RESEARCH REPORT SERIES

(Statistics #2006-7)

Data Quality:
Automated Edit/Imputation and Record Linkage

William E. Winkler

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: July 12, 2006

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Data Quality: Automated Edit/Imputation and Record Linkage

William E. Winkler 1/, william.e.winkler@census.gov
Statistical Research Division, U.S. Census Bureau, Washington, DC 20233-9100

Abstract: Statistical agencies collect data from surveys and create data warehouses by combining data from a variety of sources. To be suitable for analytic purposes, the files must be relatively free of error. Record linkage (Fellegi and Sunter, *JASA* 1969) is used for identifying duplicates within a file or across a set of files. Statistical data editing and imputation (Fellegi and Holt, *JASA* 1976) are used for locating erroneous values of variables and filling-in for missing data. Although these powerful methods were introduced in the statistical literature, the primary means of implementing the methods have been via computer science and operations research (Winkler, *Information Systems* 2004a). This paper provides an overview of the recent developments.

1 Introduction

The goal of quality is straightforward. It is "suitability for use." By suitability for use, we mean that a file of microdata can be used for producing aggregates such as totals in publications and for certain auxiliary analyses. In an ideal situation, we would want all of the values of fields in a record to be consistent with some underlying 'truth.' As an example, if a field represents an individual's income, we would want the value of the field to be quite accurate. If the field is a date-of-birth, then we would want the dates-of-birth to be represented consistently across records within the file and be accurate. If the fields need to be converted to a common format of a larger reference file such as a register or data warehouse, then we would want the values in the fields to be in a form that can be converted to the reference-file format in a straightforward manner.

Statisticians who work with files can quite easily notice when values of certain fields are missing. They sometimes notice contradictory information such as a child in a household that is less than 16 and married. Delineation of the contradictory values of individual fields usually is based on subject matter expertise or knowledge of the statistical uses of a file.

Two of the most powerful methods for improving data quality originated and have been significantly developed in statistical agencies. The first method is the record linkage model of Fellegi and Sunter (1969) that is intended to facilitate the development of methods that allow location of duplicate records within a file or corresponding records across two files. Matching is most often done with *quasi-identifiers* such as name, address, and date-of-birth. The second method is the edit/imputation model of Fellegi and Holt (1976) that allows filling-in missing values or replacing contradictory values of fields in records. The intent of the filling-in is that joint distributions be preserved in a principled manner.

Generalized record linkage systems are characterized by the relative ease of applying the methods to different files. They are also sometimes characterized by the ability to automatically estimate record linkage error rates without training data (Winkler and Yancey 2006, Belin and Rubin 1995). Recent methods (Yancey and Winkler 2004) are also characterized by speed (150,000 or more pairs of records per second) of matching that includes all steps of pre-processing, sorting, and linking. Generalized edit/imputation systems are characterized by rapidity of applying the methods in new situations and the speed with which 'clean' and 'complete' output files are produced.

In this paper (that is primarily review), we highlight the significant improvements in quality that can be achieved with the generalized record linkage and edit/imputation methods. The generalized methods with suitable software almost totally eliminate the need for extensive programming staffs and the debugging and removal of programming errors. All the software is easily ported to new survey situations. All the main logic is controlled by easily changed and maintained parameters (record linkage and edit/imputation) and tables of edit rules (edit/imputation). The main computational and optimization algorithms never need to be changed. The greater accuracy and quality of the output files usually drastically reduces (and sometimes eliminates) the need for expensive, time-consuming clerical review by large teams of analysts. The time improvements are such that a sufficiently skilled analyst/programmer may produce files of higher quality in 6-12 weeks than a team of 6 programmers and 12 analysts may produce in 9-18 months using conventional methods. A key difference is that, whereas sophisticated methods and logic are built into the generalized systems (effectively raising institutional knowledge and day-to-day practice), the classical methods often are done in an ad hoc fashion in which logic rules and programs are rewritten from scratch for each survey (or even iteration of a previous survey that is slightly re-designed).

The effectiveness, ability to achieve quantifiably improved quality, and very significantly reduce programming and clerical-review resources has been repeatedly demonstrated in statistical agencies in Canada, Italy, the Netherlands, the UK, the USA, and other countries. In this paper, we will provide examples of the improvements primarily from production situations at the U.S. Census Bureau and indicate analogous (and often better) improvements at some of the European agencies. The reduction in programming and clerical-review resources and the amount of time needed to implement a survey seem quite remarkable but have been independently repeated in all of the aforementioned countries in addition to others. In all situations, the implementers of the generalized methods have been able to demonstrate that the new methods quantifiably improve the quality of microdata.

Throughout this paper, we assume that there is a single file or set of files that can be accessed and easily used by the analyst/programmer using the generalized software. We do not cover issues of data capture in which a survey form or collection instrument are used to get data in a suitable form in computer files. We do, however, touch on the skills that managers in an agency need to foster to develop generalized systems and assure that the methods can be applied by a number of individuals (possibly after suitable training). Winkler and Hidiroglou (1998) have described the individual technical skills of successful teams that developed and later implemented the generalized methods on multiple surveys.

The outline of this paper is as follows. In the second section, we give general background on the Fellegi-Sunter model of record linkage and the Fellegi-Holt model of edit/imputation. We describe why their theoretical models solved problems that earlier methods did not. We indicate the significant advances in algorithms that have made production systems exceptionally fast and some newer advances that are even more powerful. We provide more details of implementations of the methods in actual production environments. Advantages of the general production software are the ease of applying the methods in new situations, the embedding of valid, high quality methods in software, and the drastically reduced need for programming and clerical resources. In the third section, we discuss some alternative methods, summarize issues related to the profound errors that the errors can have on analyses, and give some details on management issues. In the final section, we provide concluding remarks.

2 Background

In this section, we describe background about the advantages of the Fellegi-Sunter model of record linkage and the Fellegi-Holt model of edit/imputation.

2.1 The Fellegi-Sunter model of record linkage

Fellegi and Sunter (1969) provided a theoretical model for record linkage methods introduced by Howard Newcombe (Newcombe et al. 1959, 1962). They also provided methods for automatically computing optimal matching parameters (probabilities) without training data. At the simplest level, we may wish to match two files **A** and **B** using name and address information. An agreement pattern γ might be whether a pair in the product space **A** × **B** agrees or does not agree on corresponding fields such as first name, last name, house number, and street name. We partition the product space **A** × **B** into the set of matches M and set of nonmatches U.

Newcombe, a geneticist, used odds ratios of the form

$$R = P(\gamma|M) / P(\gamma|U)\,. \qquad (1)$$

Newcombe's decision (classification) rule **R** is

R1. If $R > T_\mu$, then designate pair as a match.

R2. If $T_\lambda \le R \le T_\mu$, then designate pair as a possible match
and hold for clerical review. (2)

R3. If $R < T_\lambda$, then designate pair as a nonmatch.

Newcombe's rule agrees with intuition. If γ consists primarily of agreements on different fields, then γ is likely to occur more often among matches than among nonmatches (**R1**). If γ consists primarily of disagreements, then γ is likely to occur primarily among nonmatches rather than matches (**R3**). Fellegi and Sunter proved (1969, Theorem 1) that, given a fixed upper bound μ of the rate of false matches and fixed upper bound λ on the rate of false nonmatches, the decision rule is optimal in the sense that it minimizes the in-between (i.e., clerical review) region **R2**. The cutoffs T_λ and T_μ are determined by the error rates λ and μ, respectively. Further, Fellegi and Sunter (1969) showed, under a conditional independence assumption, how to estimate the simple yes/no (agree/disagree) probabilities when only three fields are considered. They also provided other methods of estimating matching parameters (probabilities) without training data.

2.1.1 A real-world problem and new methods

For the 1990 U.S. Decennial Census, Hogan and Wolter (1988) introduced a method of adjusting the Census for undercount and overcount. The method required that all individuals in a sample of Census blocks (geographic region representing approximately 70 households) be re-enumerated and matched to the main Census. A capture-recapture method (Sekar and Deming 1949) would be used to estimate population counts. Hogan and Wolter initially believed that the computer matching error would exceed the total of the remaining error due to six other error sources that they had identified. The solutions and their successes in the initial application and later applications are described below.

Winkler (1988) provided an EM algorithm for estimating simple yes/no probabilities for situations of more than 3 matching fields, gave methods for automatically estimating frequency-based weights for some strings that were adjusted to the overall EM weights, and gave an enhancement to string comparator methods introduced by Jaro (1972). With frequency-based weights (called value-specific weights by Newcombe), names such as 'Garcia' and 'Martinez' might automatically be down-weighted in areas such as Southern California where they are among the most common. In areas such as Minneapolis, the names would be automatically up-weighted because 'Garcia' and 'Martinez' are relatively rare in those regions. The string comparators were necessary because of the high typographical error rate due to transcription, keying, and other errors in some files being matched. For instance, based on 1988 Decennial Census Dress Rehearsal data (truth decks from clerical review, field work, and adjudication), Winkler (1990a) noted that approximately 25% of first names and 15% of last names among matches could not be brought together with exact character-by-character matching. Winkler (1989a) had also noted that the optimal matching parameters of the form P(agree field | M) varied significantly across regions (particularly from an urban region to an adjacent suburban region).

Table 1 provides an illustration of the cost and time-savings of the production software. The numbers from 1990 (that are complete) are extrapolated to the numbers from the 1988 Dress Rehearsal and to numbers suggested by Hogan and Wolter if only clerical review instead of automated procedures were used.

Table 1. Resources for US Decennial Census Matching

	clerical	computerized	
		1988	1990
# clerks	3000	600	200
# month	6	1.5	1.5
false match rate	5%	0.5%	0.2%
computer match proportion	0%	70%	75%

With the (semi-)automated procedure of 1988 and 1990, match status of clerical pairs (i.e., those below the upper cutoff and above the lower cutoff) were determined via a quick review of individuals in corresponding households. Most of the clerical pairs consisted of individuals in the same household who were missing both first name and age. The only way to distinguish individuals in the same household (who typically agree on characteristics such as last name, street name, house number, and telephone number) was to compare first name or age.

2.1.2 Other applications

Combining a set of lists to produce a frame for an agricultural survey or census is known to be a significant problem. Agricultural entities can be corporations, partnerships, or individuals. The U.S. Department of Agriculture (USDA) introduced separate methods (Coulter 1977) for matching corporations, partnerships, and individuals but had difficulty matching across the respective sublists. The Census Bureau (which previously did the Agriculture Census that is now in USDA) began with twelve lists containing approximately 16 million records. The initial unduplication was via a corporate id (Employer Identification Number or EIN) for companies or SSN (social security number) for individuals and most partnerships. After the initial phase of unduplication on the unique identifiers, 6 million records remained that needed to be unduplicated via name and address matching.

Table 2 illustrates the improvements between the Agriculture List Development in 1987 and 1992. In 1992 a modified version of the generalized matching software was used along with name standardization software developed for the 1982 and 1987 agriculture operation and address standardization software from the 1990 Census. The first improvement was in the speed of the operations. The entire computer matching operations needed less than three weeks and the overall matching (including semi-automated clerical review of potential duplicates developed by Agriculture programmers) added a few more weeks. In comparison, the entire 1987 operation lasted more than 3 months. Based on large field validations, the final files in 1987 contained approximately 10% duplicates whereas the final files in 1992 contained approximately 2% duplicates. In fact, the 1992 files contained less duplication after the automated matching and prior to the clerical review than the 1987 files. With 10% duplicates in a survey frame, it is exceptionally difficult to produce reliable totals and other estimates.

Table 2. Updating and unduplicating an agricultural survey frame
Identify duplicates in 6 million records from 12 lists

duplicates	6.6%	12.8%
potential duplicates	28.9%	19.7%
final file duplication	~10%	~2%
clerical resources	75 clerks for 3 months	6500 person hours

As a further quality improvement, the high-quality name standardization routines (embedded) in thousands of lines of 1987 Agriculture Census FORTRAN processing code were extracted and placed in parameter-driven C routines. These general routines make it straightforward for economists and statisticians to use the software because they only need to know the name of the input file and the location of the free-form name information. The standardized outputs are placed in a set of columns at the end of the output file. These generalized routines (and generalized routines developed by the Geography Division for address standardization) are available to individuals within and without the Census Bureau. Prior to the creation of the name standardization routines, virtually all individuals in other areas of the Census Bureau were unaware of the high quality name standardization routines developed by the Agriculture Division.

2.2 Edit/Imputation

Edit/Imputation is a methodology for filling-in missing values and for replacing contradictory values with values that do not cause a record to fail edits. The edit methods of Fellegi and Holt

(1976) were the first to assure that a record could be ‘corrected’ in one pass. Prior to Fellegi and Holt, records would be ‘corrected’ for failing certain edits and then fail other edits that they did not originally fail. By ‘correcting’ a record, we mean changing a number of values in fields that would cause the ‘modified’ record to pass all edits. Fellegi-Holt generalized systems have significant advantages. The first is that the edits are all contained in easily modified tables and that the logical consistency of the entire set of edits can be checked prior to the receipt of production data. The second is that the main optimization routines (integer programming for finding the minimal number of fields to change) never need to be changed for new data from different surveys. Classical edit/imputation often required hundreds (or thousands) of edit rules that took months of time for teams of subject-matter specialists to develop and for programmers to write the code. The systems sometimes did not assure that ‘replacement’ values in edit-failing records would cause a record to pass edits. Instead, the systems required that analysts (sometimes large teams for months) clerically review, change a record, resubmit it to the edit software, and iterate until the record passed edits.

2.2.1 Fellegi-Holt systems are exceptionally fast to implement and improve data quality

Early Fellegi-Holt systems were Statistics Canada’s GEIS (Generalized Edit/Imputation System) system and the later SPEER (Structured Programs for Economic Editing and Referral) system from the Census Bureau. Both were designed for continuous economic data. In a head-to-head comparison, Kovar and Winkler (1996) demonstrated that both GEIS and SPEER could be installed in less than one day. The primary reason for the rapid edit/imputation system creation is that the set of edits were known and could be quickly placed in the edit tables. This rapid system development is in contrast to many months of time to write a system from scratch as is typical in the classical situation. Further, without any clerical follow-up, each system produced comparably high quality output files that satisfied all edits. As a further example of rapid system development, Winkler (1999) applied the DISCRETE edit/imputation system to two small demographic surveys in less than half a day on each.

Statistical agencies often did not develop Fellegi-Holt systems due to the difficulty of integer-programming methods for the main optimization algorithms and the speed of the resultant software. These situations are no longer true. Statistics Canada (BANFF for continuous – successor to GEIS – Mohl et al. 2005 and CANCEIS – Bankier 1991, 2000 for discrete data and some other types of data), Census Bureau (SPEER – Draper and Winkler 1999 and DISCRETE – Winkler 1997), Statistics Netherlands (CherryPi, De Waal 2003a-d for general data – continuous and discrete simultaneously), and ISTAT (Bruni et al. 2001a,b, 2003, 2004, 2005- primarily discrete) have all developed systems that are fast, can be maintained, and are applied to multiple surveys.

In a very dramatic demonstration, Garcia and Thompson (2000) showed that a Fellegi-Holt system was able to edit a large economic survey with complicated, interlaced edit patterns in less than 24 hours. In contrast, a team of 12 analysts took 6 months to ‘correct’ the data and changed three times as many fields.

2.2.2 Additional implementation issues and research problems

Fellegi and Holt (1976) indicated (for discrete data) that, after the minimal number of fields to change had been determined, hot-deck could be used for imputation of missing values. Much recent work has shown that hot-deck does not reliably preserve joint distributions of data (see e.g., Little and Rubin 2002). As a hot-deck example, a record of 10 fields may need to have three values imputed which necessitates that matching be performed on seven fields (associated with non-missing values). Typically, there are no donors that match on seven fields, on an a priori selected set of five fields, and often on three fields. The method of reducing the number of fields on which to match is typically done in ad hoc guesses with no detailed evaluation of the effects on joint distributions. If the matching is only on a subset of variables, then the joint distributions are compromised. The method of Little and Rubin (2002) for building a multinomial model of the data under the missing-at-random assumption yields a valid model from which matching draws are assured to satisfy joint distributions and allow variance estimation. Variance estimation is not always easy or even valid with hot-deck.

Winkler (2003) has shown how to connect (actually embed in a larger modeling framework) the imputation methods with the edit methods. He has further shown how the missing-at-random assumption can be eliminated when there is auxiliary data that is available for the nonignorable missing data situation because the generalized fitting algorithms (Winkler 1990, 1993 – originally developed for more difficult record linkage problems) allow additional (logical) restraints.

An open issue related to the imputation is speeding up the iterative-fitting algorithms in the basic multinomial modeling when there are more than 12 fields and extending the modeling ideas to continuous data. Both Bruni (2004, 2005) and Winkler (2003) have suggestions for putting the continuous data in discrete form by separating it into a suitable number of equal-sized intervals. The advantage of the ideas is that general loglinear modeling methods (e.g. Bishop, Fienberg, and Holland 1975) can be used in a 'turn-the-crank' method to get valid limiting distributions. At present, there are no general methods (i.e., turn-the-crank) for continuous distributions. Di Zio et al. (2004, see also Thibaudeau and Winkler 2003) have provided more elementary methods of imputation based on Bayesian networks that (nearly) automatically create the models. These Bayesian networks methods demonstrably improve over hot-deck because they approximately preserve joint distributions. Bayesian network software is readily available (sometimes as free-ware).

3 Discussion

We break the discussion into several components. In the first, we discuss some widely-used alternatives (or supplemental methods) to the Fellegi-Holt edit/imputation methods. In the second, we provide two rules about data error and a vision for how a suitably skilled analyst/programmer (almost single-handedly) might 'clean' the data. The third component covers management issues in relation to developing technical skills.

3.1 Other edit/imputation methods

Many individuals have suggested using selective editing (Hidiroglou and Berthelot 1986, Latouche and Berthelot 1992, Des Jardins 1998) that are based on exploratory-data-analysis (EDA Tukey 1977) and closely related graphical principles. The basic idea of the first two methods is to target the most important aggregates (or closely related alternative aggregates) produced by a survey. To significantly limit clerical review, only the most important records are sent to follow-up. All three methods will also sometimes locate 'errors' or anomalies that might not be located via Fellegi-Holt methods where the edit rules are determined a priori by subject-matter specialists.

Both De Waal (2003d) and Winkler (1999) have suggested using the selective editing methods for additional review after a set of data has been processed by Fellegi-Holt methods. The

Fellegi-Holt methods provide a complete, ‘corrected’ data file in which all records satisfy the a priori set of edits that are typically specified by subject-matter specialists. At present, there is considerable anecdotal evidence that analysts do not know how to ‘correct’ records in a reliable, consistent manner that improves aggregates in a set of microdata. There have been absolutely no studies that demonstrate the situations in which a suitably trained set of analysts might reliably ‘correct’ data to improve its overall quality. These observations suggest that it is best to reduce the amount of manual editing (from the amounts that are currently done on many surveys) unless there is some quantifiable improvement in certain estimates that must be produced from the data file.

3.2 Two rules of thumb and a vision

The first rule is:

T1. Frame errors due to omissions and duplication can yield greater errors in data than all other sources of error combined.

The second rule is:

T2. Edit/imputation error (when there are no frame errors) can yield greater errors in data than all other sources of error combined.

The first rule of thumb is consistent with Hartley (1962, 1974) in terms of agriculture frames and even with the frame/unduplication projects with which this author has been involved. The second rule of thumb can occur when a substantial amount of values of data fields are in error. This particularly happens with some economic surveys when a small proportion (say 1-5%) of records has values in fields that are in error by a factor of ten (both too large in some records and too small in other records).

Most individuals who use the data for analysis assume that the frame and the associated values in data fields are correct. They typically have no way of determining whether frame (coverage/unduplication) and edit/imputation errors are occurring. We do note that errors are often discovered by analysts when the computer outputs associated with a study clearly contain severe errors.

Vision: A suitably skilled analyst/programmer with generalized record linkage and edit/imputation software can reasonably quickly ‘clean’ a data file or set of data files so that the resultant data are of higher quality than data produced using the classical methods.

The vision is consistent with the author’s experience in which Fellegi-Holt edit/imputation methods and Fellegi-Sunter record linkage methods have been developed (primarily installed) in a few days for smaller survey situations. We note that none of these methods can (easily) overcome situations when there are significantly problems in data capture. Data-capture problems can occur when a survey form, keypunch software, or computer-assisted interviewing software are not as well designed or implemented as they might be. Other problems similar to data-capture problems can occur when files are obtained from different sources (such as other areas of an agency, other statistical agencies, or commercial organizations) that have neither the resources nor the know-how to effectively clean-up their data. There are many situations where certain data fields are not cleaned up because the fields are not used in a day-to-day basis.

3.3 Management issues

Statistical agencies need to develop individuals and teams with suitable skills for building generalized systems that can be applied to a variety of surveys. Winkler and Hidiroglou (1998) describe some of the skills and several successful projects at Statistics Canada and the U.S. Census Bureau. In the edit/imputation area, agencies may need one or more individuals with substantial knowledge (theoretical and methodological) and programming skills in integer programming, set covering algorithms, and logic programming. For imputation, it is likely that statisticians may need to develop some new methodological skills and possibly do associated programming. For instance, developing discrete-data imputation models as in Little and Rubin (2002) primarily requires basic knowledge of loglinear modeling (Bishop, Fienberg, and Holland 1975) and iterative proportional fitting using EM-type methods. While such skills are widely used among computer scientists and university statisticians, the skills are often little known in statistical agencies.

In the record linkage area, although the Fellegi-Sunter model is based on straightforward hypothesis testing, almost all the skills needed for implementation are computer science. The skills involve extraction of information (i.e., comparable strings) from semi-free-form text such as names, addresses, and dates-of-birth (particularly when not in comparable formats), string comparators for dealing with typographical error, methods for comparing records in enormous files ($10^{11} - 10^{13}$), unsupervised methods of learning without training data, and other skills.

Winkler and Hidiroglou (1998) further observed that it is often difficult for different groups within an agency to use methods and code from other parts of the agencies. Individuals (computer programmers, statisticians, and others) are seldom given training and direction on how to develop modules of code that can be relatively easy to use even with other members of their own programming teams. It is often difficult for the programmers and statisticians to communicate concepts effectively because the statisticians have little understanding of the programming concepts and the programmers have little understanding of many of the advanced methods. Neither programmers nor statisticians have understanding of the types of algorithms needed to implement fast production systems.

In terms of basic institutional knowledge, it is very unusual in most agencies to find individuals (either statisticians or programmers) who can explain how hot-deck (a main component of the widely used classical methods) was developed and refined on a particular survey. Although it is well-known that simple variants of hot-deck can significantly improve hot-deck implementations with particular types of survey data, seldom are there individuals who can explain if different variants were tested. In many situations, it is difficult to find individuals who have any training from other individuals who have developed valid hot-deck methods or know how to eliminate some of the errors that often occur in hot-deck.

If agencies do not have certain basic technical skills in terms of the best available theory and methods and of ways of developing efficient computer algorithms, it is not clear how the agencies can identify important improvements that they need to make in the overall methods that they use. As an instance, an agency may have significant need for suitable i/o (industry and occupation) coding methods that can be used on a variety of surveys. The agency may be totally unaware of the most efficient new methods in the machine learning literature and how to implement them effectively. Although agencies do considerable sampling (one variable typically), most surveys need to provide multiple estimates. Can the agencies determine effective multi-variable sampling methods and how to implement them?

4. Concluding Remarks

This paper provides background on the significant savings in terms of skilled person resources, clerical review resources, and time of using generalized systems based on the Fellegi-Sunter model of record linkage and the Fellegi-Holt model of edit/imputation. The methods also produce final data files that are of quantifiably higher quality than most data files produced by classical methods.

1/ Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U. S. Census Bureau.

References

Bankier, M. (1991), “Alternative Method of Doing Quantitative Variable Imputation,” Statistics Canada Memorandum.

Bankier, M. (2000), “2001 Canadian Census Minimum Change Donor Imputation Methodology,” U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Cardiff, UK, October 2000 (also available at http://www.unece.org/stats/documents/2000.10.sde.htm).

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694-707.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Bruni, R. (2003), “Solving Error Correction for Large Data Sets by Means of a SAT Solver,” in (E. Giunchiglia and A. Tachella, eds.) *Lecture Notes in Computer Science 2919*, Springer-Verlag, 229-241.

Bruni, R. (2004), “Discrete Models for Data Imputation,” *Discrete Applied Mathematics*, 144, 59-69.

Bruni, R. (2005), “Error Correction for Massive Datasets,” *Optimization Methods and Software*, 20, 297-316.

Bruni, R., Reale, A., and Torelli, R. (2001), “Optimization Techniques for Edit Validation and Data Imputation,” Statistics Canada Symposium 2001, Ottawa, Ontario, Canada, October 2001.

Bruni, R., and Sassano, A. (2001) “Logic and Optimization Techniques for an Error Free Data Collecting,” Dipartimento di Informatica e Sistemistica, Universita di Roma “La Sapienza.”

Coulter, R. A. (1977), “An Application of a Theory for Record Linkage,” National Agricultural Statistical Service, USDA, available in http://www.fcsm.gov/working-papers/1367_1.pdf .

De Waal, T. (2003a), “Solving the Error Localization Problem by Means of Vertex Generation,” *Survey Methodology*, 29 (1), 71-79.

De Waal, T. (2003b), “A Fast and Simple Algorithm for Automatic Editing of Mixed Data,” *Journal of Official Statistics*, 19 (4), 383-402.

De Waal, T. (2003c), “Computational Results with Various Error Localization Algorithms,” UNECE Statistical Data Editing Worksession, Madrid, Spain, http://www.unece.org/stats/documents/2003/10/sde/wp.22.e.pdf.

De Waal, T. (2003d), *Processing of Erroneous and Unsafe Data*, ERIM Research in Management: Rotterdam.

DesJardins, D. (1998), “A New Graphical Techniques for the Analysis of Census Data”, *Statistics Canada Conference Proceedings.*

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., and Ponti, A. (2004), “Bayesian Networks for Imputation,” *Journal of the Royal Statistical Society, A*, 167 (2), 309-322.

Draper, L., and Winkler, W.E. (1997), “Balancing and Ratio Editing with the new SPEER system,” *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 570-575 (also available as Statistical Research Division Report rr97/05 at http://www.census.gov/srd/www/byyear.html).

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

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210.

Garcia, M., and Thompson, K. J. (2000), “Applying the Generalized Edit/Imputation System AGGIES to the Annual Capital Expenditures Survey,“ *Proceedings of the International Conference on Establishment Surveys, II*, 777-789 (also http://www.census.gov/srd/papers/pdf/rr2000-01.pdf).

Hartley, H. O. (1962), “Multiple Frame Surveys,” *American Statistical Association, Proceedings of the Section on Social Statistics*, 203-206.

Hartley, H. O. (1974), “Multiple Frame Methodology and Selected Applications,” *Sankhya, Series C*, 36, 99-118.

Hidiroglou, M.A., and Berthelot, J.-M. (1986), “Statistical Editing and Imputation of Periodic Business Surveys,” *Survey Methodology*, *12*, 73-83.

Hogan, H. H. (1992), “The Post Enumeration Survey: An Overview,” *American Statistician*, 46, 261-269.

Hogan, H. H., and Wolter, K. M. (1988), “Measuring Accuracy in a Post Enumeration Survey,” *Survey Methodology*, 14, 99-116.

Jaro, M. A. (1972), “UNIMATCH – A Computer System for Generalized Record Linkage Under Conditions of Uncertainty,” AFIPS – Springer Conference Proceedings, 40, 523-540.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 89, 414-420.

Kovar, J. G., and Winkler, W. E. (1996), “Editing Economic Data,” *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 81-87 (also available as Statistical Research Division Report rr00/04 at http://www.census.gov/srd/www/byyear.html).

Latouche, M., and Berthelot, J.-M (1992), “Use of a Score Function to Prioritize and Limit Recontacts in Business Surveys” *Journal of Official Statistics*, 8 (3), 389-400.

Little, R. A., and Rubin, D. B., (2002), *Statistical Analysis with Missing Data (2nd edition)*, John Wiley: New York.

Mohl, C., Deguire, Y., Kozak, R., and Marquis, C. (2005), “The Transition from GEIS to BANFF,: UNECE Worksession on Statitical Data Editing, Ottawa, Ontario, Canada, May 2005, http://www.unece.org/stats/documents/2005/05/sde/wp.34.e.pdf.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

Newcombe, H.B., and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, .5, 563-567.

Sekar, C. C., and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, 44, 101-115.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading: MA.

Thibaudeau, Y., and Winkler, W. E. (2002), “Bayesian Network Representations, Generalized Imputation, and Synthetic Data Satisfying Analytic Restraints,” (Research report RRS 2002/09 http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.

Winkler, W. E. (1989b), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783.

Winkler, W. E. (1990a), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

Winkler, W. E. (1990b), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, 18, 1410-1415.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," Pr*oceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

Winkler, W.E. (1997), "Set-Covering and Editing Discrete Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 564-569 (also available as Statistical Research Division Report rr98/01at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1999), "The State of Statistical Data Editing," in *Statistical Data Editing*, Rome: ISTAT, 169-187 (also available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2003), "A Contingency Table Model for Imputing Data Satisfying Analytic Constraints," *American Statistical Association, Proc. Survey Research Methods Section*, CD-ROM, also research Report SRS 2003/07 at http:/www.census.gov/srd/www/byyear.html.

Winkler, W. E. (2004a), "Methods for Evaluating and Creating Data Quality," *Information Systems* (2004), 29 (7), 531-550.

Winkler, W. E. (2004b), "Approximate String Comparator Search Strategies for Very Large Administrative Lists," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM (also report 2005/06 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2006), "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf .

Winkler, W. E., and Hidiroglou, M. (1998), "Developing Analytic Programming Capability to Empower the Survey Organization," Statistical Research Division report 98/04, http://www.census.gov/srd/papers/pdf/rr9804.pdf .

Winkler, W. E., and Yancey, W. E. (2006), "Automatically Estimating Record Linkage False Match Rates ," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.

Yancey, W.E., and Winkler, W. E. (2004), "BigMatch Software," computer system, documentation available at http://www.census.gov/srd/www/byyear.html .