UNITED NATIONS STATISTICAL COMMISSION
and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS
METHODOLOGICAL MATERIAL

GLOSSARY OF TERMS ON STATISTICAL DATA EDITING

UNITED NATIONS
Geneva, 2000

CONTENTS

PREFACE

The methodological material "Glossary of Terms on Statistical Data Editing" was prepared based on the request of countries participating in the activities on Statistical Data Editing organised by UN/ECE Statistical Division within the framework of the programme of work of the Conference of European Statisticians. It represents an extensive voluntary effort on the part of the participants of the UN/ECE Work Sessions on Statistical Data Editing.

The document was reviewed at the Work Session on Statistical Data Editing in June 1999. National Statistical Offices of the UN/ECE member countries and the Food and Agriculture Organisation (FAO) participated in this meeting. The material reflects the outcome of the discussion on the document.

At its 2000 plenary session, the Conference of European Statisticians agreed to have this document reproduced and distributed to interested statistical offices as methodological material.

EXECUTIVE SUMMARY

By its very nature, a glossary of terms will never be complete. This is particularly true when it concerns a topic that is relatively new, as editing and imputation certainly is. This notwithstanding, the Work Sessions on Statistical Data Editing under the Conference of European Statisticians programme of work and its predecessor, the Joint Group on Data Editing under the Statistical Computing Project of the United Nations Development Programme, took it upon itself to come up with such a glossary as part of the programme of work of the Conference of European Statisticians. The present document is the fruit of that labour over the past decade. Many of the group's members contributed, but particular thanks are owed to Dania Ferguson (U.S. National Agricultural Statistics Service) and William Winkler (U.S. Bureau of the Census) for their efforts in, respectively, initiating and completing the work.

The reasons for starting this endeavour were varied. Some no longer hold, but most do. The primary concern was that many different editing and imputation activities were often referred to by the same name. By contrast, other activities carried many different names, even within a single organization, let alone between different countries. In fact even the term 'editing' was, and still is, understood by many to mean many different things. Standardization of the editing terminology is clearly needed. It is only when we have a common understanding of the terminology that we can pursue meaningful and efficient discussion of the topic. We hope this document will help experts working on data editing in national statistical offices.

This glossary is a repository of terms related to statistical data editing and imputation. It contains definitions of concepts, principles, techniques as well as methods. The current version defines over 180 terms. By making the glossary available to novices and experts alike, it at the same time instructs as well as invites critique. It also promotes a common understanding of terminology and thus initiates a framework for sharing expertise and experience.

What this glossary does not contain are practical examples and specific solutions. In particular, a list of computer systems that put into practice the defined techniques is notably absent. Related terms need to be linked, missing terms added, outdated definitions modified. As noted above, this is clearly not a complete document, but its usefulness will be derived only if it is made public, and if the public is invited to make it grow. Proposed additional terms, commentaries and examples are invited from readers in order to enable this 'living document' to evolve in a way that would make it remain as current and up to date as possible. The comments and suggestions should be sent to: info.stat@unece.org.

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 (N-dimensions of real numbers). Also called a feasible region.

ACCEPTANCE RULE

Logical or arithmetic condition applied to a data item or data group that must be met if the data are to be considered correct.

ACTIVE FIELD

A data item (field) for which some values of this data item create a conflict in combination with values of other data items.

ANALYSIS OF CORRECTION RULE SPECIFICATIONS

Verifying consistency of correction rule specifications, mainly in an extensive set of check and correction rules.

ANALYSIS OF EDIT RULE SPECIFICATIONS

An activity by which the consistency of a set of check rules is ascertained, implied (derived) check rules are created, and an economical form (reduction) of specifications of the originally large number of edit (check) rules is determined.

ANALYTICAL EDITING

Edit rule proceeding from a logical reasoning.

AUDIT TRAIL

A method of keeping track of changes to values in a field and the reason and source for each change. Audit trails are generally begun after the initial interview is completed.

AUTOCORRECTION

Data correction performed by the computer without human intervention. It makes particular use of redundancy. Exclusion (elimination) of incorrect records or substitution of a record or its part by data from other records or the correction base. Auto-correction is generally done according to rules that assure the final (corrected) record fails no edits.

AUTOMATED DATA ADJUSTMENTS occur as a result of computer actions. A desirable option in any system allowing computer actions is to allow for the overriding of those actions at some level. Batch data adjustment results in a file of corrected (edited/imputed) records with accompanying messages to report on the computer actions taken to make the adjustments.

AUTOMATED DATA REVIEW may occur in a *batch* or *interactive* fashion. It is important to note that data entered in a heads-down fashion may later be corrected in either a batch or an interactive data review process.

AUTOMATED IMPUTATIONS generally fall into one of six categories:

a. **Deterministic imputation** - where only one correct value exists, as in the missing sum at the bottom of a column of numbers. A value is thus determined from other values on the same questionnaire.

b. **Model based imputation** - use of averages, medians, regression equations, etc. to impute a value.

c. **Deck imputation** - A donor questionnaire is used to supply the missing value.

   ***Hot-deck* imputation** - a donor questionnaire is found from the same survey as the questionnaire with the missing item. The **"nearest neighbour"** search technique is often used to expedite the search for a donor record. In this search technique, the deck of donor questionnaires comes from the same survey and shows similarities to the receiving record, where similarity is based on other data on the questionnaire that correlates to the data being donated. For example: similar size and location of farm might be used for donation of fuel prices.

   ***Cold-deck* imputation** - same as hot deck except that the data is found in a previously conducted similar survey.

d. **Mixed imputation** - In most systems there usually is a mixture of categories used in some fixed rank fashion for all items. For example, first a deterministic approach is used. If it is not successful, then a hot deck approach is tried. This is followed by a model-based approach. If all these approaches fail, then a manual

imputation occurs through a human review process.

e. **Expert Systems** - An expert system is an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution. Every expert system consists of two principal parts: the **knowledge base** and the **inference engine**. The knowledge base contains both factual and heuristic knowledge. Factual knowledge consists of items commonly agreed upon by spokesmen in a particular field. **Heuristic knowledge** is the less rigorous, more experiential and more judgmental knowledge of performance or what commonly constitutes the rules of "good judgement" or the art of "good guessing" in a field. A wisely used representation for the knowledge base is the rule or if /then statement. The "if part" lists a set of conditions in some logical combination. Once the "if part" of the rule is satisfied, the "then part" can be concluded or problem solving action taken. Expert systems with knowledge represented in rule form are called **rule-based systems**. The inference engine makes inferences by determining which rules are satisfied by facts, ordering the satisfied rules, and executing the rule with the highest priority.

   Expert data editing systems make so-called **intelligent imputations** based on a specified hierarchy of methods to be used in imputing an item. One item may use a deterministic approach followed by a hot-deck approach, while another item might require a model-based approach. Each item on the questionnaire would be resolved according to its own hierarchy of approaches, the next being automatically tried when the previous method has failed.

f. **Neural networks** - information processing paradigm based on the way the mammalian brain processes information. The neural network is composed of a large number of interconnected parallel processing elements tied together with weighted connections. These connection weights store the knowledge necessary to solve specific problems. The network is prepared for solving a problem by "training", i.e. the connection weights are iteratively adjusted based on examples or a verified set of input/output data. In data editing neural networks allow to create an edit system directly from the data and to develop the edits over time through periodic retraining.

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.

BATCH DATA REVIEW

A review of many questionnaires in one batch occurs after data entry. It generally results in a file of error messages. This file may be printed for use in preparing corrections. The data records may be split into two files. One file containing the "good" records and one containing data records with errors. The other file can be corrected using an interactive process.

CAI-COMPUTER-ASSISTED INTERVIEWING uses the computer during interviewing. Any contradictory data can be flagged by edit routines and the resultant data can be immediately adjusted by information from the respondent. An added benefit is that data capture (key-entry) is occurring at interview time. CAI assists the interview in the wording of questions and tailors succeeding questions based on previous responses. CAI has been mainly used in **Computer-Assisted Telephone Interviews (CATI) or Computer-Assisted Personal Interviewing (CAPI)**.

CASIC

The acronym “CASIC” stands for computer assisted survey information collection. This encompasses computer assisted data collection and data capture. CASIC may be more broadly defined to include the use of computer assisted, automated, or advanced computing methods for data editing and imputation, data analysis and tabulation, data dissemination, or other steps in the survey or census process.

CHECKING RULE

A logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct. In various connections other terms are used, e.g. edit rule.

CLASS ATTRIBUTE CHECK
Verifying whether the value of the common attributes (class attributes) of a logical unit or its components are identical.

CODE LIST
LIST OF CODE WORDS
List of all allowed (admissible) values of a data item.

CODE REDUNDANCY
When a character or group of characters in a code word can be partially or completely deduced from the remaining characters of the code word.

CODE SPACE
A set of all combinations of admissible values of a particular record of data. Cartesian product of the code word lists of individual data in a record.

CODE STRUCTURE VALIDATION
CODE STRUCTURE CHECK
Verifying whether the characters of the correct type (e.g., digits, letters) are at the correct positions of the code word.

COLD-DECK
A correction base for which the elements are given before correction starts and do not change during correction. An example would be using prior year's data. A modified cold-deck may adjust cold-deck values according to (possibly aggregate) current information (see also HOT-DECK).

COMPLETE SET OF CONFLICT RULES
A set of explicitly given conflict rules and of all implied conflict rules.

COMPLETE SET OF EDITS
The union of explicit edits and implied edits. Sufficient for the generation of feasible (acceptance) regions for imputation (that is if the imputations are to satisfy the edits).

COMPLETENESS CHECKING
... at **survey level** ensures that all survey data have been collected. A minimal completeness check compares the sample count to the questionnaire count to insure that all samples are accounted for, even if no data were collected.

... at **questionnaire level** insures that routing instructions have been followed. Questionnaires should be coded to specify whether the respondent was inaccessible or has refused, this information can be used in verification procedures.

COMPOSITION CHECK
Verifying whether the structure of a logical unit (e.g. - household) is consistent with the definition (e.g., at least one adult).

CONDITIONAL EDIT
An edit where the value of one field determines the editing relationship between other fields and possibly itself. 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.

CONFLICT RULE
REJECTION RULE
A logical condition or a restriction to the value of a data item or a data group which must not be met if the data is to be considered correct.

CONSISTENCY CHECK
Detecting whether the value of two or more data items are not in contradiction.

CONSISTENCY EDIT
A check for determinant relationships, such as parts adding to a total or harvested acres being less than or equal to planted acres.

CONSISTENCY ERROR
Occurrence of the values of two or more data items which do not satisfy some predefined relationship between those data items.

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.

CORRECTION BASE
A set of correct data or records from which date or records are retrieved for imputation in the (probably) erroneous data or records.

CORRECTION CYCLE
A cycle in which corrections (changes) are made to a record according to an existing edit/imputation strategy. If the record fails at the end of a cycle, it is sent through the edit/imputation

software subroutines until it passes or until a pre-determined number of cycles has been exceeded.

CORRECTION RULE

A rule for correcting certain types of errors. Its general form is as follows: "If errors are detected by the checks e_1, ..., e_k, make a correction in this way: ".

CREATIVE EDITING

A process whereby manual editors (i.e. those doing the manual review) invent editing procedures to avoid reviewing another error message from subsequent machine editing.

DATA CAPTURE

The process by which collected data are put in a machine-readable form. Elementary edit checks are often performed in sub-modules of the software that does data capture.

DATA CHECKING

Activity through which the correctness conditions of the data are verified. It also includes the specification of the type of the error or condition not met, and the qualification of the data and its division into the "error free" and "erroneous data". Data checking may be aimed at detecting error-free data or at detecting erroneous data.

DATA COLLECTION

The process of gathering data. Data maybe observed, measured, or collected by means of questioning, as in a survey or census response.

DATA CORRECTION
CORRECTION OF ERRORS IN DATA

Activity of checking data which was declared (is possibly) erroneous.

DATA EDITING

The activity aimed at detecting and correcting errors (logical inconsistencies) in data.

DATA IMPUTATION

Substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits.

DATA ITEM
DATA FIELD

The specific sub-components of a data record. For instance, in a population census, specific data items might be last name, first name, sex, and age.

DATA IMPUTATION

Substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits.

DATA REDUNDANCY

When the value of data items (fields) can be partially or completely deduced from the values of other data items (fields).

DATA REVIEW / DATA CHECKING

Activity through which the correctness conditions of the data are verified. It also includes the specification of the type of the error or condition not met, and the qualification of the data and its division into the "error-free" and "erroneous" data. Data checking may be aimed at detecting error-free data or at detecting erroneous data. Data review consists of both *error detection and data analysis*, and can be carried out in *manual* or *automated* mode.

Data review/error detection may occur at many levels:

a) within a questionnaire

Item level / editing of individual data - the lowest logical level of checking and correction during which the relationships among data items are not considered. *Validations* at this level are generally named "range checking". Example: age must be between 0 and 120. In more complex range checks, the range may vary by strata or some other identifier. Example: if strata = "large farm operation", then the number of acres must be greater than 500.

Questionnaire level / editing of individual records - a logical level of checking and correction during which the relationships among data items in one record/questionnaire are considered. Example 1. If married = 'Yes' then age must be greater than 14. Example 2.

Sum of field acres must equal total acres in farm.

Hierarchical - This level involves checking items in sub-questionnaires. Data relationships of this type are known as "hierarchical data" and include situations such as questions about an individual within a household. In this example, the common household information is on one questionnaire and each individual's information is on a separate questionnaire. Checks are made to ensure that the sum of the individual's data for an item does not exceed the total reported for the household.

b) across questionnaires / editing of logical units

A logical level of checking and correction during which the relationships among data in two or more records are considered, namely in a group of records that are logically coupled together. The across questionnaire edits involve calculating valid ranges for each item from the survey data distributions or from historic data for use in *outlier* detection. *Data analysis* routines that are usually run at summary time may easily be incorporated into data review at this level. In this way, summary level errors are detected early enough to be corrected during the usual error correction procedures. The "across questionnaire" checks should identify the specific questionnaire that contains the questionable data.

Across questionnaire level edits are generally grouped into two types: *statistical edits* and *macro edits.*

DATA VALIDATION

An activity aimed at verifying whether the value of a data item comes from the given (finite or infinite) set of acceptable values. For instance, a geographic code (field), say for a Canadian Province, may be checked against a table of acceptable values for the field.

DATA VALIDATION ACCORDING TO A LIST

Verifying whether the data value is in the list of acceptable values of this data item.

DEDUCTIVE IMPUTATION

An imputation rule defined by a logical reasoning, as opposed to a statistical rule.

DETECTION OF ERRORS IN DATA (ERROR DETECTION)

An activity aimed at detecting erroneous data. Usually predefined correctness criteria are used.

DETERMINISTIC CHECKING RULE

A checking rule which determines whether data items are incorrect with a probability of 1.

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, given specific values of other fields, when only one value of a field will cause the record to satisfy all of the edits. For instance, it might occur when the items that are supposed to add to a total do not add to the total. If only one item in the sum is imputed, then its value is uniquely determined by the values of the other items. This may be the first situation that is considered in the automated editing and imputation of survey data.

DONOR (imputation)

In hot-deck edit/imputation, a donor is chosen from the set of edit-passing records based on its similarity to the fields in the record being donated to (being imputed within). Values of fields (variables) in the donor are used to replace the corresponding contradictory or missing values in the edit-failing record that is receiving information. This type of replacement may or may not assure that the imputed record satisfies edits.

EDIT RULE SPECIFICATION
CHECK RULE SPECIFICATION

A set of check rules that should be applied in the given editing task.

EDITING BOUNDS

Bounds on the distribution of a measure used on survey data so that when raw, unedited data are outside the bounds the data is subject to review and possible correction (change).

EDITING EFFICIENCY

An efficient edit is good at identifying suspicious data. Edits that incorrectly flag large amounts of valid data are not very efficient. If such edits require an analyst to re-contact the respondent to verify the data, the analyst will have less time to perform other tasks, such as converting non-respondents. The hit rate may be used as an indicator of editing efficiency.

EDITING INDICATORS (editing flag/editing code)

A flag or code that is added during the edit process. The flag might indicate that, for example: a field in a record was targeted for change because it failed an edit, the field in the record was changed, or an override code was entered so that the edit system would no longer fail the record or the field in the record.

EDITING MATCH (matching fields/statistical match)

For hot-deck imputation, the fields in an edit-failing record that do not fail edits are matched against edit-passing records. Often the edit-passing record that is closest to the edit-failing record in terms of some metric is chosen as the donor record. If the edit-passing records are in random order, then the (possibly erroneous) assumption is that the donation is at random from a valid set of donors. This type of matching is sometimes referred to as statistical matching.

EDIT(ING) MATRIX

A matrix used in editing. In hot-deck imputation, the matrix contains information from edit-passing records that is used to donate information to (impute values in) the edit-failing record. Typically, there will be a variety of matrices that correspond to the different sets of variables that are matched on. The matrices are updated continuously as a file of records is processed and additional edit-passing records become available for updating the matrices.

EDITING MEASURE

A formal measure such as the Hidiroglu-Berthelot statistic that allows delineation of a targeted subset of records for manual follow-up. The statistic might be a measure such as size that allows the most important respondents to be manually reviewed.

EDITING OF INDIVIDUAL DATA

The lowest logical level of checking and correction during which the relationships among data items are not considered.

EDITING OF INDIVIDUAL RECORDS

Logical level of checking and correction during which the relationships among data items in one record are considered.

EDITING OF LOGICAL UNITS

A logical level of checking and correction during which the relationships among data in two or more records are considered, namely in a group of records that are logically coupled together.

EDITING PROCEDURE

The process of detecting and handling errors in data. It usually includes three phases:

- the definition of a consistent system of requirements,
- their verification on given data, and
- elimination or substitution of data which is in contradiction with the defined requirements.

EDITING RATIONALITY

An editing process that focuses on improving the incoming data quality and hence the overall quality of the survey data. This may include moving most of the editing as close as possible to the collection of data, limiting manual follow-up to those flagged records with the heaviest potential impact on estimates, and applying a total quality management approach to data editing.

ELECTRONIC QUESTIONNAIRE

Questions are in a software system that can be answered by an individual. Examples might be a software system that is on a laptop computer where respondents can answer questions directly into the laptop (possibly without knowledge on the part of interviewers of the details of the answers) or through queries on an Internet page.

ERROR DETECTING CHARACTER

Character added to the basic characters of the code word. Its relationship to the basic characters is specified previously. The relationship is specified so that typical errors in the code word transmitted break this relationship.

ERROR DETECTING CHARACTERS
CHECK-DIGIT

Code or given set of code words whose relationship to the set of valid codes is known, such that when a transcription error occurs, the relationship is violated and the error is detected with certainty or very high probability.

ERROR DETECTION

An activity aimed at detecting erroneous data, using predefined correctness criteria. The correctness criteria can be defined through various *checking rules*.

ERROR LOCALIZATION

The (automatic) identification of the fields to impute in an edit-failing record. In most cases, an optimization algorithm is used to determine the minimal set of fields to impute so that the final (corrected) record will not fail edits.

ERROR STATISTICS

A statistical report of errors found. It usually provides the error rate in individual data items and the occurrence of particular kinds of errors.

EVALUATION OF EDITING

There are several methods available to evaluate an edit system. One method is to compare the raw (unedited) data file with the edited file for each question. By ordering the changes by descending absolute magnitude, the cumulative impact of the editing changes to the total editing change or to the estimated item total can be displayed in graphs and tables. This technique can be used to identify questionnaire problems and respondent errors. When forms are reviewed before data capture or when general editing changes are noted on the questionnaires, evaluation studies can be carried out by selecting a sample of forms and analysing the effect of the editing procedures on individual data items. A widely used technique to evaluate new editing methods is to simulate the new process using a raw data file from the survey. By replacing values flagged according to the new method with the values from the tabulation file containing the data edited by the alternative editing process, it becomes possible to compare estimates from this newly edited file with estimates from the tabulation file.

EXPERT SYSTEM

Computer system that solves complex problems in a given field using knowledge and inference procedures, similar to a human with specialized knowledge in that field.

EXPLICIT EDIT

An edit explicitly written by a subject matter specialist. (Contrast explicit edits with implied or implicit edits.)

EXPLICITLY DEFINED CONFLICT RULE

A conflict rule which in defined by the people responsible for the correctness of the data.

FATAL ERRORS

Errors identified by fatal edits.

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.

FATAL EDIT

Identifies data errors with certainty. Examples are a geographic code for a Canadian province that does not exist in a table of acceptable geographic codes.

FELLEGI- FACTOR CHECK

Check of measuring units verifying whether the data has been given in correct units.

FELLEGI-HOLT SYSTEMS, TENETS, PRINCIPLES

In reference to assumptions and editing and imputation goals put forth by Fellegi and Holt in their 1976 *Journal of the American Statistical Association* paper. A key feature of the Fellegi-Holt model is that it shows that implied edits are needed to assure that a set of values in data fields that are not imputed always lead to final (imputed) records that satisfy all edits.

FIXED CONSTRAINT CHECK
RANGE CHECK

Verifying whether the data item value is in the previously specified interval.

FUNCTIONAL CHECK
ARITHMETIC EDIT
Verifying whether the given functions of two or more data items meet the given condition.

GENERATED (versions of) QUESTIONAIRES
There are software systems that assist in designing and electronically generating paper questionnaires.

GRAPHICAL EDITING
Using graphs to identify anomalies in data. While such graphical methods can employ paper, the more sophisticated use powerful interactive methods that interconnect groups of graphs automatically and retrieve detailed records for manual review and editing.

"HEADS-DOWN" DATA ENTRY
Data entry with no error detection occurring at the time of entry. Data entered in a heads down mode is often verified by re-keying the questionnaire and comparing the two-keyed copies of the same questionnaire. Data entered in a "heads-down" fashion may later be corrected in either a "batch" or an "interactive" data review process.

"HEADS-UP" DATA ENTRY
Data entry with a review at time of entry. Heads up data entry requires subject-matter knowledge by the individuals entering the data. Data entry is slower, but data review/adjustment is reduced since simple inconsistencies in responses are found earlier in the survey process. This mode is very effective when the interviewer or respondent enter data during the interview (CAI).

HIT RATE
The "success" rate of an edit; the proportion of error flags that the edit generates which point to true errors.

HOLT METHOD FOR AUTOCORRECTION
Automatic correction method in which the least possible number of data items is changed and the Fellegi-Holt model is used to determine acceptable sets of values or ranges for the items that are imputed. Sequential or simultaneous imputation via cold-deck or hot-deck method may be applied.

HOT-DECK
A correction base for which the elements are continuously updated during the data set check and correction. Typically edit-passing records from the current database are used in the correction database (see COLD-DECK).

HOT-DECK IMPUTATION
A method of imputation whereby values of variables for good records in the current (hot) survey file are used to impute for blank values of incomplete records (see COLD-DECK).

IMPLIED CONFLICT RULE
Conflict rule which can be deduced from the explicitly given conflict rules.

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

IMPUTATION
A procedure for entering a value for a specific data item where the response is missing or unusable.

IMPUTATION VARIANCE
A component of the total variance of the survey estimate introduced by the imputation procedure.

INLIER
An *inlier* is a data value that lies in the interior of a statistical distribution and is in error. Because inliers are difficult to distinguish from good data values they are sometimes difficult to find and correct. A simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius.

INPUT-EDITING
Editing that is performed as data is input. The editing may be part of a data entry system.

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.

INTERACTIVE DATA REVIEW/
INTERACTIVE EDITING/
ONLINE CORRECTION

Checking and correcting data in dialogue mode using video terminals. It can be applied during data entry or on data that are already in machine-readable form. The questionnaire is immediately reviewed after adjustments are made. The results are shown on a video terminal and the data editor is prompted to adjust the data or override the error flag. This process continues until the questionnaire is considered acceptable by the automated review process. Then results of the next questionnaire's review by the auto review processor are presented. A desirable feature of Interactive Data Editing Software is to only present questionnaires requiring adjustments.

LEAST SQUARES METHOD FOR AUTOCORRECTION
FEERD-HASTLY METHOD

An automatic correction method in which:

1. the least possible number of data items are changed
2. the changed record is the closest one (measured by the weighted sum of squares of deviations of the changed data) to the original (incorrect) record.

LINEAR EDITS

Edits arising from linear constraints. For example, if v_1, v_2, v_3 are variables and a, b, and c are real constants, the linear inequality edits are given by:

1. $a <= v_1 / v_2 <= b$ (This is two edits. Each can be converted to linear inequality.).
2. $a\ v_1 + b\ v_2 \# c$.
3. $v_1 + v_{1 =} v_{3.}$

LOGICAL CONDITION CHECK

Verifying whether the given logical condition is met. It is usually employed to check qualitative data.

LOGICAL LEVEL OF THE CHECKING RULE

The logical level of the data structure to which the given checking rule refers (individual data item, record, logical group of records, and the like).

MACRO-EDIT
SELECTIVE EDIT

Detection of individual errors by: 1) checks on aggregated data, or 2) checks applied to the whole body of records. The checks are typically based on the models, either graphical or numerical formula based, that determine the impact of specific fields in individual records on the aggregate estimates.

MANUAL CORRECTION

A human activity aimed at changing the values of data items deemed erroneous. The correction specified usually on the diagnostic list is entered into the data set by means of a program specially written for this purpose.

MANUAL DATA REVIEW

May occur prior to data entry. The data may be reviewed and prepared/corrected prior to key-entry. This procedure is more typically followed when heads-down data entry is used.

MICROEDITING

Finding errors by inspection of individual observations. Editing done at the record, or questionnaire level.

MINIMAL SET OF CONFLICT RULES

A minimal subset of the complete set of conflict rules expressing the same erroneous data combinations as the complete set of conflict rules.

MINIMAL SET OF FIELDS TO IMPUTE

The smallest set of fields requiring imputation that will guarantee that all edits are passed. See also *weighted minimal set.*

MONITORING OF EDITING

Analyzing the audit trail and evaluating the edits for efficiency.

MONTE-CARLO METHOD FOR AUTOCORRECTION

An automatic correction method in which the corrected data value is randomly chosen on the basis of a previously supplied probability distribution for this data item. The method employs computer algorithms for generating pseudo-random variables with the given probability distribution.

MULTI-LEVEL MODELING IMPUTATION

An imputation rule defined by a sequence of decisions each based on exclusive sets of observations.

MULTIVARIATE EDIT

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

NONLINEAR EDITS

Edits from non-linear constraints. For example, if v_1 and v_2 are variables and b are real constants, then nonlinear edits are:

1. v_1 v_2 # a.
2. v_1 # exp(v_2).
3. conditional edits.
4. Mahalanobis-distance edits with multivariate normal data.

The importance of nonlinear edits is that they occur often but are not amendable to theory in the determination of a minimal set. Some nonlinear edits, such as ratio edits, can be cast in a linear form.

NIM (*new imputation methodology*)

A generalization of the hot-deck that employs sophisticated matching methods to choose potential donors from a pool of edit-passing records that most closely resemble the edit-failing record being donated to. The method uses additional metrics for comparing numeric data and specific logic to assure that records satisfy edits that are not available with traditional hot-deck methods.

NORMAL FORM OF CONFLICT RULE

A conflict rule which is defined by the logical product of conditions on the values of individual data items in a record. For example, the conflict (branch - (101, 107,112); production 104, 180; efficiency 0.8) is a conflict in the normal form (CAN-EDIT).

ON-LINE CORRECTION

Correcting the values in erroneous data items of a previously checked data set on a video terminal.

OUTLIER

An *outlier* is a data value that lies in the tail of the statistical distribution of a set of data values. The intuition is that outliers in the distribution of uncorrected (raw) data are more likely to be incorrect. Examples are data values that lie in the tails of the distributions of ratios of two fields (ratio edits), weighted sums of fields (linear inequality edits), and Mahalanobis distributions (multivariate normal) or outlying points to point clouds of graphs.

OVEREDITING

Editing of data beyond a certain point after which as many errors are introduced as are corrected.

PROBABILISTIC CHECKING RULE

A checking rule causing, with some small probability, incorrect qualification of data (i.e., it may identify actually correct data as incorrect and identify incorrect data as correct).

PROBABILISTIC IMPUTATION

An imputation rule that is in part a function of a randomization process exogenous to the experimental observations.

QUALITATIVE DATA

Data describing the attributes or properties that an object possesses. The properties are categorized into classes that may be assigned numeric values. However, there is no significance to the data values themselves, they simply represent attributes of the object concerned.

QUALITY CONTROL

... of the **data collection process** assures that the underlying statistical assumptions of a survey are not violated, i.e. the meaning of the principal statistical measures and the assumptions which condition their use is maintained.

... in data review process measures the impact of data adjustment on the data.

QUANTITATIVE DATA

Data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data, e.g. meters, in the case of the height of a person. It makes sense to set boundary limits to such data, and it is also meaningful to apply arithmetic operations to the data.

QUERY EDIT

Points to suspicious data items that may be in error. An example could be a value that, compared

to historical data, seems suspiciously high. Contrast query edit to *fatal edit* where data item is known with certainty to be in error.

QUERY ERRORS
Errors identified by query edits.

RATIO EDIT
An edit in which the value of a ratio of two fields lies between specified bounds. The bounds must be determined through a priori analyses (possibly involving data sets in which truth data are available) or via exploratory data analysis methods.

REJECTION RULE / CONFLICT RULE
A logical condition or a restriction to the value of a data item or a data group which must not be met if the data is to be considered correct. In various connections other terms are used, e.g. Y-rule.

REPEATABILITY
The concept that survey procedures should be repeatable from survey to survey and from location to location; the same data processed twice should yield the same results. (Also called reproducibility.)

REPORT ON ERRORS
ERROR DIAGNOSTICS
A report which usually contains the record identification, violated conditions, items that are probably erroneous, etc.

SCORE FUNCTION
A numerical indicator used to prioritize micro data review in selective editing.

SELECTIVE EDITING
A procedure which targets only some of the micro data items or records for review by prioritizing the manual work and establishing appropriate and efficient process and edit boundaries.

SEQUENTIAL CORRECTION
SEQUENTIAL IMPUTATION
A correction where the items intended for the correction are corrected sequentially.

SIMULTANEOUS CORRECTION
SIMULTANEOUS IMPUTATION
A correction in which all the data in a record intended for correction is corrected at the same time (e.g., by using the record from a hot deck or cold deck).

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 and routes to be applied to the data.

STATISTICAL EDIT
A set of checks based on statistical analysis of respondent data, e.g., the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters. A statistical edit may incorporate cross-record checks, e.g., 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 also use historical data on a firm-by-firm basis in a time series modeling procedure.

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. Compare to *query* edit.

STOCHASTIC IMPUTATION
Probabilistic imputation.

STRUCTURE ERROR
The absence or presence of a data record not following the hierarchical order into which the data file is organized.

SUBJECT-BASED EDIT
Checks incorporating real-world structures which are neither statistical nor structural.
Example: wages paid/hours worked $ minimum wage.

SUBSTANTIAL EDIT
Edit rule proceeding from knowledge of the substance of the subject matter.

SUM CHECK

Verifying whether the sum of the values of the given data group equals the value of the corresponding data item which should represent their total.

SURVEY MANAGEMENT

The processes used to monitor, administer and control the survey. Survey Management includes preparing mailing labels and calling lists, making enumerator assignments and tracking the status of each questionnaire through the data editing processes (data collection, data capture and review/correction). Survey Management also includes monitoring the status and quality of the data editing processes.

SYSTEMATIC ERROR

Errors reported consistently over time and/or between responding units (generally undetectable by editing). A phenomenon caused either by the consistent misunderstanding of a question on the survey questionnaire during the collection of data or by consistent misinterpretation of certain answers in the course of coding. The systematic error does not lead necessarily to validity or consistency errors but always seriously compromises statistical results.

TOTAL QUALITY MANAGEMENT (TQM) is an optimisation and integration of all the functions and processes of an organisation in order to satisfy customer demands through a process of continuous improvement. Total Quality revolves not only producing a good product, but on improving the competitiveness, effectiveness and flexibility of the whole organisation in providing better products and services to customers. In data editing, the TQM approach aims to prevent errors rather than correct them, and to learn from past (editing) experiences in order to improve the data collection process. One element of the approach is to use the information on errors and error causes collected at the editing process to evaluate the performance of the survey questionnaire, and to feed back the findings for improvement in staff training, system design and form design. The other element is the principle of getting the data right the first time.

TRANSPOSITION CHECK

Detecting whether the data has been recorded at the correct position (i.e., whether the data in various character positions have been transposed).

VALIDATION EDITS

Edit 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 that entries are consistent with each other.

VALIDITY ERROR

An occurrence of the value of a data item which is not an element of the set of permissible codes or values assigned to that data item.

WEIGHTED MINIMAL SET

A minimal set in which fields are weighted according to reliability in generating imputations. If fields (variables) are given weights, then it is the set of fields that give the minimum weight. There may be two of more sets that give the same minimum.

WINSORISATION

An imputation rule limiting the influence of the largest and smallest observations in the available data.