entry may be necessary in order to reduce the number of error signals produced by the machine edit. In an interactive environment, where error signals are efficiently and dynamically presented, the hand edit may become unnecessary. In systems where the computer not only detects errors but also corrects them, the specialist may offer resistance because of mistrust of the system or the perception of being replaced by a machine. The resolution of these issues is as much a personnel management question as a statistical one.

The resources that the survey organization has to draw on in the automation of editing systems includes the general computer literacy of its staff, numbers of people it can devote to editing, the availability of research and development personnel, the computer hardware available, the software environment, and systems support. Constraints include the size and types of the surveys it conducts, the time frames of the surveys, the amount of data that must be handled, and the number and types of errors to be treated. A large number of questionnaires to be processed in a limited time may preclude personal review of each form. An organization with tight deadlines may not be able to let specialists enter data and correct them simultaneously, as the efficiency of high speed data entry is required. On the other hand, an organization with declining staff numbers and tightening deadlines may be forced to adopt previously unnecessary technologies. It may have to improve the productivity of its personnel in their handling of each record, or allow the computer to handle more routine errors without review.

2.4. *The data and the edits*

Different types of imputations have different effects on the marginal and joint distributions of the data. In item nonresponse, for example, 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 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. 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. The validity of these assumptions should be stated and critically examined. If record level data must be released to other parties, 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. At a minimum, if a large amount of imputation is done, imputations should be flagged and a description of imputation procedures and assumptions included with the data. The types of surveys being processed, such as economic or social surveys, will each present their own peculiarities in the automation of editing. Economic data is predominantly continuous whereas social data is predominantly categorical. Algorithms that work for one data type may not work for the other. Additionally, hot-deck imputation is more difficult to implement in the continuous case because scale must be taken into account.

Other complicating factors in automation that are independent of survey type include the numbers and complexity of routes (branches) in the questionnaire, the degree of interrelation between the fields as expressed through edits, and the type and complexity of the edits themselves. A large number of routes in a questionnaire, in

association with complicated decision rules for determining the correct route, will complicate automation in two ways. First, there is the problem of ascertaining whether or not the routing structure of the editing program follows that of the questionnaire. This is of particular concern where routing edits are incorporated into computer assisted data collection instruments. Mistakes in this software can introduce profound systematic errors into the data (House 1985). Second, each route may require its own special treatment as edits may be conditioned upon the route taken. Additionally, each route will include different sets of variables, the edits of which must be considered as one group. This problem is manifest in systems that analyze edits, determine fields to be corrected, and then make corrections.

The degree of interrelation between variables as expressed through the edits refers to the number of variables that must be considered as a group and to the number of edits that connect this group of variables. Far fewer resources will be required in analyzing the edits of five groups of 10 variables than in analyzing those of one group of 50 variables. This is because the number of combinations of variables increases dramatically as the number of variables in a group increases. Systems that analyze edits require that the edits are linear or of certain types only. For example, ratio edits can be analyzed as one group. Conditional edits can be especially difficult to deal with. Some can be approximated by a linear expression, others can be restated linearly keeping in mind the intent of the original edit. Still other conditional edits can only be dealt with by splitting the file into two or more parts and treating each part separately (Kovar, MacMillan, and Whitridge 1988).

2.5. *Purposes and costs of editing*

See Granquist (1988a) and Pullum, Harpham, and Ozsever (1986) for good discussions on the purposes of editing systems. The authors address the trade-offs between improvements in data quality and costs of editing. In the former paper, Granquist estimates that editing takes from 20 to 40% of survey budgets in periodic surveys at Statistics Sweden and questions whether the benefits are worth the expenditures. The latter paper, which discusses the editing of the World Fertility Survey, reports that estimates derived from *raw* data tapes in six 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. These authors 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 itself could be improved.

Pullum et al. (1986) cite five reasons why the World Fertility Survey implemented stringent editing practices: (a) to increase the yield of the fieldwork, (b) to improve the validity of the findings, that is, to remove systematic errors that may lead to bias, (c) 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, (d) users have more confidence in data that are internally consistent since such consistency reflects on the entire process of data collection and preparation, and (e) the perception that editing is a hallmark of professional survey research.

In his review of the World Fertility Survey paper, Granquist (1988a) maintains that only the third and fourth reasons are real benefits of editing given the way the editing was carried out, that is, through a generalized edit system. Granquist (1984a) describes the following purposes of editing: (a) to give

detailed information about the quality of the survey, (b) to provide basic data for the improvement of the survey, and (c) 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.

Another way to consider the purposes of editing is to assess its role in the reduction of nonsampling errors. This can be done by constructing an error profile of the survey, and for each type of potential error, noting whether the editing system can address it. An editing system cannot address errors of concept or design; it cannot make up for an inadequate sample size nor for massive unit nonresponse; it cannot be applied to data that are not collected due to undercoverage on the sampling frame, and it may give only the barest of indications of respondents' misunderstandings. There are, of course, many problems that a program can detect. However, even if an error is detected, the correction may not be obvious. Often all that can be done is to substitute a plausible answer for the wrong one.

The economic incentive to automation is seen by comparing the rapidly declining costs of computing against labor costs that are either constant or increasing. Kinds of automation considered too expensive five to ten years ago may now be less expensive to apply than retaining a labor intensive status quo. There is, however, a problem in judging the improvements in productivity gained in the implementation of new systems. Accounting systems do not always give information on the costs of each survey task, including that of editing. This information must be gained through special studies such as the one carried out at the Netherlands Central Bureau of Statistics (Bethlehem 1987).

2.6. The future role of CATI and CAPI

Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) are technologies which can perform checks 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 a generalized edit system. A solution to this problem is to develop a system that can be used both for computer assisted data collection and for editing in the office. The Blaise system from the Netherlands has been developed with this goal in mind. If, on the other hand, the organization must collect data via mail, (as must the U.S. Bureau of the Census), or otherwise continue to use paper questionnaires, then further development of in-house validation programs is justified. Another consideration is how many and what kind of checks should be implemented in the CATI and CAPI instruments. For example, the use of historical checks in data collection may bias responses (Pafford 1988). Suspicious errors that are flagged often in the office may unnecessarily slow down the interview in the field. However, if there are edits that cannot be used in CATI or CAPI instruments, then it must be questioned whether they should be used later in the office.

3. The Fellegi and Holt Approach

The literature emanating from this school of thought is concerned primarily with the stage of editing known as data validation. This approach 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 and the U.S. Bureau of the Census are implementing this approach. (A Spanish system called DIA is not covered in this paper. DIA applies the Fellegi and Holt approach to qualitative data. See Garcia-Rubio and Villan (1990).)

Fellegi and Holt (1976) 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. *Implied edits* are 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, pers. com. 1988). Implicit in this approach is that there is only one level of edit failure. Two levels of edit failure, such as the Blaise system's "dirty" and "suspicious" errors are not allowed.

In the 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 goal of this school of thought that all corrections and imputations will pass all edits, although this may not be strictly adhered to in practice.

One of the major objectives 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. This may be done, for example, through the use of hot-deck imputation. 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, to preserve the univariate distributions, or all variables at once, 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.

Implementation of the Fellegi and Holt approach 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 requires operations research methods (Sande 1979). In hot-deck imputation, procedures from operations research 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 categorical imputations are, 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 nonlinear edits (such as conditional checks) are not. 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 of each field. Thus if multiple minimal fields are found, the least reliable set of fields is updated.

3.1. Two manifestations

Both Statistics Canada and 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 system, or are referred to the specialist according to certain pre-determined criteria such as size of firm.

3.1.1. United States Bureau of the Census

The U.S. Bureau of the Census uses an editing system called SPEER (Structured Program for Economic Editing and Referrals) which is based on the philosophy of Fellegi and Holt. SPEER handles continuous data under ratio edits, and has six main components: Edit Generation, Edit Analysis, Edit Checking, Error Localization, Imputation, and Diagnostics. From survey to survey, it is the Imputation module which requires greatest change. 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 Standard Industrial Classification code 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. Greenberg has extended the approach of Fellegi and Holt into the realm of ratio edits. This is done by considering the ratio edits as a set per se, doing what is possible within that set, and then sending the result to the broader system for further processing.

Imputation modules are applied one field at a time. These consist of a series of rules that are used 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 sent 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.

Greenberg (pers. com. 1988) 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 the review. The analyst's review should be done in an interactive mode working with the computer. The objectives of the analyst's job would not fundamentally change though the mechanics and logistics might.

The Bureau of the Census has processed several surveys and censuses with SPEER including the 1987 Census of Construction Industries, the 1987 Census of Manufactures, the 1986 and 1988 Annual Survey of Manufactures, the 1982 and 1987 versions of the Enterprise Summary Report and the Auxiliary Establishment Report, and the 1982 Economic Census of Puerto Rico. The processing of the Census of Construction Industries was done on a mainframe because of the several hundred thousand records involved. For the 1987 Enterprise Summary Report and the 1987 Auxiliary Establishment Report, the bureau employed a combination of mainframe and microcomputers. The mainframe was used for batch processing and the specialists handled referrals on a microcomputer. All of the processing for these two surveys could have been done on the microcomputer, however. The number of cases handled by specialists depends on the referral criteria. Criteria can include the magnitude of changes made by SPEER or the size of the firm involved.

3.1.2. 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 is the Data Collection and Capture (DC2) module, the second is the Generalized Edit and Imputation System (GEIS) module. DC2 is 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 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 would be 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 importance would be passed to the Generalized Edit and Imputation System as a last resort, at which point an effort would be made to resolve all problems by imputation (Kovar, MacMillan, and Whitridge 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 ensure that numeric data are entered in numeric fields and the like, and that control data on the first page are correct. So far, GEIS has been used only on microcomputers for small surveys. GEIS will be used in its first major application in May or June 1990 on a file of Business Income data. 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 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 et al. 1988). This aids in monitoring the edit and imputation process.

GEIS as an editing program consists of four main parts: editing, error localization, imputation, and outlier detection (Kovar et al. 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 occurs in the analysis of the edits. 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 acceptable ranges for all variables, extreme points of the feasible region, and the set of implied edits (Kovar et al. 1988 and Sande 1979). This part of the system aids the analyst in determining whether the edits are meaningful.

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 uses 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. In other words, the error localization procedure handles deterministic cases first. Uncorrected or unimputed records are sent to the imputation procedure. In this procedure, two general methods are available, donor imputation and model-based imputation procedures. Donor imputation is implemented by hot-deck imputation. This is meant to be the primary method. Hot-deck imputation is preferred because it retains the structure of the data. Other procedures include imputation of historic values (which can be trend adjusted), imputation of means, and ratio and regression estimators. These 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 linear edit bounds for future surveys, or to identify outlying values which can be flagged for imputation or for other con-

siderations in subsequent modules (Kovar et al. 1988).

3.2. Editing system features

Each organization develops a list of features it would like to have when developing a new editing system. Desired features can be classified as methodological features, system features, and subject matter specialist features. Methodological features tend to emphasize statistical defensibility and rationality. Defensibility is encouraged by providing the user with methodologically sound modules (Kovar pers. com. 1987). Rationality includes changing as few fields as possible and maintaining the frequency structure of the data. The orderly development and use of the system may be classified as system features. For example, modular programming eases the development and updating of the system, which can proceed independently of edit specification. The system should be portable between computers of varying types and should be embedded in a relational data base management system (Sande 1988). From the viewpoint of the specialist the system should be comprehensible to and readily modifiable by users, as well as flexible and satisfying (Kovar pers. com. 1987). The specialist should be able to get feedback from the system, including an analysis of edits before the survey is conducted as well as on how the edits and imputations are affecting the data once the survey is underway. See Pierzchala (1988) for detailed lists of these features.

4. Editing as Part of an Integrated Survey System

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 one, of the automation of the overall survey process. No new theoretical tools are implemented to rationalize the editing process. The cyclic batch editing process performed on mainframes is transferred to microcomputers creating an interactive process. The emphasis is on streamlining editing practices and on integrating most computerized survey functions including computer assisted data collection. A commitment to microcomputer processing is an important part of this development. The results of a "Data Editing Research Project" carried out in the mid-1980s provided impetus for this development (Bethlehem 1987). 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. Other problems were found in the work flow of the editing process. The process was cyclic and was often carried out in more than one department on many different computers using different software. Data specification was carried out repeatedly for the various tasks, for example, for the questionnaire, the edit program, the computer assisted interviewing instrument (if any), and the summary program.

As a result of these problems, the Netherlands Central Bureau of Statistics has developed an automated survey processing system called the Blaise system. It 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 and philosopher Pascal.) The key element of this system is the Blaise questionnaire. This questionnaire is not a survey collection instrument itself but rather a specification of questions, routing instructions, and edit checks entered into a computer. Once the Blaise questionnaire is

correctly specified, software instruments with which to process the survey are automatically generated. These include a module for intelligent data entry and interactive error correction, survey instruments such as Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), and an interface with standard statistical packages such as SPSS (Bethlehem, Hundepool, Schuerhoff, and Vermeulen 1989). 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 still corrects the data, only the tools have changed, 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 system does not present options for the correction of errors but it can be programmed to carry out imputations without human intervention. An interactive coding module has been implemented in Blaise that assists in the coding of open ended questions such as occupation.

4.1. The role of the subject matter specialist

The data editors are held to possess great expertise in the subject. The aim is to make greater use of this expertise in placing the editor in a more productive environment. The subject matter specialist has two important roles, the specification of the Blaise questionnaire and the resolution of errors in the data. In the specification of the questionnaire, the expertise of the specialist is written into the Blaise data collection and processing instruments. This expertise is reflected in the specification of edits and their error messages, the particular variables that are flagged with the message, routes, and the wording of questions and possible responses. Most of these specifications are made as well in the cyclic mainframe system. Now, however, the specialist must take into account the interactive environment including how the system is to work during the heat of an interview. By projecting this expertise into the interview through properly specified edits and error messages both quality and productivity may be improved; quality because reporting errors can be cleared up immediately with the respondent, and productivity because the need for in-office editing and data entry is reduced. (See House (1985) for a good discussion of problems in constructing a CATI instrument.)

In the resolution of errors from paper forms, the data can be entered by very fast data entry personnel or by the specialist. In the former case, the raw data file and the questionnaires are passed to the subject matter specialist. The records can be edited in batch or one at a time. For those records that are in error, corrections are made on the computer screen. After the corrections are made, the record is re-edited on the spot and redisplayed with new error signals, if any. Thus the mainframe 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 specialist sees the error signals on the microcomputer screen along with all the data from the questionnaire. Error messages as well as the question statements are available through pop-up windows. The values of calculations can be displayed on the screen or in error messages, for example, "The ratio of profits this month to profits last month is *1.50* which is greater than the upper limit of 1.25." Tables in question-

naires can also be handled in Blaise. Unresolved records can be flagged for later treatment, if necessary. Blaise displays 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. In another processing scheme, 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.

4.2. Implementation

In the Netherlands Central Bureau of Statistics, the system has been in production since 1987. A conversion to Blaise for editing and data entry has been carried out for virtually all surveys and as a CATI or CAPI instrument in selected surveys (Bethlehem pers. com. 1990). This includes economic, social, and demographic surveys. There has been a dramatic decrease in processing time, up to 50% in some cases. The percentage improvement depends on how things were done before. The largest improvements are in surveys which had used the old cyclic paper printouts. There have not been any specific measurements of quality improvements. Global investigations indicate that quality is at least as good as before. In the Netherlands's CAPI experience, tremendous gains in productivity have been realized in the Labor Force Survey. It used to take two years to process data from this survey because of the massive hand editing and coding that was done in the office. The Labor Force Survey is now being collected on hand held computers. Processing time has gone down from two years to two weeks as a result. Once the information arrives in the office, only one variable has to be coded which is done the same day as the data arrive (Bethlehem pers. com. 1989).

In using Blaise, the first step in the survey process is to design a Blaise questionnaire. It acts as a knowledge base in an artificial intelligence system. The Blaise system is the reasoning mechanism which uses the knowledge base in the Blaise questionnaire to produce other software products such as the data entry and data editing module and the CATI and CAPI module. 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, Bethlehem, Hundepool, and Schuerhoff 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 and other kinds of paragraphs 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.

4.3. The desirable features of the system

Editing automation should be part of a larger automation process. This is really the crux of the matter. A Blaise questionnaire (specifications generator) is constructed from which all products are generated including data entry and edit programs, and CATI and CAPI instruments. Error checking should be an intelligent and interactive process carried out between the subject

matter specialist and the computer. Superfluous activities should be eliminated and the cyclic nature of editing should be removed. The system should be applicable to different surveys and the updating of questionnaires from one survey to another should be easy (Bethlehem, Denteneer, Hundepool, and Keller 1987). An interface with statistical packages should be possible without the necessity of respecification of the data. See Pierzchala (1988) for a detailed list of desirable features. See Netherlands Central Bureau of Statistics (1987) for an excellent description of the automation process in that organization.

5. Integrating Editing into the Survey Processing System

The theme underlying NASS's Survey Processing System (SPS) is one of integration. The term integration affects the SPS in two major ways. First, it refers to one impetus of the development of the SPS, that is, the integration of NASS surveys into a coherent sequence of surveys. This was originally done under the name of the Integrated Survey Program and is now known as the Agricultural Survey Program. Second, it refers to the integration of the distinct steps of the survey process under a unifying system. As such, the SPS performs data validation, statistical and macro-editing, deterministic imputation, analysis, and summary. It generates reports and will have further connections to a public use data base and secure agency data base.

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 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 would not allow data to be compared between records (Ferguson 1987). Beyond these system limitations, a number of new capabilities were 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 nonresponse 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 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. NASS is probably one of the few agencies in the world to have programmed its editing system in the statistical language SAS. The use of SAS for editing has interesting potential 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 1988). However, because of the way it is written, the SPS cannot be used on microcomputers without major modifications.

5.1. Differences in integration between the SPS and the Blaise system

There are two essential differences between the integration of survey activities as developed by NASS and the Netherlands Central Bureau of Statistics. First, in NASS the initial emphasis is on the integration of the latter parts of the survey process, that is between editing, analysis, machine imputation, and summary. In the Netherlands, the emphasis is on the integration of the earlier parts of the survey process, that is between editing,

data entry, and data collection. Second, NASS's system was envisioned as a mainframe based system whereas the Netherlands Central Bureau of Statistics made an early and determined commitment to decentralized microcomputer processing. However, NASS is currently procuring powerful microcomputers to be placed on the desk of virtually every statistician and clerk in its 44 field offices. In each office, the microcomputers are to be connected in a Local Area Network (LAN). Each LAN will in turn be connected to the same mainframe. The arrival of these microcomputers will compel changes in the development of the SPS. For example, *interactive* validation editing was not part of the original plans in the SPS. The arrival of the microcomputers on every desk will allow interactive editing to be performed as the agency is freed from the expense of carrying out the machine edit on one mainframe. NASS has just completed a research project on the use of interactive validation editing. Netherlands's Blaise system was used in these trials in order to determine how interactive editing could be carried out in NASS. The group that conducted the project recommended that NASS adopt interactive editing as its standard processing mode and that Blaise be purchased to do that processing (Pierzchala et al. 1990). The LANs will also allow NASS to expand its CATI operation from its current 14 sites to each field office. The manner in which all the various survey tasks are finally to be integrated, from survey specification through summary, remains to be worked out.

5.2. *Implementation of the Survey Processing System*

The Survey Processing System is being developed and documented. The Data Validation, the Sample Select, and the Analysis and Summary modules have all been completed. The completed modules are being used to process the Agricultural Surveys, the Farm Labor Survey, the Farm Costs and Return Survey, and the Prices Paid by Farmers Survey. One emphasis is on handling expanded survey requirements. These include an increase in the numbers of edits and variables allowed, and the use of cross-record and historical checks to improve data validation. The SPS also has the capability to handle many versions of a questionnaire at once, to handle sections within questionnaires differently, as well as to handle states individually. Though data correction is still by printout, error signals may be presented in English rather than in a number code. It is far easier to write edits than before and there are few limits to the number and types of edits that can be written. Menus are in place with which to generate parameters for editing and analysis. Data listings are available which states can customize for their own needs.

The SPS allows a review of the number and types of edit failure but does not allow a review of their effects on the expansions. Contributions from nonresponse for crops and livestock are made available to the Agricultural Statistics Board before estimates are set. 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 edited at the time of data collection. Additionally, statistician edits occur both before and after the data are in computer media, not solely afterwards. It is impossible to determine the effect of changes that were made before data entry except by special study.

5.3. *Statistical editing and macro-editing versus data validation*

Statistical editing and macro-editing are

integral parts of the Survey Processing System. For example, survey expansions between the current quarter and the previous quarter are generated and compared at the stratum level, that is, at a level of aggregation below the publication level. NASS's analysis packages carried out these types of edits between the validation editing and the summary parts of the survey in what were called analysis packages. The concepts of statistical editing and macro-editing in NASS have been reviewed and specifications for these procedures have been rewritten for each of the surveys processed in the SPS. The statistical and macro procedures found in the anaysis packages will be incorporated into either the validation edit or the summary. For example, statistical editing as a part of the validation edit may be invoked when a certain percentage of the questionnaires are received or on a specified date, whichever occurs first (Nealon et al. 1988).

6. Further Research

The role of the editing function in ensuring data quality needs to be assessed. This includes its role in a current survey period as well as providing feedback for improvement of periodic surveys. The type and amount of resources expended on editing and the manner in which they are expended can then be compared to the benefit gained. Granquist (1984a, b) has developed some of these ideas. See also Bethlehem (1987) and Linacre and Trewin (1989). A related idea is that editing may be a distorting process whereby data may be unjustifiably altered. Studies that compare estimates based on raw data to those based on edited data may be valuable in evaluating the purpose of the editing function (Pullum et al. 1986; and Greenberg and Petkunas 1986). However, the effect of data correctly reported but in the wrong units should be taken into account. For example, if a respondent gives an answer in dollars rather than in thousands of dollars as requested, it is not a distortion of the data to convert the answer into the correct units. Also to be taken into account is the effect of hand editing before the data are entered into the computer. The development of automated audit trails and the automatic generation of reports at the end of each survey period should continue. This would reduce the need to conduct special research into the effect of editing on the data.

Research should continue into making the editing function more productive. Better ways of setting error limits may reduce needless and distracting error signals. Gains may be realized by replacing some validation edits with statistical or macro-edits. Further development of concepts of statistical editing (Biloq 1989; Dinh 1987; Hidiroglou and Berthelot 1986; Mazur 1989; Pierce and Bauer 1989) and macro-editing (Granquist 1987a, 1988b) is necessary. Within the interactive environment, more productive ways of presenting information to the data editor should be investigated. Determination should be made of activities that can be reduced or eliminated through the implementation of new technology or procedures, including the hand edit before data entry. Also important is the determination of the proper domains of action for person and machine and how the boundary between these domains may shift with the advent of new technology. Another area of investigation in this realm is the extent to which expert systems can resolve error signals or guide inexperienced data editors through difficulties (Greenberg 1985; and Magnas 1989). Research should continue into the extent to which the editing function as a separate survey task can be reduced and integrated into other survey tasks. The role of relational data bases in the

integration of the editing function into other functions should also be investigated.

Further theoretical development is needed in those systems that automatically correct errors subject to the constraints of the edits. Especially confounding is the treatment of conditional edits. Ways to include variables that may take on negative values should be found. Procedures of choosing one minimal set where many are generated need further elaboration. The presence of blanks (do they represent zeroes or item nonresponse) is also a source of concern.

7. Summary

Great strides are being made in the further automation of editing and imputation systems. Advances in technology and theory have permitted these strides. Concerns with statistical defensibility as well as economic forces have provided incentives for this development. At the same time researchers have questioned the purposes of the editing function in the reduction of nonsampling errors. Editing methods such as statistical or macro-editing, as opposed to validation editing, have been proposed and developed. The editing function has been integrated into data collection through CATI and CAPI, and into analysis and summary. Statistical organizations will develop their own systems according to the "terms of reference" set out in Section 2 and according to other criteria as well. None of the four systems presented does everything, however, they may serve as useful models for parts of the editing process.

Benefits from all of this activity may be great. The (sometimes large) resources that are spent on editing may be reduced allowing a shift of resources to other survey functions. Subject matter specialists may be free from the tedious review of many unimportant error signals in order to concentrate on the more difficult errors. Defensibility may be made easier by providing the user with statistically sound procedures and by the implementation of automatically generated audit trails. Reports generated from the audit trails and from error signal counts may give feedback on survey performance allowing improvement in future surveys. Gone are the days when data quality is considered good because data have passed through a rigorous edit. Now one must recognize the editing function for what it can do in reducing nonsampling errors, perform this function efficiently, and concentrate on other methods for improving data quality.

Appendix

A glossary of some editing terms

These are definitions of terms used in the paper and are not intended to be taken as a general terminology

- Balance edit: A type of consistency edit which checks that a total equals the sum of its parts. Example: closing inventory = opening inventory + purchases − sales (Kovar per. com. 1987).
- Between-record edit: An edit carried out on fields involving more than one record in the survey. Some statistical edits are examples of between-record edits as when distributions are generated on sets of fields over all the records in the survey.
- Categorical edits: An edit applied to fields measured on a categorical scale.
- Complete set of edits: The union of explicit edits and implied edits. It is necessary for the generation of feasible regions for imputation in the Fellegi and Holt approach.
- Conditional edit: An edit where the value of one field determines the editing relationship between other fields. For

example, suppose there are three fields A, B, and C. A conditional edit would exist if the relationship between fields B and C as expressed through the edits depended on the value in A.

- Consistency edit: A check for determinant relationships such as parts adding to a total or harvested acres always less than or equal to planted acres.
- 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.
- 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.
- Deterministic imputation: The situation when only one value of a field will cause the record to satisfy all of the edits. It is the first solution to be checked for in the automated editing and imputation of survey data. For example, when parts of a total are listed but not the total, the correct solution is to calculate and impute the total.
- Explicit edit: An edit explicitly written by a subject matter specialist.
- Historical edit: An edit which compares data from a previous survey period to current data.
- Hot-deck imputation. A method of imputation in which values of fields are taken from donor records from the current file of sample data and are imputed into fields that are blank or incorrect in the receiving record.
- Implied edit: An unstated edit derived logically from explicit edits.
- Imputation: The assignment of a value to a field either for nonresponse or to replace a recorded value determined to be inconsistent with a set of edits. In this paper, imputation refers to both hand and machine correction.
- Linear edit: An edit arising from linear constraints. Examples: $a \leqslant F \leqslant b$, or, $a + b = c + d$.
- Macro-edit: An edit that detects individual errors by: (1) a check on aggregated data, or (2) a check applied to the whole body of records. The checks are based on the effect on the estimates (Granquist 1987b).
- Micro-edit: An edit performed on record level data.
- Nonlinear edit: An edit arising from nonlinear constraints. For example: (a) ratio edit, (b) conditional edit, and (c) mixed edit. 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.
- Quantitative edit: An edit 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-to-field ratio check or a year-to-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.
- Statistical edit: A check 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 Surid 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.
- Statistically defensible survey: A survey

whose procedures and specifications can stand up to court challenge, official investigation, or scientific scrutiny. The procedures and specifications must be adequate to meet the purpose of the survey. Attributes of a defensible survey include: a sample size sufficient to give the necessary precision, randomly selected respondents, carefully worded questions, professional and neutral interviewing (where applicable), reasoned editing practices, correct summarization, and appropriate publication. Arbitrariness must be removed and all bias should be recognized and diminished. Not only must the survey be conducted according to these criteria, the organization must be able to show that it has been conducted thusly.

- Structural edit: A check based on the routing relationship between two or more fields. For example, in a skip pattern in a questionnaire, two variables lying on disjoint paths cannot both be positive. It is a check that the structure of the questionnaire is maintained in the data record.
- Subject-based edit: A check incorporating strictures met in practice which are neither statistical nor structural, for example, the ratio of wages paid to hours worked must exceed the minimal wage.
- Suspicious edit: An edit which if violated points to an error in the data with probability less than one. Example: $80 <$ maize yield < 120.
- Validation edit: An edit that is made between fields within 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.

8. References

Atkinson, D. (1988). Travel Notes – Budapest, Hungary. (Documentation of 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), Internal memorandum, National Agricultural Statistics Service, U.S. Department of Agriculture.

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

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

Bethlehem, J.G., Hundepool, A.J., Schuerhoff, M.H., and Vermeulen, L.F.M. (1989). Blaise 2.0, An Introduction. Staff report, Netherlands Central Bureau of Statistics.

Biloq, F. (1989). Analysis on Grouping of Variables and on Detection of Questionable Units. Staff report of the Business Surveys Method Division, Statistics Canada.

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

Dinh, K.T. (1987). Application of Spectral Analysis to Editing a Large Data Base. Journal of Official Statistics, 3, 431–438.

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

Ferguson, D. (1987). Why a New Edit System. Internal memorandum, National Agricultural Statistics Service, U.S. Department of Agriculture.

Garcia-Rubio, E. and Villan, I. (1990). DIA System: Software for the Automatic Editing of Qualitative Data. Paper presented at the Bureau of the Census 1990 Annual Research Conference, March 18–21, 1990, Washington D.C., National Statistical Institute of Spain.

Giles, P. and Patrick, C. (1986). Imputation Options in a Generalized Edit and Imputation System. Survey Methodology, 12, 49–60.

Granquist, L. (1984a). On the Role of Editing. Statistisk tidskrift, 22, 106–118.

Granquist, L. (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.

Granquist, L. (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.

Granquist, L. (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.

Granquist, L. (1988a). On the Need for Generalized Numeric and Imputation Systems. Report by Statistics Sweden presented at the Conference of European Statisticians, Seminar on Statistical Methodology, Geneva, February 1–4, 1988.

Granquist, L. (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.

Greenberg, B. (1985). Edit and Imputation as an Expert System. Paper presented at the Workshop on Statistical Uses of Microcomputers in Federal Agencies, Session on Expert Systems.

Greenberg, B. (1987). Discussion on 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.

Greenberg, B. and Surdi, R. (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, B. and Petkunas, T. (1986). An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Census in Business Division. Report number: Census/SRD/RR-86/04, U.S. Bureau of the Census, Statistical Research Division.

Hidiroglou, M.A. and Berthelot, J.M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73–83.

House, C.C. (1985). Questionnaire Design With Computer Assisted Telephone Interviewing. Journal of Official Statistics, 1, 209–219.

Kovar, J.G., MacMillan, J.H., and Whitridge, P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Working paper no. BSMD-88-007E, Business Survey Methods Division, Methodology Branch, Statistics Canada.

Linacre, S.J. and Trewin, D.J. (1989). Evaluation of Errors and Appropriate

Resource Allocation in Economic Collections. Proceedings of the U.S. Bureau of the Census Fifth Annual Research Conference, 197–209.

Magnas, H.L. (1989). An Expert System to Assist in the Disposition of Computer Edit Error Flags. Paper given in the short course: Quality Assurance in the Government, Washington Statistical Society, Washington, D.C., Energy Information Administration, U.S. Department of Energy.

Mazur, C. (1989). A Statistical Edit for Livestock Slaughter Data. NASS research report number SRB-90-01, National Agricultural Statistics Service, U.S. Department of Agriculture.

Nealon, J., Bass, B., Dillard, D., Dantzler, M., Ferguson, D., Hohenbrink, K., Kott, P., Pafford, B., Pense, R., Roehrenbeck, M., and Van Lahr, C. (1988). Specifications for Integrating the Edit and Analysis. Internal working paper, National Agricultural Statistics Service, U.S. Department of Agriculture.

Netherlands Central Bureau of Statistics (NCBS) (1987). Automation in Survey Processing. CBS select 4, Voorburg/ Heerlen, Netherlands.

Pafford, B.V. (1988). The Influence of Using Previous Survey Data in the 1986 April ISP Grain Stocks Survey. Staff report number SRB-88-01, National Agricultural Statistics Service, U.S. Department of Agriculture.

Pierce, D.A. and Bauer, L.L. (1989). Tolerance-Width Groupings for Editing Banking Deposits Data: An Analysis of Variance of Variances. Finance and Economics Discussion Series number 72, Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board, Washington, D.C.

Pierzchala, M. (1988). A Review of the State of the Art in Automated Data Editing and Imputation. Staff report number SRB-88-10, National Agricultural Statistics Service, U.S. Department of Agriculture.

Pierzchala, M., Hohenbrink, R., Matthews, W., Stewart, R., Schuchardt, R., Yost, M., Ferguson, D., and Graham, M. (1990). Interactive Editing Research: A Report and Recommendations by the Interactive Editing Working Group, National Agricultural Statistics Service, U.S. Department of Agriculture.

Pullum, T.W., Harpham, T., and Ozsever, N. (1986). The Machine Editing of Large-sample Surveys: The Experience of the World Fertility Survey. International Statistical Review, 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, Philippines.

Sande, I.G. (1988). A Statistics Canada Perspective on Numerical Edit and Imputation in Business Surveys. Paper presented at the Conference of European Statisticians, Geneva, February 1–4, 1988.

Shanks, M.J. (1989). Information Technology and Survey Research: Where Do We Go From Here? Journal of Official Statistics, 5, 3–21.

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

Received November 1988
Revised May 1990