vrije Universiteit amsterdam

BMI PAPER

AUTOMATIC SURVEY DATA EDITING

BASED ON THE GENERALIZED FELLEGI-HOLT PARADIGM

Nicolaas Nobel

August 2012

Supervised by Zoltán Szlávik

Abstract: Surveys on business, social or census studies contain human made errors, which can be problematic when analyzing and finally reporting data. This paper sets out an overview of different procedures for the detection and correction of these errors based on the (generalized) Fellegi-Holt paradigm, as well as some practical aspects of current trends. Also a detailed overview on the types of errors and the prevention is provided.

Keywords: Error Localization; Imputation; Generalized Fellegi & Holt Paradigm; Edits; Set-covering problem; Branch & Bounds; Hot Deck Imputation;

PREFACE

This paper is part of the master Business Analytics at the VU University of Amsterdam. The goal of the paper is to do research on a particular subject by choice that is of practical business importance and contains elements of mathematics and/or informatics. This paper contains several mathematical optimization problems that can be solved using algorithms in the field of operations research. Anyone that is interested in the field of (survey) data editing and has a basic mathematics and/or informatics background should be able to understand the content of this paper.

I would like to thank Zoltán Szlávik (Department of Computer Science at VU University) for his support and supervision, Willem Sluis (Statistics Netherlands) informing me about the field of Automatic Data Editing and Wim Nobel (Logica) for some useful comments concerning mathematical formulas and notation.

In order to write this paper I have studied the book "Handbook of Statistical Data Editing and Imputation" by Ton de Waal et al. (1) extensively.

TABLE OF CONTENT

INTRODUCTION

In today's information society, the survival of most companies depend on the knowledge of their business, customers and competitors due to the decisions that can be made related to product innovation, employees satisfaction, customer service and company reputations. Administrative registers containing detailed individual data records on a population, like businesses or persons, are used, but cannot always deliver the field-specific information that is required for a particular study. Conducting surveys is a method of retrieving information directly from people and businesses and is widely used by statistical firms and businesses due to its flexibility to ask specific questions. It is used in many fields, i.e. marketing research, psychology, health professionals and sociology (2). The result is that questionnaires are sent out on a massive scale about numerous subjects from door-to-door, online, by phone and other. Subsequently the raw data, i.e. the table containing questionnaire answers, is returned and can fill up millions of records in a table, which are then analysed, aggregated and published.

Depending on the quality of the questionnaire, the survey process and the cooperation of respondents, data is collected, hopefully with a high quality, in some cases followed up by the data being entered in a computer system. These are mainly manual operations and human mistakes are quickly made. When respondents fill in a questionnaire, they can forget, misread, ignore or misinterpret questions which lead to incomplete or erroneous answers. In case of paper questionnaires mistakes are also made by employees or computers while copying data from an original questionnaire into a computer or database. Leaving these errors in the data can result in unwanted and wrong information and analysing incomplete questionnaires can even result in technical problems (3). Therefore the erroneous questionnaires have to be processed before being analysed.

There are two manual methods for processing erroneous questionnaires. The first one is simply by removing all incomplete or erroneous questionnaires. With a high number of respondents this is a fast way to achieve a 'clean' dataset. The second method is by re-contacting respondents asking for the correct answers.

However, both of these methods have some practical and theoretical weaknesses. Firstly, by deleting questionnaires containing errors, information is lost and the group of deleted questionnaires can be underrepresented. Secondly, larger questionnaires with more questions have a greater chance of containing faulty or incomplete answers. This way just a small subset of the total number of questionnaires can be used. This is even less fortunate in case of paid questionnaires. Correcting data by re-contacting respondents is a time-consuming and costly task and cannot even be performed when there is no contact information available.

A relatively fast and economical method to obtain a 'clean' dataset is by means of automatic error detection and correction procedures and programs, better known as automatic data editing and imputation. The increasing computation power, together with better and more efficient algorithms causes businesses and statistics firms to invest large shares of their total budgets on these automated systems, it has been estimated that National Statistics Institutes (NSIs) spend approximately 40% of their resources on editing and imputing data (1). The United Nations consider statistical data editing to be such an important topic

that they organize a so-called work session on statistical data editing every 18 months (1). Automatic data editing does not mean that wrong answers are replaced by real values. In fact, it is impossible to know what the real answers are without knowledge of the respondent. Furthermore it is not always certain which answers are incorrect. Automatic data editing means that the most probable wrong answers are corrected with a consistent value. Rules like "a male cannot be pregnant", but also "no negative age" define which answers are inconsistent and supposedly wrong. These rules are called edits.

Automatic Data Editing and Imputation consists of two steps: the first one is the process of identifying supposedly wrong answers, referred to as the error localization problem; the second step is imputation, which deals with the actual replacing or filling of answers with a consistent value.

This paper sets out an overview on the problem of random errors in surveys, along with two procedures based on the (generalized) Fellegi & Holt paradigm. The first is the original proposed procedure by Fellegi & Holt in 1967 (4) which involves a set-covering problem. The second describes a relatively recent branch and bound algorithm developed at Statistics Netherlands by de Waal and others (1). The focus of this paper is on the error localization problem, which is by far the most complex of the two steps and has many mathematical as well as computational challenges.

1. THE SURVEY PROCESS

Conducting large surveys such as economic, social or production surveys involve more than just preparing a questionnaire. Sometimes millions of people need to be surveyed, within a tight time schedule with a limited budget. The survey process describes the different phases for acquiring information from individuals through sampling from a population with the intention of making statistical inference. There are many different types of surveys, but this paper concentrates on the bigger ones like (inter-)national companies, household population or demographic data, that actually need automatic editing. The purpose of this chapter is to introduce the multiple phases in which a survey is created, rolled out and analysed, to give a general background on the topic, with as main focus the errors and reasons for non-response.

A general process for accumulating information by business surveys can be divided into five phases (5):

1. Setting the survey objectives
2. Preparing the survey operations
3. Sampling, data collection and data entry
4. Processing and analysis
5. Publication and dissemination

The same process could also be used for surveys in general (1).

1.1 SETTING SURVEY OBJECTIVES

Clear survey objectives can result in clear, unambiguous and relevant results, where a vague plan results in vague outcomes. This phase forms the basis for the next ones and needs to prevent burden and non-response by clearly understanding client needs. Therefore an exploratory stage is set in which a clear understanding of client needs is developed. Also it is important to find out whether the results cannot already be found in other sources. This should be considered in a broad sense since data that had another purpose might still be useful for the current study as well. Also the usefulness is tested of administrative registers, which are systematic collections of data that can be related to individual entities. Sometimes these registers already contain (parts of) the required information. The group of potential respondents called the target group is identified and their willingness to participate is tested. These can be normal household citizens, businesses or specific client or employee groups. Practical issues are identified like the available budget, priority definitions and legal aspects. After the exploratory stage, many surveys are placed in a general framework of standards and norms, so that they can be compared to similar surveys. For instance, national statistic firms acquire data from their country in the same framework as other countries so that a global overview can be created. After that, the intended statistical output is specified, called the target population based on the surveys objectives. The respondents that are closest matching the purpose of the study are selected. The last step is to construct a table outline which contains the output variables in order to fulfil the objectives. Their definitions and

terminology are formed in such a way that they are unambiguous, clear, uniform and cope with the willingness of respondents to answer. The table outline is focused on the objectives and provides the benchmark for the forthcoming stages in the survey design and it supplies the client with a clear picture of the type, detail and accuracy of the information of the survey.

1.2 PREPARING THE SURVEY OPERATIONS

Prevention of non-response and mistakes by respondents require intensive effort in this phase. Where the previous stage concentrates on what needs to be investigated, the next stages are focussed on process level, so how the survey is carried out. This stage deals with the tools that need to be prepared. The first tool is called the sample strategy, which is required because of practical challenges with respect to timeliness and budget limitations, most likely enriched with data from an additional administrative register. The second is the questionnaire design, where the outcome variables need to coincide with the list of questions which are sent to respondents to cope with the surveys objectives. A sampling strategy is a combination of a design and an estimator. The sampling design is a set of specifications which include the target population, the sampling units, and the probabilities attached to the possible samples. The estimator is a mathematical function of which the estimate for a particular parameter is computed.

Where the table outline is meant for the client, the questionnaire is focussed on the respondent. Typical reasons why respondents do not answer a specific question is because they forget, misread, ignore or misinterpret questions. When a bad questionnaire is the cause the errors are called systematic errors, i.e. the kind of errors where the source is (easily) explained. Pre-testing the questionnaire gives the possibility to detect the major bugs and problems and to test it as a whole. Secondly, to cope with the respondents needs, much taught needs to be put into the general survey conditions, like data communication conditions, such as the data transfer mode (self-administered or interviewer-assisted), the data transfer medium (face-to-face, telephone, online, email), the number of available interviewers or contact persons and disclosure control. Interviewer-assisted surveying has the preference, since trained interviewers are more familiar with the survey objectives than respondents and can reduce the number of erroneous given answers. Digital data transfer media also have advantages. The first is that respondents can be directly informed about the most common mistakes while answering questions and that inconsistent possible answers for further questions are blocked. The second is that the data is directly available on a computer system. Furthermore, the questions need to be clear and user-friendly, cope with the knowledge of its respondents and are as short as possible. Thirdly, some output variables require a difficult calculation or advanced logics, in those cases it is better to construct multiple (easy) questions that coincide with the output variables.

1.3 SAMPLING, DATA COLLECTION AND DATA ENTRY

During this phase the sample is drawn, the questionnaires are sent out and (partly) returned by respondents. Furthermore the data is collected and entered in a data file using the tools developed earlier. This is also the phase were the success of the design of the questionnaire is largely tested, hoping that respondents will answer correctly. Besides preventing human-made errors in the questionnaire, it is important to get answers back in the first place. A high response rate requires a well-managed team of professionals informing, guiding, explaining and reminding respondents with their questionnaires firstly, and secondly remain a suitable time planning to receive the majority of the questionnaires back on time. Also a team of data enterers need to be trained, because mistakes are also made during entering. This phase deals mainly with manual input, especially when questionnaires are sent out on paper.

1.4 PROCESSING AND ANALYSIS

The topic of this paper is mainly embedded in this phase. Where in previous phases errors occur along the way, this phase attempts to identify and correct to generate an allowable data quality level. At this point the data has been collected and entered in a table on a computer. Each questionnaire is now translated to one record in a table with the table outline as format, and every answered question becomes a variable. In the rest of the paper, whenever the word 'record' is used, this means the row in the table which represents the set of answers from the questionnaire already translated to the table outline. The word 'variable' means a specific column of that table. This (semi-) raw data first needs to be processed in such a way that it can be used for analysis. There are two types of errors, the first are systematic errors, where the reason is known and many respondents have made the same mistake. These can be corrected simply by automatic logical replacement techniques by field experts. For example, if a human length in meters is larger than 100, then it needs to be replace by the original value divided by 100 assuming that in this case the respondent filled in his or her length in centimetres. The second types are random errors, where the cause is not known. These are much harder to track and the remaining chapters of this paper concentrate on these kind of errors. Other mentionable processing steps include the adjustments for seasonal change and the weighting and reweighting in case the target population is not completely presented by the sample population. More on weighting, see (6). Nevertheless most of the time is spend on data editing and imputation. Luckily it is not required to have an errorless dataset of questionnaire data to obtain reliable publication figures (7), therefore a large amount of time and money is spent on detecting and correcting only influential errors, which do have an impact on the results. Furthermore, specifying too many edits can even result in unwanted results by means of over-editing (8). There are multiple systems that perform these processing and analysis steps like GEIS, Banff, SPEER, AGGIES, CherryPi, SLICE, SCIA and DISCRETE (1).

1.5 PUBLICATION AND DATA DISSEMINATION.

The final phase starts with a complete and full table containing the processed and analysed data on its lowest aggregation level. The quality of the data is on a pre-defined sufficient level and is ready to be aggregated such that the surveys objectives are satisfied and that as many clients can be served, like government al institutions, businesses, research institutions, foreign users like the EU and UN, students, normal citizens and national statistics firms. Their needs can differ greatly, focusing on for instance census or business data in large or small amounts. For that reason the information is published in several ways, as well for the occasional client, who wants to find some specific numbers and more extensive clients, who need large amounts of data. For especially the latter it is therefore required that data is protected and sometimes anonymized. The data can be presented by numerous channels including magazines, websites and servers.

During all phases in the survey process human errors are made. Having clear objectives help to form a strong basis and decreases potential errors in later phases. Constructing and testing a user-friendly questionnaire decreases especially systematic errors and increases the chance that respondents are well-willing to cooperate. During the sampling, collecting and entering phase the bulk of the work is done manually and it is there were human mistakes occur at a large scale. Training staff, supporting respondents and keeping a functioning planning assists to prevent non-response. The processing and analysis phase contain various techniques to convert raw data into an effective warehouse, mainly focussed on data editing and imputation and last, during the publication and data dissemination phase the information is brought to as many as possible clients.

2. ERROR LOCALIZATION PROBLEM AND HOT DECK IMPUTATION

There is a substantial amount of (human) errors that occur during all phases of the survey process. A large number of these errors cannot easily be detected, since a given answer can be valid, but not present the true value. A considerable amount of errors are easier to detect, the ones that do not pass predefined rules, called edits. For instance, a negative human weight can be addressed as an error. Sometimes it is not directly possible to identify a variable as erroneous, but a detection of at least one error in a combination of variables is. For instance, a record describing a pregnant man will not pass the edit, either because it is a pregnant woman, a non-pregnant male or a non-pregnant woman. This chapter describes the Fellegi-Holt paradigm along with a mathematical optimization model. This requires a brief description on edits.

2.1 GENERALIZED FELLEGI-HOLT PARADIGM

A natural assumption is that on average a respondent makes as few mistakes as possible. In the example of the pregnant man one would intuitively want to say that or the gender or the pregnancy indicator is false, but not both. In case the respondent is also someone's wife, than it is two to one that the respondent is a pregnant woman.

The Fellegi-Holt paradigm uses this natural assumption and for that reason says that data should be made to satisfy all edits by changing the fewest possible number of variables. Based on the assumption the paradigm actually defines a way of handling data, to stay as close as possible to the given values or the raw data. One could argue that some answers hardly ever are given falsely, while others might have a higher chance. For that reason a generalized paradigm is adopted which not only stated that the number of variables is decisive, but also the reliability of the variables. Reliability weights are added (by field experts) and are a measure for the confidence level of the values of the variable. The higher the reliability weight of a certain variable, the more reliable its values are considered to be.

2.2 EDITS

Before describing the error localization problem based on the generalized Fellegi-Holt paradigm in terms of a mathematical optimization model, a basic understanding is required on a structural way of defining the allowable domain of variables, called edits. Since it is easier to describe what is allowed than what is not (9) it is therefore that modern procedures specify an edit e_i as the allowable domain or range of values that a certain variable or a combination of variables can contain:

$$e_i: x \in S_x$$

Here S_x is the set of allowable values of x and x on itself can be one or multiple variables. If x is not from S_x then the edit is violated and the corresponding record can be identified as erroneous. A whole list of edits can be constructed in order to validate one record. All errors, including simple validation and domain errors concerning a single variable, as well as advanced referential errors concerning multiple variables can be described in this form.

To describe the possible values for the variable 'gender' the following edit e_1 can be defined:

$$e_1: \text{gender} \in S_{\text{gender}} \text{ where } S_{\text{gender}} = \{\text{Male, Female}\}$$

The next shorter notation is used in the remaining part of this paper for simplicity reasons:

$$e_1: \text{gender} \in \{\text{Male, Female}\}$$

This edit is an example of a hard categorical domain edit. Hard, because there are definitely no other values for gender, categorical, since the possible values do not present numerical values and domain, since the range or domain of a single variable is described. The next example concerns a soft numerical range edit, describing the allowable domain for a human age:

$$e_2: \text{age} \in \{x \mid 0 \leq x \leq 125\}$$

In reality, it is not likely that someone becomes older than 125, but it is possible. Furthermore the range is defined by means of numerical values. Hard edits define strict rules, where soft edits define plausibility. Until now the examples concern univariate edits, which means that the defined area or domain only concerns one type of variable. The multivariate edits define certain rules between multiple variables. The next edit does not allow pregnant males:

$$e_3: \text{gender, pregnancy\_ind} \in \{x, y \mid x = \text{"Male"} \quad y = \text{"No"}\} \cup \{x, y \mid x = \text{"Female"} \quad y = \text{"No"}\} \cup \{x, y \mid x = \text{"Female"} \quad y = \text{"Yes"}\}$$

The -symbol refers to an AND-sign (intersection) and the $\vee$ to an OR-sign (union) within a set. The same holds for $\cap$(intersection) and $\cup$(union) between sets.

Another example of a multivariate edit is the so called balance edit, for instance:

$$e_4: \text{profit, turnover, totalcosts} \in \{x, y, z \mid x = y - z\}$$

One can choose to add more variables to edits stating that any value of that variable is allowed. This would not change the edit and therefore states that such a variable is not involved. In other words, if the domain of all possible values D_j for variable v_j is equal to the allowable values A_j^i for variable v_j in edit e^i then v_j is not involved in e^i.

Like described later, the used algorithms that work with categorical edits/variables and the ones involving continuous edits/variables are different and for that reason a distinction is made explicitly for most models.

2.3 MATHEMATICAL OPTIMIZATION MODEL

The generalized Fellegi-Holt paradigm says that data should be made to satisfy all edits by changing the fewest possible number of weighted variables. This can be seen as an optimization problem where the total number of changed variables needs to be minimized with the edits as constraints. De Waal (1) uses the next model for the error localization problem on a mixture of categorical and continuous variables:

$$\min \sum_{i=1}^{m} w_i^c \delta(v_i^0, v_i) + \sum_{j=1}^{n} w_j^r \delta(x_j^0, x_j)$$

Here (v_1,. . . , v_m, x_1,. . . , x_n) is a record with v_i the m categorical and x_j the n numerical variables. w_i and w_j are the so-called reliability weights of respectively variable i and j. The classification of these variables is done by field experts and is not dealt with in this paper. When the original value y^0 is changed, so when $y^0 \neq y$, then $\delta(y^0, y) = 1$ and if $y^0 = y$ then $\delta(y^0, y) = 0$. In this model a record that is barely changed has preference over an extensively changed record, when both comply with the edits.

Of course this model can give several optimal solutions. In that case one can choose the solution that fits best with the marginal and joint distribution of the (imputed) variables. Fellegi and Holt describe a similar model, only in their case all reliability weights are set to one, and only a record containing categorical variables is optimized.

2.4 IMPUTATION

The scientific area of missing data problems and even just the field of imputation methods is too big to be completely covered by this paper. This paper concentrates on Hot Deck Imputation (10) (11), which is particularly suitable for survey data. For more on imputation, consider reading (1) (12)

Hot Deck Imputation (10) uses the values from a similar record, called the donor, to replace the missing or erroneous values from a recipient record from the same dataset. The punched cards that were used extensively in the past still refer to the name "Hot Deck", since the donor and the recipient both came out from the same hot deck after creation. One speaks of a similar record when all fields but the erroneous or missing ones coincide (mostly) from the donor and recipient record. Donor records are only selected when they are consistent with all edits. For this reason recipient records have the same property after the imputation and become consistent as well.

A rather computationally effective technique is to find the last original record (donor) that coincides with the recipient and impute its values. This is called sequential imputation. In case of random imputation a list of all possible recipients is kept and randomly one is chosen to donate its values. The Nearest Neighbour Hot Deck Imputation technique can also be used for numerical variables. No categories are required and the donor and recipients values can differ slightly. A good donor is found by minimizing the total weighted distance between the two records. If there are multiple optimal neighbours, than one is randomly selected (11).

4. PROCEDURES

This chapter involves the procedures developed to solve the error localization problem described in chapter three. These procedures all depend on the generation of implied edits, based upon Fourier-Motzkin elimination (13), which is explained first. Since the total number of algorithms is too large to be described in this paper, two procedures are worked out. The original proposed procedure by Fellegi and Holt is described first involving a set-covering problem, followed up by the key elements of several improving algorithms. After that a modern branch and bound algorithm developed at Statistics Netherlands is described.

4.1 FOURIER MOTZKIN ELIMINATION

Edits are constructed by field-experts in a logical order. Based on their knowledge they construct sometimes large and difficult edits that needs to be broken up in smaller pieces in order to convert all edits to their normal form. These logical edits (in case of categorical data) and arithmetical edits (continuous data), are explicitly defined, but from combinations of edits new implied edits can be derived using Fourier–Motzkin elimination (FME) (13). FME is a mathematical algorithm for eliminating variables from a system of linear (in)equalities. Elimination of a variable x in this sense means that the remaining variables of a system of linear (in)equalities without x still needs to have the same solution as in the original system. For instance, the next pair of edits stating that $x_1 \leq x_2$ and $x_2 \leq x_3$ also implies that $x_1 \leq x_3$. The variable x_2, which is referred to the generating field or variable by Fellegi and Holt, has been eliminated and an implied edit has been derived which still describes the relation between x_1 and x_3. In mathematics this is also referred to as projection.

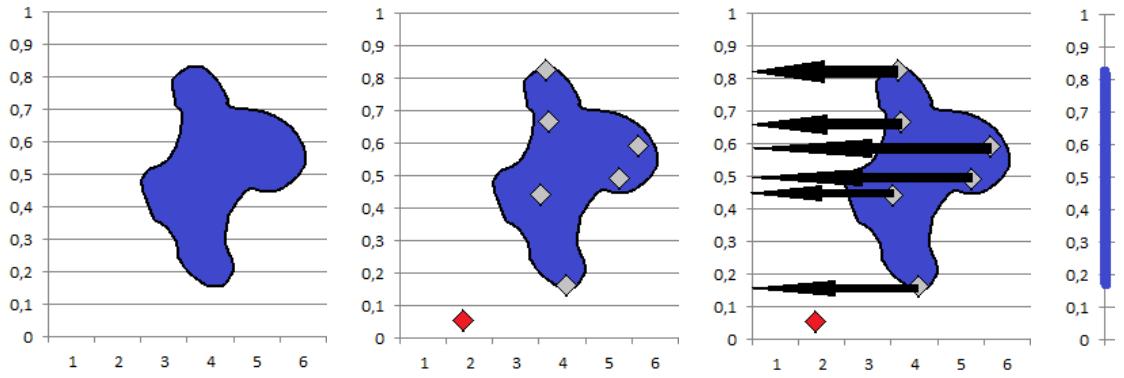

Figure 1: Elimination of a variable

In this graph the blue area represents an enclosed range defined by edit rules for two continuous variables x and y on the x and y-axis. If x is to be eliminated, only a vertical domain will remain. This means that although x is eliminated, y still needs to fulfil the newly implied edit $0{,}2 \leq y \leq 0{,}85$.

The projection can be applied to categorical edits as well. One variable is chosen to be eliminated and two edits involving that variable are combined to form an implied edit. Fellegi

and Holt were the first to do this using a lemma, involving a generating field g that is to be eliminated and contributing edits ($\in e_g$) that involve g. In their paper they describe edits as scenarios that cannot exist and refer to the normal form of edits.

Normal Form of edits (4)*:*

$$e_i = \bigcap_{j=1}^{N} A_j^i = F$$

Where A_j^k are the allowable values for field j in edit i.

Lemma 4.1 generates an implied edit e^* with generating field g and contributing edits e^i.

Lemma 4.1

$$A_j^* = \bigcap_{i,e^i \in e_g} A_j^i \quad j \in (1, \dots, m), j \neq g$$

$$A_g^* = \bigcup_{i,e^i \in e_g} A_j^i$$

Where A_j^i are the allowable values for variable j in edit i. If A_j^* exists for every $j \in \{1,\dots,m\}$ then e^* is an implied edit. If A_g^* is the total set of possible values for the generating field g, then e^* is also an essentially new implied edit (see (4) for the proof).

4.2 EXAMPLE OF AN ESSENTIALLY NEW IMPLIED EDIT

The next example is provided in order to demonstrate the generation of an essentially new implied edit using lemma 4.1 and two contributing edits. The edits are described in the normal form, taking the logical inverse modifies the edits to the form of chapter two. Edit e_5 describes the relation between the variables Adulthood (Adult) and Marital Status (MarStat) saying that non-adults cannot be married. Edit e_6 defines the relation between Marital Status and Relation to head (RelHead), where the combination not married and spouse are inconsistent.

$$e_5: (\text{Adult}, \text{MarStat}) \in \{(x, y) \mid x = \text{No} \wedge y = \text{Married}\} = F$$

$$e_6: (\text{MarStat}, \text{RelHead}) \in \{(y, z) \mid y \neq \text{Married} \wedge z = \text{Spouse}\} = F$$

In this example the variable Marital Status is used as the generating field, the field to be eliminated. Note that Relation to Head is not mentioned in e_5 and can contain any value. The same holds for the variable Adult in e_6.

Then

$$A_{\text{Adult}}^* = \bigcap_{k,e^k \in e_g} A_{\text{Adult}}^k = A_{\text{Adult}}^5 \cap A_{\text{Adult}}^6 = \{No\} \cap \{Any\ value\} = \{No\}$$

$$A^{*}_{\text{RelHead}} = \bigcap_{k,e^{k} \in e_{\text{RelHead}}} A^{k}_{\text{RelHead}} = A^{5}_{\text{RelHead}} \cap A^{6}_{\text{RelHead}} = \{Any\ value\} \cap \{Spouse\} = \{Spouse\}$$

For the generating field MarStat:

$$A^{*}_{\text{MarStat}} = \bigcup_{k,e^{k} \in e_{\text{MarStat}}} A^{k}_{j} = A^{5}_{\text{MarStat}} \cap A^{6}_{\text{MarStat}} = \{Married\} \cap \{Not\ Married\} = \{Any\ value\}$$

This generates the implied edit:

$$e^{*}: \bigcap_{j=1}^{N} A^{*}_{j} = (\text{Adult, MarStat, RelHead}) \in \{(x, y, z) \mid (x = \text{No} \wedge z = \text{Spouse}) \wedge y \in \{Any\ value\}\} = F$$

Note that Marital Status is not mentioned. Since it can contain any value it is redundant and e* is an essentially new implied edit.

Together with the explicit edits, implied edits are very useful during the algorithms to solve the error localization problem for two reasons:

1. A set of edits is consistent if and only if the set of edits after elimination of a variable is consistent. This gives an opportunity to check whether the set of specified explicit edits defined by field-experts is consistent in first sense. The list of edits after repeated use of FME needs to remain consistent. If this is not the case, then the original list of edits was not consistent either.
2. Essentially new implied edits can be generated which help to solve the error localization problem. When it is not clear which variable fails the most explicit edits, then new implied edits can give that insight, e.g. a child being a spouse and also being married.

The dual variant of FME is when instead of eliminating variables, edits are eliminated. For more details on FME, see (13).

4.3 SET COVERING PROBLEM

Fellegi and Holt describe a method for solving the error localization problem automatically by the generation of essentially new implied edits as described above. These essentially new implied edits together with the original explicit edits, together called the complete set of edits, are a necessity to translate the error localization problem into a set-covering problem (1). Fellegi and Holt were the first to use a repeated application of FME for categorical variables until no more essentially new edits can be generated. Once the complete set of edits is generated, it then can be re-used for every record separately.

A set-covering problem describes a problem in which the smallest sub-family of subsets is to be found in order to cover a total set of elements, called the universe.

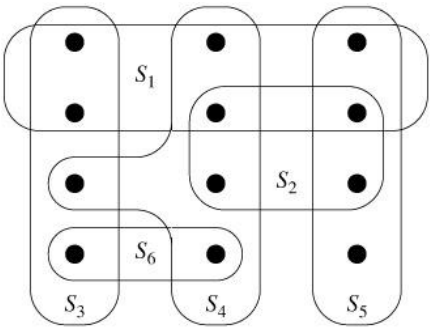

Figure 2: Graphical display of a set covering problem with subsets S_1 – S_5 that cover the universe of black dots. The sub-family {S_3, S_4, S_5} is the smallest sub-family to cover the universe (14)

The error localization problem involves a number of failed edits and a number of variables concerning one record. The set of all failed edits is in this problem the universe. Edits that all involve a particular variable together form a subset. Now the paradigm requires to find a minimal number of variables to be altered. This problem can be translated to finding a minimal sub-family of subsets, that covers the set of all failed edits.

For clarification the next failed edit matrix (4) contains 4 failed edits and 6 variables by a certain record. The (combination of) variables that make the edit fail are given a 'one' in the matrix for that edit, the others a 'zero'.

	$Variables$					
$Edits$	$v1$	$v2$	$v3$	$v4$	$v5$	$v6$
e_2	1	0	1	0	0	0
e_3	0	0	1	0	1	1
e_4	0	1	0	0	0	0
e_5	0	1	0	1	0	0

Figure 3: Example of a failed edit matrix

In this figure edit e_2 is not satisfied by the combination of variable v1 and v3. The solution to the set-covering problem is to find a minimum number of variables so that together they cover at least one failed variable per edit. These variables can be given new imputed values so that the record is valid. At this point it is also clear that implied edits play a crucial role, since they show hidden relations between variables. Since the number of variables and edits can be extremely high, solving this problem is rather an impossible task. Fellegi and Holt therefore propose the next procedure on the failed edit matrix:

1. Variables with an error in a single variable, are by definition part of the minimal set. In some cases this set of variables already covers all failed edits, which means that the minimal set is found.
2. If not, then the edit with the fewest number of variables is identified (select edit arbitrary if there is a tie).
3. For each variable in that edit a modified failed edit matrix is generated. The edits that are involved with the selected variable can be eliminated from the matrix, because

one of its involving variables will be imputed later. Repeat steps 3 and 4 until the first modified failed edit matrix vanishes totally.

4.4 EXAMPLE OF SET COVERING PROBLEM

1. Edit e_4 contains an error only because of v2. This variable contains an error, which means that if it is imputed edit e_5 will also be satisfied.

	$Variables$					
$Edits$	$v1$	$v2$	$v3$	$v4$	$v5$	$v6$
e_2	1	0	1	0	0	0
e_3	0	0	1	0	1	1

Erroneous variables:
V2

2. The matrix has not been vanished, so the edit containing the fewest variables is selected, being edit e_2.
3. The corresponding matrices for v1 and v3 are as follows:

	$Variables$					
$Edits$	$v1$	$v2$	$v3$	$v4$	$v5$	$v6$
e_3	0	0	1	0	1	1

Erroneous variables:
V2, V1 (together with v3, v5 or v6)

	$Variables$					
$Edits$	$v1$	$v2$	$v3$	$v4$	$v5$	$v6$
$Empy$	−	−	−	−	−	−

Erroneous variables:
V2, V3

In this example the minimal number of variables required to be imputed in order to satisfy all edits is 2, which are v2 and v3.

After the minimal set of variables is found, Hot Deck sequential and random imputation is used to impute new valid values to ascertain that the new values comply firstly, and secondly attempts to keep the original marginal and total distribution of the variables.

Fellegi and Holt also mention that not every record is suitable for automatic editing. Every record at first is checked whether it complies with all edits. If this is the case then this record is skipped and the next is selected. If the record contains too many errors, based upon a pre-defined maximum number of errors, then the record is not qualified for automatic editing and another technique needs to be used. If the record contains fewer errors than the pre-defined number, it is edited automatically.

4.4 IMPROVEMENTS OF THE SET-COVERING PROBLEM

Advantages of this system are that the number of edits is flexible and that it gives a guarantee for the optimal solution. A major drawback of this system is that the type of edits

that can be treated is limited to categorical edits. Another reason is that when the number of records, variables and edits increase, the computation time explodes. Especially the generation of essentially new implied edits is responsible for this steep ascent. For that reason the last couple of decades many improvements of this procedure have been developed. Garfinkel et al. (15) provided improvements to the edit generation process to the original FH-method and also developed a cutting plane algorithm which increased speed by not solving one large, but multiple small set-covering problems in sequence. Winkler (16) and Winkler and Chen (17) continued working on the edit generation procedure to increase speed even more. The main idea was that not all essentially new implied edits need to be considered, but just the non-redundant ones, which means that the complete set of edits does not contain edits that are dominated by others. An edit e_i dominates e_j if e_j is a subset of e_i. In fact Winkler discovered that the redundant edits are not only required for the complete set of edits, but neither for the generation process of the essentially new edits. Both these findings improved the generating speed extensively. Sande (18) presented an algorithm that also copes with linear numerical edits and Schopiu-Kratina and Kovar (19) and Filion and Schopiu-Kratina (20) continued working on that algorithm. Winkler and Draper (21) described a method for another important group of edits, the so called ratio and balance edits.

Another new technique is described by Bankier (22) and Bankier et al. (23) where error detection and correction occur in a single step, as well as Quere and De Waal (24), De Waal (25) and Daalmans (26) using a branch and bound algorithm, which is described next.

4.5 A BRANCH AND BOUND ALGORITHM

In 2002 De Waal (25) proposed an algorithm by means of a binary tree to solve the error localization problem. For every record a tree is used and the leaves of the tree contain a minimal list of variables that needs to be imputed. The size of the tree is determined by the number of variables and can theoretically contain $2^N + 1$ nodes in case of N variables. This still could (and usually is) a very large amount of nodes to process for modern day computers. Fortunately some branches of the tree are preliminary pruned, either because that path cannot result in a satisfied solution anymore, or because another branch already has a better solution. In literature this is mostly referred to as a branch and bound algorithm (27). Another nice feature of the described algorithm is that it also allows a combination of categorical and continuous variables to be processed. Lastly, it does not require a complete set of edits beforehand.

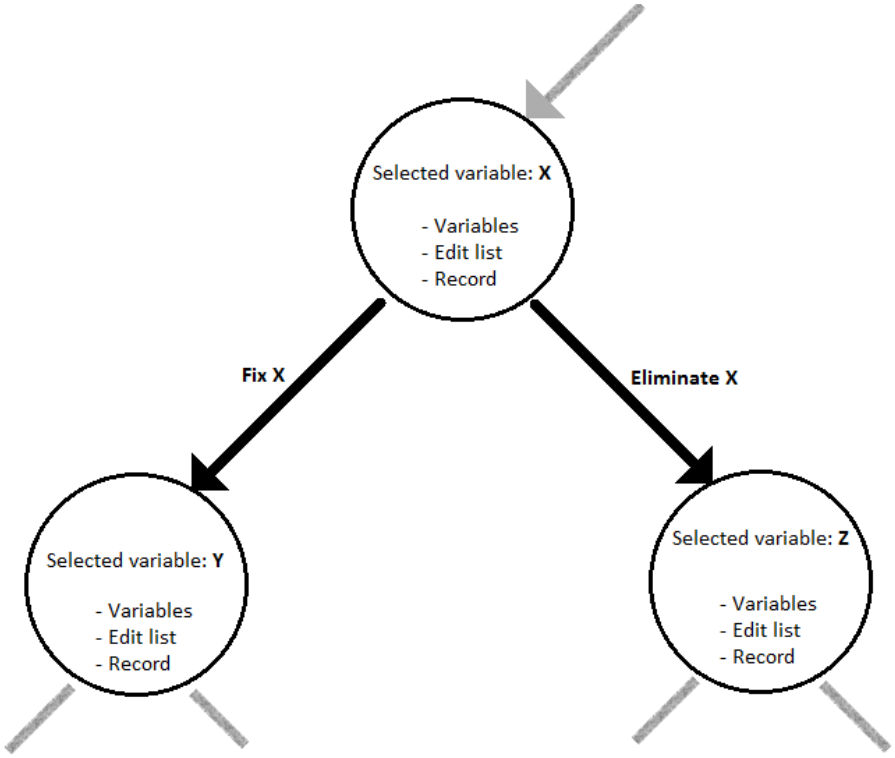

Figure 4: A binary tree used for solving the error localization problem containing the selected variable, the not handled variables, the edit list and the record number.

At first the root node is constructed, containing the total list of explicitly defined edits, together with the list of not-handled variables, which contain all but one, which is the currently selected variable. The order of selected variables is based on the suspiciousness, from most suspicious to least suspicious. A variable is suspicious when (28):

- the number of violated edits in which the variable is involved is high.
- the number of violated edits that can be satisfied by changing only the value of the variable under consideration is high.
- the number of satisfied edits in which the variable is involved is low.

This node then gets two child nodes. In the left child it is assumed that the currently selected variable in the parent is correct and in the right child it is assumed that it is incorrect. This is respectively called fixing and eliminating variables and it is a method to simplify the problem to retrieve solutions. Updating the handled variable list and more important, updating the edit list plays a crucial role. The selected variable is eliminated using FME as described earlier in this paper. Note that by eliminating a variable, newly implied edits are derived, hence the edit list needs to be updated as well (see paragraph on FME). Fixing a variable means that the variable is replaced by its current value in every edit, hence this also updates the edit list. The next figure shows an example of a variable that is fixed in a blue area enclosed by edits and retrieves its original value x = 5. From that follows that the variable on the y-axis can contain a value between 0,45 and 0,7.

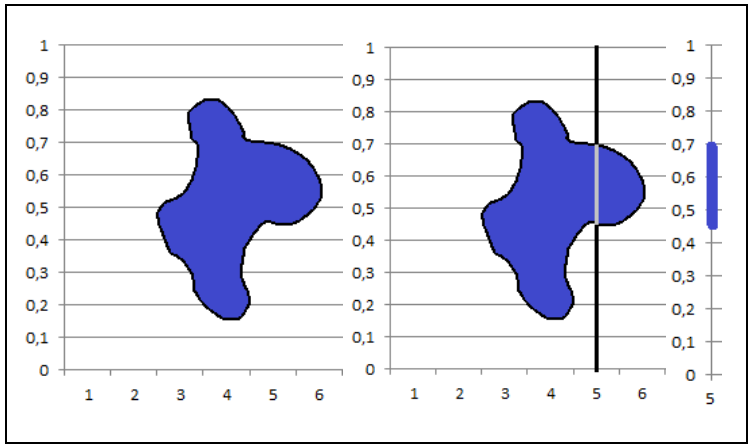

Figure 5: Fixing a variable

Sometimes tautologies are formed like $1 \geq 0$ and these edits can be discarded. In case inconsistencies are formed in the edit list ($1 \leq 0$), than this branch cannot contain a valid solution anymore and is not constructed any further. In case of missing values, only the elimination of a variable can be performed. The solutions are found in the leaves of the tree, stating for every variable if it is assumed to be correct or incorrect. The optimal solution(s) are found on the left side of the tree, since there the least number of variables are considered to be in error.

In order to deal with a mix of categorical and continuous data, the continuous variables are treated first and the categorical second. More on the mix of types of data for this algorithm see (29).

Other promising research is done based on Benders' decomposition (30). Another promising direction on mathematical logics is using solvers for the satisfiability problem (31) (32).

5. DISCUSSION

There are several methods for automatically detecting and correcting survey data. The most used and known methods include systems based on the generalized FH-paradigm. Many research has been done on the basis of set-covering problems and more recent procedures involve a so called branch and bound algorithm.

Within the set-covering problem a lot of effort is put into a better generation of implied edits. This is because the original complete set of edits is often too large to be fully generated. Another important goal was to make the procedure suitable for different kinds of edits and combinations of edits. The branch and bound algorithm efficiently only generates implied edits when this is required for the variable that is to be eliminated. This means that some implied edits never have to be generated, since the algorithm prunes unsuccessful branches of the tree. Statistics Netherlands has developed systems for both procedures and decided to continue with the branch and bound algorithm, since it is easier to maintain and shows a better performance.

An alternative to FH-systems are systems based on the Nearest Neighbour Imputation Methodology (NIM), for example used at statistics Canada. This algorithm only uses edits to classify whether a record fails and then directly choses a best fitting single donor from several good potential donors. This methodology is data-driven, since it depends on a (large) amount of possible donors.

Besides Statistics Netherlands extensive research is done on the performance of several editing and imputation systems (mostly on FH- and NIM-systems, from which the EUREDIT project in 2004 was one of the largest (33). Also from this project it seems that the best performance cannot easily be assigned to one system. The detection as well as the imputation of 'better' values strongly depends on the type of data (categorical, continuous), the type of error (random, specific) and the field of interest (geographic data, household, business). Scientists therefore propose the combination of several systems to tackle every specific type of inconsistencies separately (34) (35). In general it is said that FH-systems perform better on categorical data, while NIM-systems are better with continuous data (1). The combination of these two algorithms has also been tested with some good performance (36).

Besides the data editing tools, there is also a growth of computer-assisted data collection tools (37) (38) (39) (40). One could think of the many modern web survey software and computer-aided tools that interactively inform respondents of probable mistakes and therefore prevent errors in later phases. One can argue that these techniques would overtake the necessity to automatically edit data later. Another recent trend is that the growing reluctance of the household population to survey requests has increased the effort that is required to obtain interviews, and, thereby, the costs of data collection (41) (42) (40). Reminders of all inconsistent answers during the data collection phase (i.e. the interview) can make this reluctance by respondents grow even faster, in the sense that they abandon the survey or send in incomplete forms. Therefore it is probably wiser to interactively correct the most influential inconsistencies only, those that have a substantial influence on publication figures. The same holds for the automatic editing process, because records can also be over-

edited, meaning that too much time is spend at records that do not have a noticeable impact on the ultimately published figures (8) (43). Nowadays it is common to edit only small inconsistencies automatically and let field-experts treat the most influential inconsistencies and the records containing too many errors. This strategy is referred to as selective editing. A good survey strategy in order to attain high quality data could be that of the combination of modern computer-assisted data collection tools and selective and automatic editing. More research is required on the combination of different data editing systems and their performance, as well as a general survey strategy for obtaining high-quality survey data in order to improve the timeliness, accuracy, detail and quality of statistical information and cut the related ever increasing costs.

6. LIBRARY

1. **De Waal, T., Pannekoek, J. and and Scholtus, S.** *Handbook of Statistical Data Editing and Imputation.* Hoboken, New Jersey : John Wiley & Sons, Inc, 2011.

2. What is a survey. [Online] [Cited: 08 01, 2012.] http://whatisasurvey.info/.

3. *Internet, mail, and mixed-mode surveys: The tailored design method (3rd ed.).* **Dillman, D. A., Smyth, J. D. and and Christian, L. M.** s.l. : Hoboken, NJ: John Wiley & Sons, 2009.

4. *A Systematic Approach to Automatic Edit and Imputation.* **Fellegi, I. P. and and Holt, D.** s.l. : Journal of the American Statistical Association, 1976.

5. **Willeboordse, A.** *Handbook on the Design and Implementation of Business.* 1988.

6. *Methods of weighting for unit non response.* **Holt, D. and and Elliot, D.** s.l. : The Statistician, 1991.

7. *Data Editing and its Impact on the Further Processing of Statistical Data.* **Granquist, L.** s.l. : Workshop on Statistical Computing, Budapest., 1984.

8. *Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data.* **Di Zio, M., U., Guarnera and and Luzi, O.** ISTAT, Rome : s.n., 2005.

9. *An Introduction to the Data Editing Process.* **Ferguson, D. P.** United Nations, Geneva. : Statistical Data Editing, Volume 1: Methods and Techniques, 1994.

10. *A Review of Hot Deck Imputation for Survey Non-response.* **Little, Rebecca R. Andridge and Roderick J. A.** s.l. : International Statistical Review , 78, 1, 40–64 doi:10.1111/j.1751-5823.2010.00103.x, 2010.

11. *Statistical data editing for agricultural surveys.* **Pannekoek, J. and and De Waal, T.** s.l. : John Wiley & Sons, Ltd, 2010.

12. *MissForest—non-parametric missing value imputation for mixed-type data.* **Stekhoven, P. and and Bühlmann, D. J.** s.l. : BIOINFORMATICS, 2012.

13. *Fourier–Motzkin Elimination and Its Dual.* **Dantzig, G. B. and and Eaves, B. C.** s.l. : Journal of Combinatorial Theory (A) 14, 288–297., 1973.

14. Set Covering Problem. [Online] [Cited: 08 01, 2012.] http://serverbob.3x.ro/IA/images/fig1056_04.jpg.

15. *Optimal Imputation of Erroneous Data: Categorical Data, General Edits.* **Garfinkel, R. S., Kunnathur, A. S. and and Liepins, G. E.** s.l. : Operations Research 34, pp.744–751, 1986.

16. *Set-Covering and Editing Discrete Data.* **Winkler, W. E.** Washington, D.C. : Statistical Research Division Report 98/01, U.S. Bureau of the Census, 1998.

17. *A Comparison Study of ACS If-Then-Else, NIM, DISCRETE Edit and Imputation Systems Using ACS Data.* **Chen, B., Thibaudeau, Y. and and Winkler, W. E.** s.l. : Paper No. 7, UN/ECE Work Session on Statistical Data Editing, 2003.

18. *An Algorithm for the Fields to Impute Problems of Numerical and Coded Data.* **Sande, G.** s.l. : Technical report, Statistics Canada, 1978.

19. *Use of Chernikova's Algorithm in the Generalized Edit and Imputation System.* **Schiopu-Kratina, I. and and Kovar, J. G.** Statistics Canada : Methodology Branch Working Paper BSMD 89-001E, 1989.

20. **Fillion, J. M. and and Schiopu-Kratina, I.** *On the Use of Chernikova's Algorithm for Error Localization.* Statistics Canada : s.n., 1993.

21. *The SPEER Edit System. In: Statistical Data Editing, Volume 2: Methods and Techniques.* **Winkler, W. E. and and Draper, L. A.** United Nations, Geneva : s.n., 1997.

22. *Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses.* **Bankier, M.** s.l. : UN/ECE Work Session on Statistical Data Editing, 1999.

23. *A Generic Implementation of the Nearest-Neighbour Imputation Methodology (NIM).* **Bankier, M., et al., et al.** s.l. : Proceedings of the Second International Conference on Establishment Surveys, 2000.

24. *A Fast and Simple Algorithm for Automatic Editing of Mixed Data.* **De Waal, T. and and Quere, R.** s.l. : Journal of Official Statistics 19, pp. 383–402., 2003.

25. *Algorithms for Automatic Error Localisation and Modification.* **De Waal, T.** s.l. : Paper prepared for the DATACLEAN 2002 conference, 2002.

26. **Daalmans, J.** *Automatic Error Localization of Categorical Data.* Statistics Netherlands, Voorburg : s.n., 2000.

27. *An automatic method of solving discrete programming problems.* **Doig, A. H and and Land, A. G.** s.l. : Econometrica 28 (3): pp. 497–520. DOI:10.2307/1910129., (1960).

28. **Chung, W.H.** *Effective Automatic Error Localisation.* s.l. : Internal report Statistics Netherlands, 2003.

29. *Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project.* **Pannekoek, J. and and De Waal, T.** s.l. : Journal of Official Statistics 21, pp. 257–286., 2005.

30. *New Algorithms for the Editing and Imputation Problem.* **Riera-Ledesma, J. and and Salazar-Gonzalez, J. J.** Madrid : Working Paper No. 5, UN/ECE Work Session on Statistical Data Editing, 2003.

31. *Logical Formalisation of the Fellegi–Holt Method of Data Cleaning.* **Boskovitz, A., Goré, R. and and Hegland, M.** s.l. : Research School of Information Sciences and Engineering, 2003.

32. *Data Editing and Logic: The Covering Set Method from the Perspective of Logic.* **Boskovitz, A.** s.l. : Australian National University, 2008.

33. EUREDIT Project. [Online] [Cited: 08 01, 2012.] http://www.cs.york.ac.uk/euredit/.

34. *Designing a complete edit strategy; combining techniques.* **De Jong, W. A. M.** s.l. : Wrk Sessn Statistical Data Editing, Voorburg, 1996.

35. *Combining methodologies in an editing and imputation procedure: the survey of Balance Sheets of Agricultural Firms.* **Di Zio, M. and and Luzi, 0.** s.l. : Statist. Appl., 14, 59-80, 2002.

36. *Combining Editing and Imputation Methods An Experimental Application on Population Census Data.* **Manzari, A.** s.l. : Journal of the Royal Statistical Society, 2004.

37. *Using statistical methods applicable to autocorrelated processes to analyze survey process quality data.* **Hapuarachchi, P., March, M. and Wronski, A.** s.l. : Survey Measurement and Process Quality, 1997.

38. *Measuring Survey Quality in a CASIC Environment.* **Couper, M.** Dallas : Joint Statistical Meetings of the American Statistical Association, 1998.

39. *Macro and Micro paradata for survey assessment.* **Scheuren, F.** s.l. : United Nations Work Session Statistical Metadata, Washington DC., 2001.

40. **Cobben, F.** *Nonresponse in Sample Surveys: Methods for Analysis and Adjustment.* 2009.

41. *Trends in household survey nonresponse: A longitudinal and international comparison.* **De Leeuw, E.D and and De Heer, W.** s.l. : R.M. Groves, D.A. Dillman, J.L. Eltinge, & R.J.A. Little (Eds). Survey nonresponse. New York: Wiley, pp. 41-54., 2002.

42. *Nonresponse in Household Interview Surveys.* **Groves, R. and and Couper, M.** s.l. : Wiley series in probability and statistics. Survey methodology section., 1998.

43. **Granquist, L. and and Kovar, J.** Editing of Survey Data: How Much Is Enough? 1997.

44. *Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse.* **Rubin, D. B. and and Schenker, N.** s.l. : Journal of the American Statistical Association 81, pp. 366–374, 1986.

45. **Ferguson, D. P.** *An Introduction to the Data Editing Process.* United Nations, Geneva. : Statistical Data Editing, Volume 1: Methods and Techniques, 1994.