RESEARCH REPORT SERIES
(Statistics #2018-05)

Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation

William E. Winkler

Center for Statistical Research and Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: April 3, 2018

Cleaning and Using Administrative Lists: Enhanced Practices and Computational Algorithms for Record Linkage and Modeling/Editing/Imputation

William Erwin Winkler

1. Introductory Comments

The national statistical institutes (NSIs), including the U.S. Census Bureau, developed the original methods for processing survey data. The NSIs developed systematic, efficient generalized methods/software based on the Fellegi-Holt model of statistical data editing (1976) and the Fellegi-Sunter model of record linkage (1969). The generalized software is suitable for processing files for businesses, survey institutes, and administrative organizations. Early systems yielded high quality results but were often much too slow for even a hundred thousand records. New computational algorithms yield drastic hundred-plus fold speed increases over algorithms used outside the NSIs and previously within the NSIs. The software is suitable for processing hundreds of millions or billions of records in a few days.

The NSIs utilize population-level administrative data, such as censuses of population and housing, tax records, wage and employment records, health records, and records of program participation. In order to combine survey data with these lists, it is necessary to process all files so that they are compatible and of suitable quality. Modeling data and performing edit/imputation can be a few steps in this process. Edit/imputation is an approach to cleaning administrative lists. Record linkage is another process that can be used to enhance the information available for analysis. This paper reviews details of modeling/edit/imputation and record linkage. It also presents a review of methods that adjust statistical analyses for probable record linkage error. The statistical adjustment is a step taken after the other two processing steps in order to improve statistical inference based on the data files produced in the NSIs.

A major issue for the different NSIs is developing teams that can create and maintain the generalized systems. Fellegi-Holt systems have never been commercially available or in shareware (except two edit/imputation systems from Statistics Canada that are available for purchase). Until twenty years ago, record linkage systems were often not commercially available or in shareware. In many situations, the commercial systems can be expensive (as high as one million U.S. dollars), very time consuming to learn, and not easily adaptable for differing data because the source-code is proprietary. Five countries (Canada, Spain, Italy, the Netherlands, U.S.) have separately developed Fellegi-Holt edit/imputation systems. Three countries (Canada, U.S., Italy) have separately developed Fellegi-Sunter record linkage systems. Although the main edit ideas were in the Fellegi-Holt paper (1976), the first fully developed Fellegi-Holt system was from Statistics Canada in 1994. The first fully developed Fellegi-Sunter record linkage system was by Statistics Canada in 1984.

Section 2 provides an overview of the three main technical challenges reviewed in this paper. It also provides brief remarks about the difficulty of developing software, particularly when the initial computational algorithms are one hundred to one thousand times too slow for production (i.e., twenty-to-fifty machines need to run in twelve hours but instead run in twelve hundred hours or longer).

2. Three Illustrative Computational and Statistical Challenges

In this section, three technical examples are given and, because of its importance, an additional subsection with more details on developing systems is provided.

2.1. Example 1

The NSIs through the 1960s primarily performed data collection manually. Enumerators asked questions from survey forms and did minor corrections to the information provided by respondents. Other reviewers systematically went through the collected forms to make additional 'corrections'. The data were keyed into the computers. Processing software, beginning in the 1960s, incorporated hundreds (sometimes thousands) of 'if-then-else' rules into computer code so that 'corrections' were made more quickly and consistently. Problems with the early systems were that there could be logical errors in the specifications of the edits (i.e., an edit such as 'a child under 16 could not be married'), that there were coding errors, and that no systematic methods existed for assuring that joint distributions of variables were preserved in a principled manner.

Starting in the 1980s, five NSIs separately developed 'edit' systems based on the Fellegi-Holt (1976, hereafter FH) model of statistical data editing. The advantage of the FH systems were that edits resided in easily modified tables, the logical consistency of the system could be checked prior to the receipt of data, the main logic resides in reusable routines, and, in one pass through the data, a 'corrected record is guaranteed to satisfy edits. Prior to FH theory, an 'if-then-else' edit system might run correctly on test decks (sets of records with known properties) but fail when non-anticipated edit patterns were encountered during production. Additionally, there might be logic errors in the set of 'if-then-else' rules. NSIs hired individuals with expertise in Operations Research (typically Ph.D.s) who did theory and computational algorithms for the integer programming logic needed to implement FH systems. NSIs in Canada, Spain, Italy, the Netherlands, and the U.S. separately developed general systems. Each was able to demonstrate substantial computational efficiencies with the new systems that were portable across different types of survey data and that often ran on more than one computer architecture.

The 1997 U.S. Economic Censuses (with a total of 8 million records in three files) were each processed in less than twelve hours. Kovar and Winkler (1996) show that the Census Bureau SPEER system maintained the very high quality of Statistics Canada's Generalized Edit/Imputation System while being 60 times as fast (due to new computational algorithms – Winkler 1995b). Because of substantial hardware-speed increases in the last twenty years, most of the NSIs methods/software should be suitable for smaller administrative lists (millions or tens of millions of records). Census Bureau software (Winkler 2013c) should be suitable for hundreds of millions of records or even billions of records.

2.2. Example 2

Record linkage is the methodology of finding and correcting duplicates within a single list or across two or more lists. The means of finding duplicates involve non-unique (quasi-) identifiers such as name, address, date-of-birth, and other fields. Unique keys, such as a verified U.S. Social Security Number, often are not available on all files for all entities without error. Other variables such as name, address, and date-of-birth (called quasi-identifiers) must be used in combination to uniquely identify entities (people, businesses, etc.). Statistics Canada, the Italian National Statistical Institute (ISTAT), and the U.S. Census Bureau developed systems for record linkage based on the Fellegi-Sunter (hereafter FS) ideas. Academic researchers, many of whom were professors of computer science and

funded by various health agencies, developed other systems.

Hogan and Wolter (1984) introduce a capture-recapture procedure (called the Post Enumeration Survey, PES) in which a second list was matched against the Decennial Census to evaluate and potentially correct-for under- and over-coverage. Their procedure had seven components. They believed that the record-linkage component had more error than the other six components combined. Hogan and Wolter estimated that manual matching would require 3000 individuals for six months with a false match error rate of a 5%. Their PES estimation/adjustment procedures required at most 0.5% false match rate.

Instead of relying on manual human-powered record linkage, the FS algorithms as implemented by the U.S. Census Bureau were used. Linkage was performed in about 500 contiguous regions in the U.S. The country was broken into regions to reduce computations and due to the power of computers at that time. In each region, it was necessary to estimate optimal parameters for the record linkage problem, which is a type of unsupervised learning classification. Parameters were estimated largely based on an application of the EM algorithm (Dempster, Laird, and Rubin 1977) viewing record linkage as a missing data and latent class problem.

Using the computerized software, the U.S. Census Bureau accomplished the matching in six weeks. Approximately 200 individuals were needed for clerical matching (using additional information from paper files) and field follow-up. The clerical review region of possible matches consisted almost entirely of individuals within the same household who were missing first name and age (the only two fields for distinguishing individuals within households). The EM-based estimation procedures for obtaining optimal parameters (Winkler 1988, Winkler and Thibaudeau 1991) yield a false match rate of at most 0.2%. This number was calculated based on field follow-up, two rounds of clerical review, and one round of adjudication.

2.3. Example 3

Large national files can yield useful information in statistical, demographic and economic analyses. Combining large national files can enable additional analyses. If there are errors in files or they are merged with some inaccuracies, then analyses might not be completely accurate. Can files be cleaned-up before and after the process of merging them together? Can analyses be adjusted to compensate for possible linkage error?

A conceptual picture would link records in file $\mathbf{A} \times \mathbf{X} = (a_i, \ldots, a_n, x_1, \ldots, x_k)$ with records in file $\mathbf{B} \times \mathbf{X} = (b_1, \ldots, b_m, x_1, \ldots, x_k)$ using common quasi-identifying information $\mathbf{X} = (x_1, \ldots, x_k)$ to produce the merged file $\mathbf{A} \times \mathbf{B} = (a_i, \ldots, a_n, b_1, \ldots, b_m)$ for analyses. The variables $x_1, \ldots, x_k$ are quasi-identifiers such as names, addresses, dates-of-birth, and even fields such as income (when processed and compared in a suitable manner). Individual quasi-identifiers will not uniquely identify correspondence between pairs of records associated with the same entity; sometimes combinations of the quasi-identifiers may uniquely identify. Each of the files **A** and **B** may be cleaned for duplicates and then cleaned up for missing and contradictory data via edit/imputation.

If there are errors in the linkage, then completely erroneous $(b_1, \ldots, b_m)$ may be linked with a given $(a_i, \ldots, a_n)$ and the joint distribution of $(a_i, \ldots, a_n, b_1, \ldots, b_m)$ in $\mathbf{A} \times \mathbf{B}$ may be very seriously compromised. If there is inadequate cleanup (i.e., non-effective edit/imputation) of $\mathbf{A} \times \mathbf{X} = (a_i, \ldots,$

$a_n, x_1, \ldots, x_k)$ and $\mathbf{B} \times \mathbf{X} = (b_1, \ldots, b_m, x_1, \ldots, x_k)$, then analyses may have other serious errors in addition to the errors due to the linkage errors. For instance, each of files **A** and **B** may have 3% duplicates or edit/imputation errors. If there is 3% matching error, then the joint file $\mathbf{A} \times \mathbf{B}$ could have 9% error. If there were 9% error in a file, then how would an NSI determine if any analysis on $\mathbf{A} \times \mathbf{B}$ could be performed? If there were 5% error in either of the files **A** or **B**, how would the NSI determine that the errors existed or how to correct for them? If there were 5% error in the matching between files **A** and **B**. how would the NSI determine that there was too much error and how would the NSI correct for the error?

2.4 Skills Needed for Developing Generalized Systems

Winkler and Hidiroglou (1998) provide an overview of the teams that developed generalized systems. The teams, often consisting of five-fifteen individuals, worked two-four years to get preliminary results for the potentially most difficult parts of a new system. If initial results were promising, the NSIs added more individuals for upwards of ten years to develop the completed systems. By the mid 1990s, NSIs had failed in at least ten attempts of initial development of Fellegi-Holt systems and, by the mid 2010s, NSIs had failed in possibly thirty attempts of the initial development of Fellegi-Sunter systems.

For as long as ten years, Eurostat tried to co-ordinate European Union (EU) NSIs in development. NSIs that had developed their own systems were unwilling (even when Eurostat could provide considerable extra money) to write versions of their systems in different computer languages on completely different computer architectures. The potential users in other NSIs (who typically did not have suitable technical backgrounds) often insisted on considerable new features that they claimed were critical to the needs of their NSI. One NSI that had developed a system pointed out that the requested changes in the general software would have caused the general software to be unsuitable for most of their survey databases. Another NSI (that had developed a general system with a different computer language, different hardware, and completely different OR algorithms) corroborated that making similar changes would have also made their system unusable most of the time. Commercial companies were unwilling to develop systems because Eurostat (initially) wanted the source code for a new system to be available to EU participants. The commercial companies would be unable to copyright/patent their methods/code. It was not clear that resources would be available for maintenance of the software.

The generalized software for production record linkage in the 1990 Decennial Census began in January 1984 and lasted until July 1994. Fifteen individuals did most of the development. For some development and intermediate tests, an additional five-ten individuals were used. The final 1990 production system had thirty-plus modules. Three modules (the main matcher, name standardization, and address standardization) were completely rewritten in 1992-1994 to make the entire set of software much easier to apply in newer applications. Although many programs needed speed improvements of at least a factor of ten, a main parameter estimation program needed a speed improvement of at least 200. Two statisticians created three additional versions of the parameter-estimation software over a year. The final software was 220 times as fast as the first version and was barely fast enough for production (reducing the amount of time in each of ~500 areas from thirty-six hours to at most twenty minutes). Without the speed improvements, the entire system was too slow for production (three-six weeks for the entire U.S. using seven VAX 8700s – about the same speed as a fast 80386 PC but the VAXes were able to address much more memory).

The edit/imputation system for the 1997 U.S. Economic Censuses consisted of twenty-plus programs developed by fifteen individuals over four years. Many of the modules were variants of programs from the 1992 Economic Censuses. The legacy software needed to be ported from Univac to VAXes or rewritten from scratch. The set covering algorithms and integer programming parts of the Fellegi-Holt system were new and written for VAXes (and other Census Bureau computers). Winkler (1995b) develops new integer programming algorithms after deciding that the GEIS (Generalized Edit/Imputation System) software from Statistics Canada was too slow. The GEIS software, however, was faster than the first two versions of the SPEER edit software that Winkler had written. Because Winkler had written some record linkage modules for Statistics Canada, he received access to GEIS source code that he agreed to keep private. Statistics Canada adopted Chernikova algorithms after William Pulleyblank (OR professor at Waterloo) indicated that the Statistics Canada algorithms were slightly faster than algorithms of Pulleyblank's and far faster than standard OR algorithms on three test decks. The new Census Bureau SPEER software (Kovar and Winkler 1996) was sixty times as fast as GEIS and just as highly accurate. SPEER allowed processing each of the three Economic Censuses (containing a total of eight million records) in twelve hours or less.

Each of the Census Bureau systems is portable across IBM PCs, VAXes, and Linux servers. The software, upon recompilation, also runs without modification on MacIntoshes and IBM mainframes. Much of the software is in FORTRAN because the development teams consisted almost entirely of FORTRAN programmers. The remainder of the software is in C. Other NSIs have had to convert their software as they moved to new machines. For instance, Statistics Netherlands converted from Pascal to C++; Statistics Canada converted from PL/I on IBM mainframes to C/C++ for their newer mainframes, Linux servers and IBM PCs.

Sections 3-5 discuss the three challenges: edit/imputation, record linkage, and adjustment of statistical analysis for linkage error. The discussion provides additional background, delineation of methods, and examples. In the case of the third challenge, work in progress and some suggested methods for addressing the issue are described.

3. Edit/Imputation

Fellegi and Holt (1976, hereafter FH) propose a statistical data editing model for edit/imputation primarily for discrete data. The FH paper (Theorem 1) rediscovered one of the fundamental results of logic programming. As implemented at the U.S. Census Bureau, the imputation part of the system uses ideas of filling-in missing data according to extensions of the missing-at-random (*mar*) model of Little and Rubin (2002, Chapter 13) and further extensions by Winkler (2008). It then applies the theory of Winkler (2003) that connects editing with imputation. The computational ideas are set-covering algorithms (Winkler 1997) and three model-building algorithms based on the EMH algorithm (Winkler 1993) that extend the ECM algorithm of Meng and Rubin (1993) from linear to convex constraints. Key algorithmic breakthroughs increased speed of these four computational algorithms by factors of 100-1000. The generalized software (Winkler 2008, 2010) is suitable for use in most survey situations with discrete data and is sufficiently fast for hundreds of millions or billions of records. Without the algorithmic speed increases, the system was unusable (taking weeks/months instead of hours) on the Census Bureau censuses and many of the surveys.

3.1. Edit/Imputation Background

The intent of classical data collection and clean-up is to provide a data file that is free of logical errors and missing data. For a statistical agency, an interviewer might fill out a survey form during a face-to-face interview with the respondent. The 'experienced' interviewer would often be able to 'correct' contradictory data or 'replace' missing data during the interview. At a later time, analysts might make further 'corrections' prior to the data on paper forms being placed in computer files. The purpose is to produce a 'complete' (i.e., no missing values) data file that has no contradictory values in some variables. The final 'cleaned' file would be suitable for various statistical analyses. In particular, the statistical file would allow determination of the proportion of specific values of the multiple variables (i.e., joint inclusion probabilities).

In this section, the presentation covers Fellegi-Holt methods for continuous economic data (SPEER, Winkler 1995b) and for discrete data (DISCRETE, Winkler 1997, 2003, 2008, 2010). The SPEER system has been in use since 1997. The DISCRETE system uses far more difficult general integer programming methods for editing than the SPEER system. It also uses new theory (Winkler 1990c, 1993) and a series of new combinatorial methods for imputation. Most NSIs have developed separate FH methods for continuous and discrete data. Statistics Canada developed GEIS methods for continuous data and CANCEIS (Bankier 2000) methods for discrete data. Winkler and Chen (2002) proved that the methods for CANCEIS are consistent with FH principles. Statistics Netherlands (De Waal et al. 2011) has a hybrid methodology suitable for both discrete and continuous data that is optimal for edit properties but may not be optimal for preserving joint distributions.

Naïvely, dealing with edits is straightforward. If a child of less than sixteen years old is given a marital status of 'married', then either the age associated with the child might be changed (i.e., to older than 16) or the marital status might be changed to 'single'. The difficulty consistently arises that, as a (computerized) record r_0 was changed ('corrected') to a different record r_1 by changing values in fields in which edits failed, the new record r_1 might fail other edits that the original record r_0 had not failed. The interaction between the set of edits yields enormously large combinatorial optimization problems.

In a real-world survey situation, subject matter 'experts' may develop hundreds or thousands of if-then-else rules that are used for the editing and hot-deck imputation. Because it is exceptionally difficult to develop the logic for such rules, most edit/imputation systems do not assure that records satisfy edits or preserve joint inclusion probabilities. Further, such systems are exceptionally difficult to implement because of (1) logic errors in specifications, (2) errors in computer code, and (3) no effective modeling of hot-deck matching rules. As demonstrated by Winkler (2008), it is effectively impossible with the methods (classical if-then-else and hot-deck) that many agencies use to develop edit/imputation systems that preserve either joint probabilities or to create records that satisfy edit restraints. This is true even in the situations when Fellegi-Holt methods are used for the editing and hot-deck is used for imputation.

An edit/imputation system that effectively uses the edit ideas of Fellegi and Holt (1976) and modern imputation ideas (such as in Little and Rubin 2002) has distinct advantages. First, it is far easier to implement (as demonstrated in Winkler 2008). Edit rules are in easily modified tables. The logical consistency of the entire system can be tested automatically according to the mathematics of the Fellegi-Holt model and additional requirements on the preservation of joint inclusion probabilities (Winkler 2003). Second, the optimization that determines the minimum number of fields to change or replace in an edit-failing record is in a fixed mathematical routine that does not need to change. Third,

imputation is determined from a model (limiting distribution) and automatically preserves joint distributions. Most modeling is very straightforward. It is based on variants of loglinear modeling and extensions of missing data methods that are contained in easily applied, extremely fast computational algorithms (Winkler 2006b, 2008; also 2010). The methods create records that *always* satisfy edits and preserve joint inclusion probabilities.

3.2. Fellegi-Holt Model

Fellegi and Holt (1976) were the first to provide an overall model to assure that a changed record r_1 would not fail edits. Their theory requires the computation of all implicit edits. Implicit edits are edits that can be logically derived from an originally specified set of 'explicit' edits. If the implicit edits are available, then it is always possible to change an edit-failing record r_0 to an edit passing record r_1. The availability of 'implicit' edits makes it quite straightforward and fast to determine the minimum number of fields to change in an edit-failing record r_0 to obtain an edit-passing record r_1 (Barcaroli and Venturi 1997). In particular, the implicit edits drastically speed up standard Integer Programming algorithms such as branch-and-bound by quickly forcing solutions into more suitable regions. Fellegi and Holt indicated how hot-deck might be used to provide the values for filling in missing values or replacing contradictory values.

3.3. Imputation Generalizing Little-Rubin

The intent of filling-in missing or contradictory values in edit-failing records r_0 is to obtain records r_1 that can be used in computing the joint probabilities in a principled manner. The difficulty observed by many individuals is that a well-implemented hot-deck does not preserve joint probabilities. Rao (1997) provided a theoretical characterization of why hot-deck fails even in two-dimensional situations. The failure occurs even in 'nice' situations where individuals had previously assumed that hot-deck would work well. Other authors in the 1980s had shown empirically that hot-deck fails to preserve joint distributions. Andridge and Little (2010) provide a review of hot-deck-related literature.

Little and Rubin (2002, Chapter 13) provide a model for filling-in missing data according to the missing-at-random (*mar*) data. The modeling methods can be considered as a correct extension of 'hot-deck'. With 'hot-deck' a record with missing values is matched against the entire set of *complete* records that have no missing data. In practice, NSIs might have a record with twenty variables of which five had missing values. Typically, there would be no *donors* (complete data records that match on the fifteen non-missing values. To find matches among the set of complete data records, NSIs developed collapsing rules that would take subsets of the fifteen variables to match on. In extreme (but typical) situations, collapsing would continue until only 2-4 variables were used for matching. In all situations with collapsing, joint distributions of the final data were substantially distorted. Most NSIs have not developed systematic, efficient methods for collapsing.

3.4. Connecting Edit with Imputation

With classical edit/imputation in the NSIs, the logic in the software consists of a number (hundreds or thousands) of 'if-then-else' rules where values in fields in a record *r* associated with failing edits are replaced with values that would cause the 'corrected' record to no longer fail edits. With the development of FH systems, the NSIs, for the first time, can assure that a 'corrected' record will no longer fail edits. Because the 'substitutions' of values due to editing were not designed to assure that joint distributions were preserved, statisticians later developed imputation models for filling in missing

values or replacing 'erroneous' values associated with failing edits that assured joint distributions were preserved. Some 'corrected' records did not satisfy edits because the imputation methods did not effectively account for the entire set of edit restraints.

Winkler (2003) provides the first theory that connects imputation with the entire set of edit restraints. Winkler (1997) had earlier provided set-covering algorithms that were at least one hundred times as fast as set-covering algorithms that IBM developed for the Italian Labour Force Survey (Barcaroli and Venturi 1997 and private communications with the authors). With suitably fast set-covering algorithms all implicit edits were available prior to building imputation models that systematically and globally assures that all records passing through the system would satisfy all edits and, in aggregate, preserve joint distributions. Lemma 1 (Winkler 1990c) shows how structural zeros (edit restraints) in a contingency table could be consistent with the other restraints in a contingency table such as interactions and certain convex constraints. The theory assures that, if individuals created suitable structures using implicit edits in just the right manner, then imputation can be made consistent with the edit restraints.

Although the theory assured that it was possible to develop generalized systems, the previously existing methods for doing the computation associated with the methods of Chapter 13 in Little and Rubin (2002) were 10^3-10^6-times too slow for use on national files or large administrative lists.

3.5. Substantially Decreased Development Time and Improved Accuracy

The generalized software yields significant time-savings and dramatically increased accuracy. Prior to the FH systems, each edit/imputation system (in all NSIs) was written from scratch each time a survey form was changed slightly. Winkler and Petkunas (1997) showed that the edit part of the system could each be created in less than four hours for three different small survey systems. Winkler (2008) indicates that the imputation part of the system can be created in less than one week with most Census Bureau demographic surveys. This a dramatic reduction for most small-to-intermediate size surveys where as many as ten analysts (subject matter specialists) and five programmers took a year to create the new version of the system. The main computational algorithms in a FH system are constant across different survey applications. The software does not contain the numerous errors in contrast to when systems are rewritten from scratch.

3.6. Achieving Extreme Computational Speed

The set covering algorithms (Winkler 1997) were 100-1000 times as fast as algorithms developed by IBM for the Italian Labour Force Survey at ISTAT (Barcaroli and Ventura 1997 and private communications with the authors). Barcaroli and Ventura provided the Italian Labour Force data for advanced testing of later versions of the algorithms of Winkler (1997). The data were non-confidential because they consisted of edit restraints. The key to the speed of the set covering algorithms was certain data structures and software loops that allowed the computational algorithms to learn both likely successful and likely unsuccessful computational paths. In the initial testing, computation on the Decennial Census short-form implicit edits dropped from 24 hours to six minutes. Subsequent testing on the Decennial Census long-form (still shorter than the Italian Labour Force data) allowed computation of implicit edits in six hours. Based on additional testing at the Census Bureau and ISTAT with the Italian Labour Force data, the new set covering algorithms took 6-24 hours on a 200 Mhz Pentium Pro PC. The IBM algorithms might have taken 800+ days on the largest IBM mainframe (based on testing with smaller subsets of the Italian Labour Force data and e-mail

communications with ISTAT). It was more than ten years (Winkler 2003, 2008, 2010) before theory connecting edit and imputation were developed and a series of computational algorithms were created that were suitable for imputation in many of the largest situations.

The current loglinear modeling algorithms (Winkler 2008, 2010) do the EM fitting as in Little and Rubin (2002) but with computational enhancements that scale subtotals exceedingly rapidly and with only moderate use of memory. The algorithms also deal with the edit restraints (Winkler 2003, 1993, 1990c). The computational speed for a contingency table of size 600,000 cells is 45 seconds on a standard Intel Core I7 machine and for a table of size 2.0 billion cells is approximately 500 minutes (each with epsilon 10^{-12} and 200 iterations). In the larger applications, 16 Gb of memory are required. The key to the speed is the combination of effective indexing of cells and suitable data structures for retrieval of information so that each of the respective margins of the M-step of EM-fitting are computed rapidly. Naive algorithms in SAS or R programs might not converge to a solution in the 600,000-cell situation in less than one week. Based on extrapolation in the 2-billion-cell situation, the new EM algorithms might take five days on a standard Census Bureau server with Intel chips while taking centuries with SAS Proc Catmod.

A nontrivially modified version of the indexing algorithms allows near instantaneous location of cells in the contingency table that match a record having missing data. This is the key logical extension of Little and Rubin (2002, Chapter 13). An additional algorithm nearly instantaneously constructs an array that allows binary search to locate the cell for the imputation (for the two algorithms: total < 2.0 milliseconds cpu time). For instance, if a record has 12 variables and 5 have missing, it is likely to need to delineate all 100,000+ cells in a contingency table with 0.5 million or 0.5 billion cells and then draw a cell (donor) with probability-proportional-to-size (pps) to impute missing values in the record with missing values. This type of imputation assures that the resultant 'corrected' microdata have joint distributions that are consistent with the model. A naively written SAS search and pps-sample procedure might require as much as a minute cpu time for each record being imputed.

4. Record Linkage

Record linkage is the methodology of finding and correcting duplicates within a single list or across two or more lists via quasi-identifiers. In the first subsection, background on the Fellegi-Sunter model of record linkage is provided. In the second subsection, a method for estimating 'optimal' parameters is presented; it was used for approximately five hundred regions without training data (unsupervised learning) in a production system for the 1990 Decennial Census. In the third subsection methods for estimating false match rates without training data are shown. The methods are not as accurate as the semi-supervised methods of Larsen and Rubin (2001) and Winkler (2002). The fourth subsection focuses on how one can achieve extreme computational speed suitable for sets of national files and administrative lists with hundreds of millions or billions of records. The last two sections deal with data quality and with name and address standardization. The name and address standardization can have a greater effect on matching efficacy than the parameter-estimation methods.

4.1. Fellegi-Sunter Model

Fellegi and Sunter (1969) provide a formal mathematical model for ideas that had been introduced by Newcombe et al. (1959, 1962). They introduced many ways of estimating key parameters without training data. To begin, notation is needed. Two files **A** and **B** are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files **A** and **B** into M, the set of true matches, and U, the set of

true nonmatches. Fellegi and Sunter, making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \quad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith" and "Zabrinsky" occur. Then P(agree "Smith" | M) < P(agree last name | M) < P(agree "Zabrinsky" | M) which typically gives a less frequently occurring name like "Zabrinsky" more distinguishing power than a more frequently occurring name like "Smith" (Fellegi and Sunter 1969, Winkler 1995a). Somewhat different, much smaller, adjustments for relative frequency are given for the probability of agreement on a specific name given U. The probabilities in (1) can also be adjusted for partial agreement on two strings because of typographical error (which can approach 50% with scanned data (Winkler 2004)) and for certain dependencies between agreements among sets of fields (Larsen and Rubin 2001, Winkler 2002). The ratio R or any monotonely increasing function of it such as the natural log is referred to as a *matching weight* (or score).

The decision rule is given by:

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

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

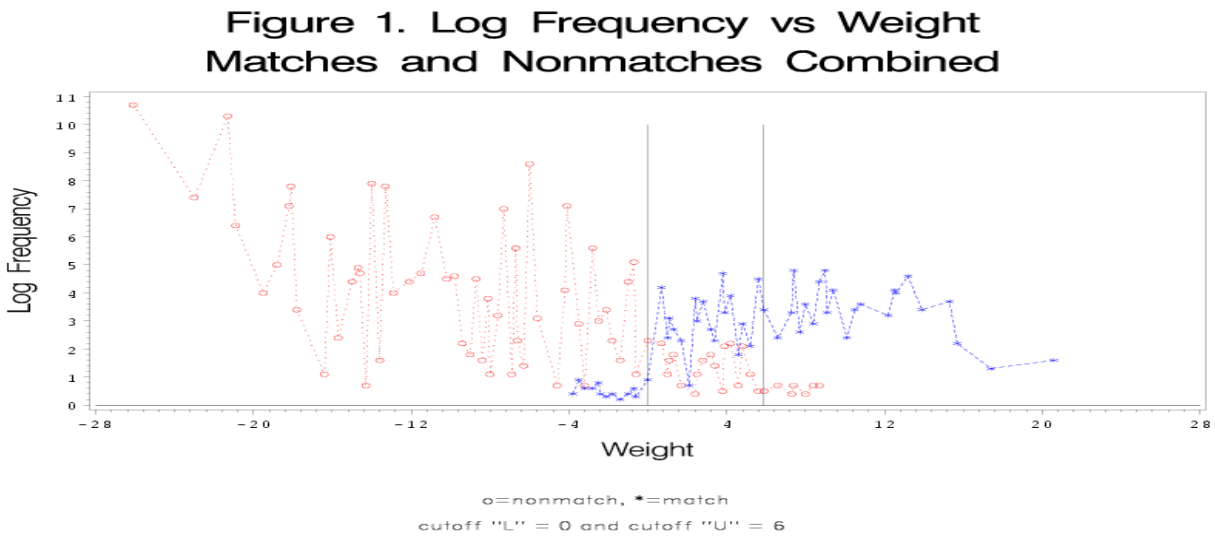
If $R < T_\lambda$, then designate pair as a nonmatch.

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint subregions. The region $T_\lambda \le R \le T_\mu$ is referred to as the *no-decision region* or *clerical review* region. In some situations, resources are available to review pairs clerically. If no resources are available for clerical review, then the cutoffs T_μ and T_λ can be chosen equal (Herzog et al. 2007).

Fellegi and Sunter (1969, Theorem 1) proved the optimality of the classification rule given by (2). Their proof is very general in the sense in it holds for any representations $\gamma \in \Gamma$ over the set of pairs in the product space $\mathbf{A} \times \mathbf{B}$ from two files. As they observed, the quality of the results from classification rule (2) were dependent on the accuracy of the estimates of P($\gamma \in \Gamma$ | M) and P($\gamma \in \Gamma$ | U).

Figure 1 provides an illustration of the curves of log frequency versus log weight for matches and nonmatches, respectively. The two vertical lines represent the lower and upper cutoffs thresholds T_λ

(denoted by L in the Figure) and T_μ, (denoted by U) respectively. The x-axis is the log of the likelihood ratio R given by (1). The y-axis is the log of the frequency counts of the pairs associated with the given likelihood ratio. The plot uses pairs of records from a contiguous geographic region that were matched in the 1990 Decennial Census. The clerical review region between the two cutoffs primarily consists of pairs within the same household that are missing both first name and age (the only two fields that distinguish individuals within a household).



4.2. Estimating Parameters

This section summarizes current extensions of the EM procedures for estimating false match rates. With any matching project, a primary concern is with false match rates among the set of pairs among designated matches above the cutoff score T_μ in (2) and the false nonmatch rates among designated nonmatches below the cutoff score T_λ in (2). Very few matching projects estimate these rates although valid estimates are crucial to understanding the usefulness of any files obtained via the record linkage procedures (see section 4.3 below).

If a small amount of training data is available, then it may be possible to improve record linkage and provide good estimates of error rates. Larsen and Rubin (2001) combined small amounts of (labeled) training data with large amounts of unlabeled data to estimate error rates using an MCMC procedure. In machine learning (Winkler 2000), the procedures are referred to as *semi-supervised learning*. In ordinary machine learning, the procedures to get parameters are "supervised" by the training data that is labeled with the true classes into which later records (or pairs) will be classified. Winkler (2002) also used semi-supervised learning with a variant of the general EM algorithm.

Both the Larsen and Rubin (2001) and Winkler (2002) methods were effective because they accounted for interactions between the fields and were able to use labeled training data that was concentrated between the lower cutoff T_λ and the upper cutoff T_μ. Winkler (1988) observed that the conditional independence model (naive Bayes) in the paper was an approximation of a general interaction model. Winkler (2014, also Herzog et al. 2010) provided more complete explanations of why the 1988 procedure was a best naïve Bayes approximation of a general interaction model. The

model-approximation methods were re-discovered by Kim Larsen (2005) who used much-slower-to-compute general-additive models.

Because the EM-based methods of this section serve as a template of other EM-based methods, the following provides details of the unsupervised learning methods of Winkler (2006a) that are used for estimating the basic matching parameters and, in some situations, false match rates. The basic model is that of semi-supervised learning that combines a small proportion of labeled (true or pseudo-true matching status) pairs of records with a very large amount of unlabeled data. The conditional independence model corresponds to the naïve Bayesian network formulization of Nigam et al. (2000). The more general formulization of Winkler (2000, 2002) allows interactions between agreements (but is not used in this paper).

The development is similar theoretically to that of Nigam et al. (2000). The notation differs very slightly because it deals more with the representational framework of record linkage. Let γ_i be the agreement pattern associated with pair p_i. Classes C_j are an arbitrary partition of the set of pairs D that is a subset in $\mathbf{A} \times \mathbf{B}$. Later, the assumption is that some of the C_j will be subsets of M and the remaining C_j are subsets of U. Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns γ_i may have many pairs $p_{i(l)}$ associated with them. In record linkage, the parameter Θ generally represents a Dirichlet-Multinomial model. Specifically,

$$P(\gamma_i \mid \Theta) = \sum_i^{|C|} P(\gamma_i \mid C_j; \Theta)\, P(C_j ; \Theta) \qquad (3)$$

where (i is a specific pair, C_j is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or conditional independence (**CI**), the following equality holds

$$P(\gamma_i \mid C_j ; \Theta) = \Pi_k \quad P(\gamma_{i,k} \mid C_j ; \Theta) \qquad (4)$$

where the product is over the k^{th} individual field agreement γ_{ik} in pair agreement pattern γ_i. In some situations, a Dirichlet prior is used

$$P(\Theta) = \Pi_j \, (\Theta_{Cj})^{\alpha-1} \qquad \Pi_k \quad (\Theta \gamma_{i,k \mid Cj})^{\alpha-1} \qquad (5)$$

where the first product is over the classes Cj and the second product is over the fields (quasi-identifiers such as first name, last name, age, etc.). The symbol D_u to denotes unlabeled pairs and D_l to denote labeled pairs. Given the set D of all labeled pairs D_l and unlabeled pairs D_u that partition D, the log likelihood is given by

$$\begin{aligned} l_c(\Theta \mid D; z) = \quad & \log (P(\Theta)) \quad + \\ & (1-\lambda) \sum_{i \in Du} \quad \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \quad P(C_j ; \Theta)) + \\ & \lambda \sum_{i \in Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta) \quad P(C_j ; \Theta)). \end{aligned} \qquad (6)$$

where $0 \leq \lambda \leq 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs. In the third terms of equation (6), the sum is over the observed z_{ij}. The estimation is using the EM algorithm (Winkler 1993, 2002). In the second term, expected values for the z_{ij} based on the initial estimates $P(\gamma_i \mid C_j; \Theta)$ and $P(C_j; \Theta)$ are put in (i.e., the E-step in the EM algorithm). After

re-estimating the parameters $P(\gamma_i \mid C_j; \Theta)$ and $P(C_j; \Theta)$) during the M-step (that is in closed form under condition (**CI**)), new expected values are put in and the M-step is repeated. The computer algorithms are easily monitored by checking that the likelihood increases after each combination of E- and M-steps and by checking that the sum of the probabilities add to 1.0. If λ is 1, then only training data are used and the methods correspond to naïve Bayes methods in which training data are available. If λ is 0, then the situation is for unsupervised learning situations of Winkler (1988, 2006a). Winkler (2000, 2002) provides more details of the computational algorithms.

4.3. Estimating False Match Rates

This section provides a summary of current extensions of the EM procedures for estimating false match rates. To estimate false-match rates, it is possible to obtain reasonably accurate estimates of the right tail of the curve of nonmatches and the left tail of the curve of matches such as given in Figure 1. Similar methods, but with less data, were given in Belin and Rubin (1995). With any matching project, the concern is with false match rates among the set of pairs among designated matches above the cutoff score T_μ in (2) and the false nonmatch rates among designated nonmatches below the cutoff score T_λ in (2). Very few matching projects estimate these rates although valid estimates are crucial to understanding the usefulness of any files obtained via the record linkage procedures. Sometimes, reasonable upper bounds for the estimated error rates can be obtained by experienced practitioners and the error rates are validated during follow-up studies (Winkler 1995a). If a moderately large amount of training data is available, then it may be possible to get valid estimates of the error rates using training data only.

Belin and Rubin (1995) were the first to provide an unsupervised method for obtaining estimates of false match rates. The method proceeded by estimating Box-Cox transforms that would cause a mixture of two transformed normal distributions to closely approximate two well separated curves such as given in Figure 1. They cautioned that their methods might not be robust to matching situations with considerably different types of data. Winkler (1995a) observed that their algorithms would typically not work with business lists, agriculture lists, and low quality person lists where the curves of nonmatches were not well separated from the curves of matches. Scheuren and Winkler (1993), who had the Belin-Rubin EM-based fitting software, observed that the Belin-Rubin methods did work reasonably well with a number of well-separated person lists.

The Data Files

Three pairs of files were used in the analyses. The files are from 1990 Decennial Census matching data in which the entire set of 1-2% of the matching status codes that were believed to have been in error for these analyses have been corrected. The corrections reflect clerical review and field follow-up that were not incorporated in computer files originally available to us.

A summary of the overall characteristics of the empirical data is in Table 1. In the following, the totals only consider pairs that agree on census block id (small geographic area representing approximately 50 households) and on the first character of surname. Less than 1-2% of the matches are missed using this set of blocking criteria. They missing matches (pairs) represent a fixed lower bound on the estimated false match rates. They are not considered in the analysis of this paper.

Table 1. Summary of Three Pairs of Files

	Files		Files		Files	
	A1	A2	B1	B2	C1	C2
Size	15048	12072	4539	4851	5022	5212
# pairs	116305		38795		37327	
# matches	10096		3490		3623	

The matching fields are:

Person Characteristics: First Name, Age, Marital Status, Sex
Household Characteristics: Last Name, House Number, Street Name, Phone

Typically, everyone in a household will agree on the household characteristics. Person characteristics such as first name and age help distinguish individuals within household. Some pairs (including true matches) have both missing first name and age.

In the following, various partial levels of agreement in which the string comparator values are broken out as [0, 0.66], (0.66, 0.88], (0.88, 0.94], and (0.94, 1.0]. The intervals were based on knowledge of how string comparators were initially modeled (Winkler 1990b, also 2006c) in terms of their effects of likelihood ratios (1). The string comparators take values between 0 – total disagreement and 1 – exact character-by-character agreement. The first interval is referred to as disagreement. The disagreement is combined with the three partial agreements and blank to get five value states (base 5). The large base analyses consider five states for all characteristics except sex and marital status; the smaller base considers only three (agree/blank/disagree). The total number of agreement patterns is 140,625. In the earlier work (Winkler 2002), the five levels of agreement worked consistently better than two levels (agree/disagree) or three levels (agree/blank/disagree).

The pairs naturally divide into three classes: C_1- match within household, C_2 - nonmatch within household, C_3 – nonmatch outside household. In the earlier work (Winkler (2002) considered two dependency models in addition to the conditional independence model. In that work in which small amounts of labeled training data were combined with unlabeled data, the conditional independence model worked well and the dependency models worked slightly better. Newcombe and Smith (1975) and later by Gill (2001) first used the procedures for dividing the matching and estimation procedures into three classes without the formal likelihood models given by equations (3)-(6).

For data sets, ‘pseudo-truth’ is created in which matches are those unlabeled pairs above a certain high cutoff and nonmatches are those unlabeled pairs below a certain low cutoff. Figure 1 illustrates the situation using actual 1990 Decennial Census data in which log of the probability ratio (1) is plotted against the log of frequency. With the datasets of this paper, high and low cutoffs in a similar manner are chosen that do not include in-between pairs in the designated ‘pseudo-truth’ data sets. These ‘designated’ pseudo-truth data sets are used in a semi-supervised learning procedure that is nearly identical to the semi-supervised procedure where actual truth data was used. A key difference from the corresponding procedure with actual truth data is that the sample of labeled pairs is concentrated in the difficult-to-classify in-between region where, in the ‘pseudo-truth’ situation, there is no way to

designate comparable labeled pairs. The sizes of the 'pseudo-truth' data are given in Table 2. The errors associated with the artificial 'pseudo-truth' are given in parentheses following the counts. The *Other* class gives counts of the remaining pairs and proportions of true matches that are not included in the 'pseudo-truth' set of pairs of 'Matches' and 'Nonmatches'. In the *Other* class, the proportions of matches vary somewhat and would be difficult to determine without training data.

Table 2. 'Pseudo-Truth' Data with Actual Error Rates

	Matches	Nonmatches	*Other*
A pairs	8817 (.008)	98257 (.001)	9231 (.136)
B pairs	2674 (.010)	27744 (.0004)	8377 (.138)
C pairs	2492 (.010)	31266 (.002)	3569 (.369)

To determine how accurately it is possible to estimate the lower cumulative distributions of matches and the upper cumulative distribution of nonmatches. This corresponds to the overlap region of the curves of matches and nonmatches. If it is possible to accurately estimate these two tails of distributions, then it is possible to accurately estimate error rates at differing levels. The comparisons consist of a set of figures with different plots of the cumulative distribution of estimates of matches versus the true cumulative distribution with the truth represented by the 45 degree line. As the plots get closer to the 45 degree lines, the estimates get closer to the truth. The plotting is only for the bottom 30% of the curves given in Belin and Rubin (1995, Figures 2, 3). Generally, the only concern is with the bottom 10% of the curves for the purpose of estimating false match rates. Because of the different representation with the 45 degree line, it is much better for comparing three different methods of estimation for false match rates.

The primary results are from using the conditional independence model and 'pseudo-semi-supervised' methods of this section with the conditional independence model and actual semi-supervised methods of Winkler (2002). With the 'pseudo-truth' data, the best sets of estimates of the bottom 30% tails of the curve of matches with conditional independence and λ=0.2 (equation (6)) can be obtained. Figure 2a,b,c illustrates the set of curves that provide quite accurate fits for the estimated values of matches versus the truth. The 45 degree line represents the truth whereas the curve represents the cumulative estimates of matches for the right tails of the distribution. The plots are for the estimates of the false match probabilities divided by the true false match probabilities. Other at results were considered for λ=0.1, 0.5, and 0.8 and various interactions models. The results under conditional independence (**CI**) were the best with λ=0.2 (equation (6)). Several different ways of constructing the 'pseudo-truth' data were also looked at. Additionally, other pairs of files in which all of the error-rates estimates were better (closer to the 45 degree line) than those for the pair of files given in Figure 2 were considered.

Figures 2d,e,f are the corresponding curves using the methods of Belin and Rubin (1995). The curves are substantially farther from the 45 degree lines because they are only using the distributions of weights (natural logs of likelihood ratio (1)). Using the detailed breakout of the string comparator values, the fact that the 3-class EM (Winkler 1993) provides much better estimates, and the breakout available from equation (6), it is possible to use more information that allows the estimated curves in Figure 2a,b,c to be closer to the truth than the corresponding curves in Figure 2d,e,f.

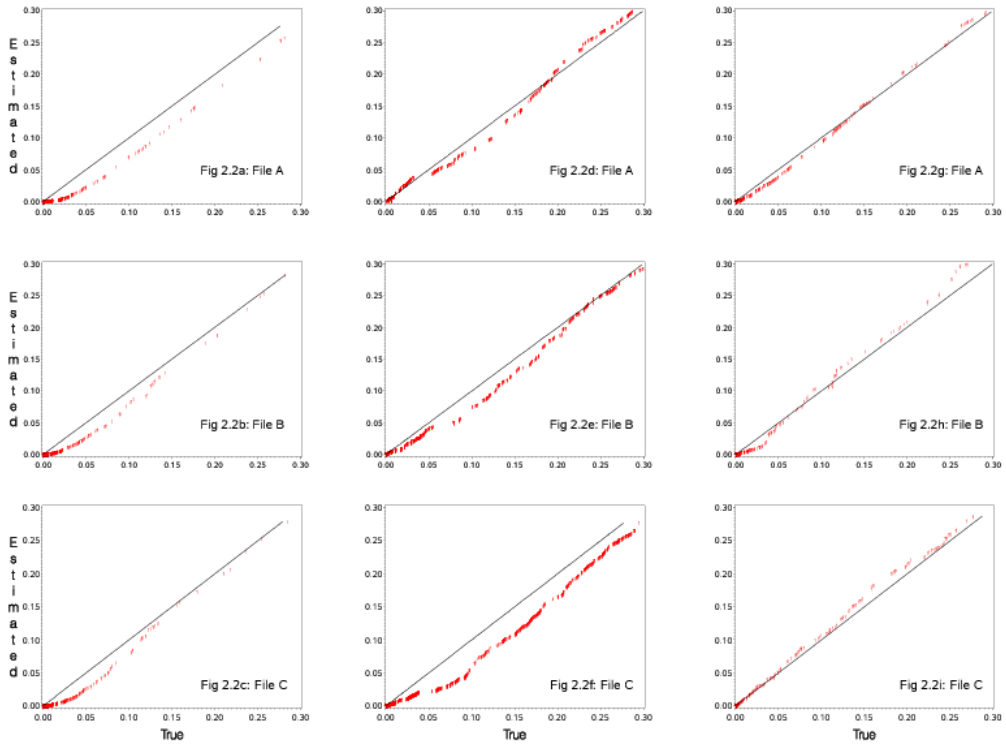


Figure 2. Comparison of Error Rate Estimation Procedures of Winkler (2006, first column, unsupervised), Belin and Rubin (1995, second column, unsupervised), and Winkler (2002, third column, semi-supervised).

The final sets of curves (Figure 2g,h,i) are similar to the semi-supervised learning of Winkler (2002) that achieved results only very slightly worse than Larsen and Rubin (2001) but for which the EM computational speeds (10 minutes in each of ~500 regions) were at least 100 times as fast as the MCMC methods of Larsen and Rubin. It is difficult for the unsupervised methods to perform as well as the semi-supervised methods because the relatively small sample can be concentrated in the clerical review region between the lower cutoff T_λ and the upper cutoff T_μ. Because underlying truth data were available, in some regions only 1/40 of the 'truth' sample was truly a match whereas in other regions 1/10 of the 'truth' sample was truly a match. In the 1990 Decennial Census, the clerical region consisted almost entirely of individuals within the same household who were missing both first name and age (the only two fields for distinguishing within the household). Because of the requirement to match all 457 regions of the U.S. in 3-6 weeks to provide estimates required by law, it was not possible to perform clerical review in each region or use approximations across certain regions because the optimal parameters vary significantly from region to region (Winkler 1989a).

4.4. Achieving Extreme Computational Speed

Early record linkage (Newcombe 1959, 1962) had difficulty dealing with the computation associated with files having a few thousand records each. To deal with the computational burden, individuals used *blocking*. With blocking, individuals sorted two files on a key such as a postal code and first character of surname. Only pairs agreeing on the sort (blocking) key are considered. To deal with situations where there are typographical error in either the postal code or first character of the surname. A second blocking pass may use part of the postal code and house number.

By the 1990s, NSIs might be working with millions of records to which they might apply 5-10 blocking passes. For the first blocking pass, individuals sorted the two files on the first blocking criteria and then performed matching. Prior to a second pass, individuals created residual files from the two passes after removing the records associated with pairs that were believed to be matches. Individuals sorted the two residual files according to the second blocking criteria and the matching repeated. With five blocking passes, individuals need to read and process each file fourteen times. Individuals needed to sort each of the two files five times. With a large administrative file with 0.5 billion records, ~3.5 times the size of an original file was needed for each sort. The sort might take 12+ hours. Processing of a pair of files of 0.5 billion records each with five blocking passes might take two-eight weeks. Often the matching could not be performed because too much disk space was needed.

To alleviate the situation and very significantly speed up processing, Yancey and Winkler (2004, 2009) created BigMatch technology. With BigMatch, there is only one pass against each file. There is no need to sort any of the files. One file is read (a smaller file then can be broken into small subparts) into memory with suitable indices created according to the blocking criteria. The suitable indices are in memory. The creation of ten sets of indices with each corresponding to blocking criteria takes about the same amount of time as a single sort of the file. Each record in the larger file is read only once. All pairs corresponding to all the blocking passes are written into appropriate files. The U.S. Decennial Census is broken into ~500 contiguous regions. The overall matching strategy is controlled by Linux scripts. Each region is matched against ~500 regions. This natural, but crude, parallelization can easily extend to administrative lists.

Winkler (2004) provides some methodological insights into BigMatch and for putting bounds on the false nonmatch rates. Winkler et al. (2010) provide details related to the 2010 Decennial Census matching system that processed 10^{17} pairs (300 million × 300 million) using four blocking criteria. The four blocking criteria were determined by three individuals who had developed better blocking criteria than Winkler (2004). With the four blocking criteria detailed processing was performed on only 10^{12} pairs. Auxiliary testing in 2005 and 2006 using 2000 Decennial Census data provided estimates of properly matching 97.5% of all duplicates.

The BigMatch software ran in 30 hours using 40 cpus of an SGI Linux machine with 2006 Itanium chips. Each cpu processed 400,000+ pairs per second. Most new High Performance Computer (HPC) machines would be 2-8 times as fast due to more memory, more cpus, and faster cpus. Based on the number of pairs processed within a given time, BigMatch is 50 times as fast as parallel PSwoosh software (Kawai et al. 2006) developed by computer scientists at Stanford. BigMatch is 10 times as fast as recent Dedoop software (Kolb and Rahm 2013) that uses sophisticated load balancing methods. The later software and two other sets of parallel software (Yan et al. 2013, Karapiperis and Verykios 2014) would not have been available for the 2010 Decennial Census. BigMatch is 500 times

as fast as software used in some statistical agencies (Wright 2010).

4.5. Lack of Quality in Files

The definition ‘good quality’ is that almost all quasi-identifiers are not missing and that the quasi-identifiers must be in suitably comparable forms. In combination, the set of quasi-identifiers should be able to delineate most matches accurately. The following examples of ‘errors’ in seemingly ok data are very difficult to overcome.

Following the 1990 Decennial Census, several small tests of matching administrative records from two localities were performed in which lists were matched against the 1990 Census files. Laws were changed to allow the matching and review. In one instance, a school board list was compared against the Census. Being able to count children under fifteen has often been difficult. For one school, half the students could not be matched. Review showed that in half of the classrooms (students were broken out by school and classroom), every first name, every last name, and every date-of-birth contained typographical error; in the other half of the classrooms, every first name, every last name, and every date-of-birth appeared to have been correctly keyed into the computer. The assumption was that two individuals had keyed in the data. It was not possible to overcome severe errors of this type in representation of the quasi-identifiers associated with the students.

The following provides an example of where quasi-identifiers are severely in error. One record in one file contained ‘Susan K Smith, 950901’ and the other record in the other file contained ‘Karen Jones, 780322’ where the second quasi-identifier is the date-of-birth in YYMMDD format. The first first name is correct. The second first name of Karen is the name that the woman uses most of the time. The first last name is her maiden name. The second last name is her name that was changed (possibly after marriage). The second date-of-birth is completely wrong. The completely wrong date-of-birth can occur when the date-of-birth is copied from an adjacent line on paper such in the situation with voter registration files and databases. With some pairs of lists only a small portion of representations of the quasi-identifiers will have these types of errors. With other lists, twenty-plus-percent of the representations will have these types of errors (making accurate matching effectively impossible).

4.6. Name and Address Standardization

Most of the methods of preprocessing files prior to parameter-estimation and matching involve name and address standardization. If there is 5% error in name standardization in File **A** and in File **B** and there is 5% error in address standardization in File **A** and in File **B**, then the number of true matches erroneously missed during matching could be as much as 20%. These errors can be by far the largest source of error when matching ‘good quality’ files. Only the name standardization is described. The address standardization has substantial similarity.

The name standardization software will break up a free-form name into components (called parsing) and sometimes change commonly occurring words to a consistent spelling. The quasi-identifier ‘Reverend John K Smoth Junior DD’ might have each of the separate words (components) put in locations so that the corresponding components from different records can be compared. The words ‘Reverend’ and ‘Junior’ might be replaced by consistent abbreviations ‘Rev’ and ‘Jr’. With another record ‘John Smith’ it might be possible to pull off the first name ‘John’ could be compared with the first name ‘John’. The name ‘Smith’ could be compared with the last name ‘Smoth’ (that has typographical error). String comparator metrics (Winkler 1990b) could be used to compare the pairs

of strings that have typographical error. Between 1988 and 1994, almost all the improvements in Census matching were due to the improvements in standardization. The basic parameter-estimation methods described in the first four subsections of this section did not need improvement.

For the 1990 Decennial Census production, there were several modules for name and address standardization that were felt to be suitable for production matching at the time. By 1994, the Census Bureau had completely rewritten and substantially extended the name and address standardization software. The address standardization was formally compared to the two best-selling commercial address standardization packages using a test deck of twenty thousand control addresses and eighty thousand addresses that were difficult to standardize. Whereas the two commercial address standardizers were unable to exceed a fifty percent standardization rate, the Census Bureau standardization exceeded a ninety-eight percent standardization rate. Ad hoc tests showed that the Census Bureau name standardizer exceeded alternative name standardizers.

5. Models for Adjusting Statistical Analyses for Linkage Error

An early model for adjusting a regression analysis for linkage error is due to Scheuren and Winkler (1993). By making use of the Belin-Rubin predicted false-match rates, Scheuren and Winkler were able to give (somewhat crude) estimates of regressions that had been adjusted for linkage error to correspond more closely with underlying 'true' regressions that did not need to account for matching error. Many papers subsequent to Scheuren and Winkler have assumed that accurate values of false match rates (equivalently true match probabilities) are available for all pairs in $\mathbf{A} \times \mathbf{B}$. The difficulty in moving the methods into practical applications is that nobody has developed suitably accurate methods for estimating all false match rates for all pairs in $\mathbf{A} \times \mathbf{B}$ when no training data are available.

For additional information, Reiter (2018) provides an overview on adjusting statistical analyses for linkage error and some elementary examples.

Lahiri and Larsen (2005) later extended the model of Scheuren and Winkler with a complete theoretical development. In situations where the true (not estimated) matching probabilities were available for all pairs, the Lahiri-Larsen methods outperformed Scheuren-Winkler methods and were extended to more multivariate situations than the methods of Scheuren and Winkler (1993). Variants of the models for continuous data are due to Chambers (2009) and Kim and Chambers (2012a,b) using estimating equations. The estimating equation approach is highly dependent on the simplifications that Chambers et al. made for the matching process.

Trancredi and Liseo (2011) applied Bayesian MCMC methods to discrete data. The Trancredi-Liseo methods are impressive because of the number of simultaneous restraints with which they can deal. The Trancredi-Liseo methods are extraordinarily compute intense (possibly requiring as much as 12 hours computation on each block (approximately 50-100 households). There are eight million blocks in the U.S where people live.

Goldstein et al. (2012) provide MCMC methods for adjusting analyses based on very general methods and software that they developed originally for imputation (Goldstein et al. 2009). They provide methods of estimating the probabilities of pairs based on characteristics of pairs from files that have previously been matched. They are able to leverage relationships between vector $x \in \mathbf{A}$ to $y \in \mathbf{B}$ based on a subset of pairs on which matching error is exceptionally small and then extend the

relationships/matching-adjustments to the entire set of pairs in **A × B**. The Goldstein et al. (2012) methods are highly promising, possibly in combination with other methods. It is not clear that it is possible to encounter situations similar to Goldstein et al. (2012) where estimates of matching probabilities are very highly accurate and where it is possible to obtain highly accurate estimates of relationships for (x, y) pairs on **A × B** (particularly from previous matching situations),

Hof and Zwinderman (2012, 2015) provide EM-based methods for adjusting a statistical analysis for matching error. Like the other methods, their procedure successively and iteratively makes adjustment for linkage error and errors in the statistical models. They provide a considerable number of details related to the likelihoods and how the various terms are approximated. For individuals familiar with EM-based methods, their procedure appears fairly straightforward to apply.

The natural way of analyzing discrete data is with loglinear models on the pairs of records in **A** × **B**. Performing analyses for discrete data ais more difficult than for situations with certain regression models where the form of the model gives us considerable simplifying information. Chipperfield et al (2011) gives certain insights that are not available in many of the other papers. The Chipperfield et al. methods are closely related to Winkler (2002) which contains a full likelihood development. For consistency with Chipperfield et al. (2011) the notation is slightly changed. Rather than break out **A** and **B** as $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_m)$, it is possible to merely enumerate A with x in **A** and **B** with y in **B**. All observed pairs have probabilities p_{xy} where p_{xy} represents all pairs of records in **A × B** with x in **A** and y in **B**. If the truth is known, all the p_{xy} would be known. It is possible to estimate the p_{xy} in a semi-supervised fashion as with the likelihood equation given in (6). Chipperfield et al. (2011) take a sample of pairs s_c for which they can determine p_{xy} exactly (no estimation error) for all pairs x and y associated with s_c. In some situations, the methods work very well. In other situations, a very large truth sample may be needed; else there would not be sufficient information for the estimation.

The methods for adjusting regression analyses for linkage error with continuous regression situations have been more successful than the methods for adjusting statistical analyses of discrete data because of various inherent simplifications due to the form of regression models of continuous data.

6. Concluding Remarks

The paper provides background and an overview of methods of edit/imputation and record linkage that have been developed and used in the National Statistical Institutes. The generalized systems apply the theory of Fellegi and Holt (1976) and Fellegi and Sunter (1969), respectively. While improving quality, the generalized systems also yield significant cost- and time-savings. The methods of adjusting statistical analyses for linkage error are a large and difficult area of current research. A crucial aspect of developing generalized software and investigating new methods is having highly skilled teams that can do new theory when needed. The teams also need to develop a large group of new computational algorithms that are sufficiently powerful and fast for production systems.

7. Issues and some related questions

a. If an NSI does not have the skill and/or resources to develop their own systems, then how does the NSI create suitable test decks to evaluate the quality of record linkage and edit/imputation software developed by vendors or in shareware?

b. File A contains name and address of persons. File B contains name and date-of-birth for the same

population of persons. Is it possible to do the matching? What would be an upper bound on the false match rate?

c. Two files A and B represent the same population. Files A and B both contain the quasi-identifiers first name, last name, and date-of-birth. If file A has missing first name for twenty-five percent of the records and file B has missing first name for twenty-five percent of the records, what is the maximum matching rate that can be achieved?

d. Two files A and B are matched where the main output has all pairs along with matching weight and quasi-identifiers in a form that is reasonably easy to use. There are resources available to review a certain number of pairs to determine whether they are a true or false match? How does one efficiently draw a sample that can be used in determining the false match rate? Is it possible to determine the false nonmatch rate? Why?

References

Andridge, R. A. and Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review*, 78 (1), 40-64.

Barcaroli, G., and Venturi, M. (1997), "DAISY (Design, Analysis and Imputation System): Structure, Methodology, and First Applications," in (J. Kovar and L. Granquist, eds.) *Statistical Data Editing, Volume II*, U.N. Economic Commission for Europe, 40-51.

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.

Chambers, R. (2009), Regression Analysis of Probability-Linked Data, Statisphere, Volume 4, http://www.statisphere,govt.nz/official-statistics-research/series/vol-4.htm.

Chipperfield, J. O., Bishop, G. R., and Campbell, P. (2011), Maximum Likelihood estimation for contingency tables and logistic regression with incorrectly linked data, *Survey Methodology*, 37 (1), 13-24.

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.

Gill, L. (2001), *Methods of Automatic Record Matching and Linking and their use in National Statistics*, National Statistics Methodological Series, No. 25, http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_25_v2.pdf.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009), Multilevel models with multivariate mixed response types, *Statistical Modeling*, 9 (3), 173-197.

Goldstein, H., Harron, K., Wade, A. (2012), The analysis of record-linked data using multiple imputation with data prior values, *Statistics in Medicine*, DOI: 10.1002/sim.5508.

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013), "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs," *J. Amer. Stat. Assn*., 108 (501), 34-47.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques*, New York, N. Y.: Springer.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2010), "Record Linkage," in (D. W. Scott, Y. Said, and E. Wegman, eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*, New York, N. Y.: Wiley, 2 (5), September/October, 535-543 .

Hof, M. H. P., and Zwinderman, (2012), Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables, *Statistics in Medicine*, 31, 4231-4232.

Hof, M. H. P., and Zwinderman, (2015), A mixture model for the analysis of data derived from record linkage, *Statistics in Medicine*, 34, 74-92.

Hogan, H. H. and Wolter, K. (1984), “Research Plan on Adjustment”, https://www.census.gov/srd/papers/pdf/rr84-12.pdf.

Karapiperis, D., and Verykios, V. (2014) Load-balancing the distance calculations in record linkage, ACM SIGKDD Explorations, 17 (1), 1-7.

Kawai, H., Garcia-Molina, H., Benjelloun, O., Menestrina, D., Whang, E., and Gong, H. (2006), “P-Swoosh: Parallel Algorithm for Generic Entity Resolution,” Stanford University CS technical report.

Kim, G. and Chambers, R. (2012a), Regression Analysis under Incomplete Linkage, *Computational Statistics and Data Analysis*, 56, 2756-2770.

Kim, G. and Chambers, R. (2012b), Regression Analysis under Probabilistic Multi-linkage, *Statistica Neerlandica*, 66 (1), 64-79.

Kolb, L. and Rahm, E. (2013), “Parallel Entity Resolution with Dedoop,” *Datenbank-Spektrum*, 13 (1), 23-32, http://dbs.uni-leipzig.de/file/parallel_er_with_dedoop.pdf.

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 http://www.census.gov/srd/papers/pdf/rr2000-04.pdf).

Larsen, K. (2005), Generalized Naïve Bayes Classifiers, SIGKDD Explorations, 7 (1), June 2005, 76-81, doi>10.1145/1089815.1089826 .

Larsen, M. D., and Rubin, D. B. (2001), Alterative Automated Record Linkage Using Mixture Models, *Journal of the American Statistical Association*, 79, 32-41.

Lahiri, P. A., and Larsen, M. D. (2005) Regression Analysis with Linked Data, *Journal of the American Statistical Association*, 100, 222-230.

Liseo, B. and Tancredi, A. (2011), Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets, *Survey Methodology*, 27 (3), 491-505.

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

Neter, J., Maynes, E. S., and Ramanathan, R. (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 60, 1005-1027.

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.

Newcombe, H. B., and Smith, M. E. (1975), “Methods for Computer Linkage of Hospital Admission- Separation Records into Cumulative Health Histories,” *Methods of Information in Medicine*, 14 (3), 118-125.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T., (2000), “Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, 39, 103-134.

Rao, J. N. K. (1997), “Developments in Sample Survey Theory: An Appraisal,” *The Canadian Journal of Statistics, La Revue Canadienne de Statistique*, 25 (1), 1-21.

Reiter, J. P. (2018), Assessing Uncertainty when Using Linked Administrative Lists, Chapter 10 in (A. Y. Chun and M. L. Larsen, eds. *Administrative Records for Survey Methodology*), John Wiley & Sons, New York, New York.

Scheuren, F.,and Winkler, W. E. (1993), Regression analysis of data files that are computer matched, *Survey Methodology*, 19, 39-58, also at http://www.fcsm.gov/working-papers/scheuren_part1.pdf .

Scheuren, F.,and Winkler, W. E. (1997), Regression analysis of data files that are computer matched, II, *Survey Methodology*, 23, 157-165, also at http://www.fcsm.gov/working-papers/scheuren_part2.pdf.

Tancredi, A., and Liseo, B. (2011), A Hierarchical Bayesian Approach to Matching and Size Population Problems, *Ann. Appl. Stat.*, 5 (2B), 1553-1585.

Tancredi, A., and Liseo, B. (2015), Regression Analysis with Linked Data, Problems and Possible Solutions, *Statistica*, https://rivista-statistica.unibo.it/article/view/5821/0.

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, also at http://www.census.gov/srd/papers/pdf/rr2000-05.pdf .

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), "Documentation of record-linkage software," unpublished report, Washington DC: Statistical Research Division, U.S. Bureau of the Census.
Winkler, W. E. (1990b), "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 (available at *www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf*).
Winkler, W. E. (1990c), "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, also http://www.census.gov/srd/papers/pdf/rr93-12.pdf .
Winkler, W. E. (1995a), "Matching and Record Linkage," in B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, Colledge, M. A., and P. S. Kott (eds.) *Business Survey Methods*, New York: J. Wiley, 355-384 (as of January 2015, also available at http://fcsm.sites.usa.gov/files/2014/04/RLT97.pdf on pages 359-403).
Winkler, W. E. (1995b), "SPEER Economic Editing Software," Unpublished technical report.
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 http://www.census.gov/srd/papers/pdf/rr9801.pdf).
Winkler, W. E. (2000), "Machine Learning, Information Retrieval, and Record Linkage," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 20-29. (http://www.amstat.org/sections/SRMS/Proceedings/papers/2000_003.pdf, also available at http://www.niss.org/sites/default/files/winkler.pdf).
Winkler, W. E. (2002), Record Linkage and Bayesian Networks, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM (also 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 http://www.census.gov/srd/papers/pdf/rrs2003-07.pdf .
Winkler, W. E. (2004), 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. (2006a), Automatic Estimation of Record Linkage False Match Rates, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM, also at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf .
Winkler, W. E. (2006b), "Statistical Matching Software for Discrete Data," computer software and documentation.
Winkler, W. E. (2006c), "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. (2008), General Methods and Algorithms for Imputing Discrete Data under a Variety of Constraints, http://www.census.gov/srd/papers/pdf/rrs2008-08.pdf .
Winkler, W. E. (2009), Using General Edit/Imputation and Record Linkage Methods and Tools to Enhance Methods for Evaluating and Minimizing Uncertainty in Statistical Matching, unpublished technical report.
Winkler, W. E. (2010), General Discrete-data Modeling Methods for Creating Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties, http://www.census.gov/srd/papers/pdf/rrs2010-02.pdf .
Winkler, W. E. (2013a). "Record Linkage," in *Encyclopedia of Environmetrics*. J. Wiley.
Winkler, W. E. (2013b, Methods for adjusting statistical analyses for record linkage error, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.
Winkler, W.E. (2013c), Cleanup and Analysis of Sets of National Files, Federal Committee on Statistical Methodology, Proceedings of the Bi-Annual Research Conference,

http://www.copafs.org/UserFiles/file/fcsm/J1_Winkler_2013FCSM.pdf., https://fcsm.sites.usa.gov/**files**/2014/05/J1_**Winkler_2013**FCSM.pdf

Winkler, W. E. (2014), Matching and Record Linkage, *Wiley Interdisciplinary Reviews: Computational Statistics*, https://wires.wiley.com/WileyCDA/WiresArticle/wisId-WICS1317.html, DOI: 10.1002/wics.1317.

Winkler, W. E. and Chen, B.-C. (2002), "Extending the Fellegi-Holt Model of Statistical Data Editing," (available at http://www.census.gov/srd/papers/pdf/rrs2002-02.pdf).

Winkler, W. E. and Hidiroglou, M. (1998), "Developing Analytic Programming Ability to Empower the Survey Organization," http://www.census.gov/srd/papers/pdf/rr9804.pdf .

Winkler, W. E., and Petkunas, T. (1997), "The DISCRETE Edit System," in (J. Kovar and L. Granquist, eds.) *Statistical Data Editing, Volume II*, U.N. Economic Commission for Europe, 56-62, also http://www.census.gov/srd/papers/pdf/rr96-3.pdf .

Winkler, W. E. and Scheuren, F. (1991), "How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis," U.S. Bureau of the Census, Statistical Research Division Technical Report.

Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical Report 91-9, http://www.census.gov/srd/papers/pdf/rr91-9.pdf.

Winkler, W. E., Yancey, W. E., and Porter, E. H. (2010), "Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, CD-ROM.

Wright, J. (2010), "Linking Census Records to Death Registrations," Australia Bureau of Statistics Report 131.0.55.030.

Yan, W., Xue, Y, and Malin, B. (2013) Scalable Load Balancing for Map-Reduced Based Record Linkage, Performance Computing and Communications Conference (IPCCC), 2013 IEEE 32nd International, San Diego, CA, DOI:10.1109/PCCC.2013.6742785 .

Yancey, W. E. and Winkler, W. E. (2004), "BigMatch Record Linkage Software," Statistical Research Division Research Report.

Yancey, W. E. and Winkler, W. E. (2009), "BigMatch Record Linkage Software," Statistical Research Division Research Report.