Cleaning and using administrative lists: Methods and fast computational algorithms for record linkage and modeling/editing/imputation

William E. Winkler
U.S Bureau of the Census, william.e.winkler@census.gov 1/

Abstract - Administrative lists offer great opportunity for analyses that provide quantities for policy decisions. This is particularly true when groups of administrative lists are combined with survey and other data. To produce accurate analyses, data need to be cleaned and corrected according to valid subject matter rules. This paper describes methods and associated computational algorithms that, while often being easier to apply, are sometimes 40-100 times as fast as classical methods. This means that moderate-size administrative files can be cleaned (via modeling/edit/imputation) to eliminate contradictory or missing quantitative data to yield valid joint distributions, unduplicated within files, and matched and merged across files in a matter of weeks or months.

Keywords: quality, merging, computational algorithms

1. Introduction

Well collected and processed administrative data can be of great use for providing enhanced aggregates and microdata for analytic purposes. In this paper, we assume that data are collected and processed in a manner that minimizes error. We describe three methods for processing data. The first are modeling/edit/imputation methods for filling in missing data and 'correcting' erroneous or contradictory data. The second are record linkage (entity resolution) methods for matching files using common quasi-identifiers such as name, address, date-of-birth, and other characteristics. The third are methods for adjusting analyses of merged files for linkage error.

The modeling/edit/imputation methods are based on the theoretical model and suggested algorithms of Fellegi and Holt (1976, hereafter FH). Versions of generalized software for editing and certain types of imputation have been in use in a few statistical agencies for more than ten years. What is new is a rigorous method of theoretically connecting editing with modern imputation such as given in Little and Rubin (2002). Winkler (2003) introduced the theory for discrete data and provided extremely fast computation algorithms (2008, 2010b) in highly automated, parameter-driven software. The set-covering algorithms (Winkler 1997) for enumerating all implicit edits are 100 times as fast as those of IBM based on the ideas of Garfinkel, Kunnathur, and Liepins(1986). The modeling, imputation, and imputation-variance are on the order of 100 times as fast as those in commercial or experimental university software.

The record linkage algorithms (Yancey and Winkler 2005-2009, Winkler, Yancey, and Porter 2010) are 40+ times as fast as recent parallel software from Stanford and Penn State (Kawai et al. 2006, Kim and Lee 2007) and 500+ times as software used in some government agencies (e.g., Wright 2010).

The analysis-adjustment methods for merged files are still quite preliminary (Scheuren and Winkler 1993, 1997; Lahiri and Larsen 2005, Chambers 2009) with the main difficulties being seen as properly creating an overall model of the record linkage process and having suitable generalized methods for adjusting analyses for error. The methods of Chambers (2009) appear to show great promise in drastically simplified record linkage situations and simple simulations but may not extend to the more general and far more realistic situations of Lahiri and Larsen (2005). At issue in all of the work are methods for estimating suitable probabilities of matching for all pairs (typically without training data). Lahiri and Larsen (2005) and Chambers (2009) assume that extremely large resources and time may be available for follow-up on an exceptionally large number of pairs to determine matching probabilities. Scheuren and Winkler (1993) made simplifications in the adjustment procedures because they were able to make use of methods due to Belin and Rubin (1995) for estimating match probabilities. A more general method for estimating match probabilities (Winkler 2006) mimics ideas from semi-supervised learning (see e.g. Larsen and Rubin 2001, Winkler 2002, Nigam et al. 2000) but also does not use training data.

A conceptual picture would link records in file

$$\mathbf{A} = (a_i, \ldots, a_n, x_1, \ldots, x_k)$$

with records in file

$$\mathbf{B} = (b_1, \ldots, b_m, x_1, \ldots, x_k)$$

using common identifying information $(x_1, \ldots, x_k)$ 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. Survey files routinely require cleanup via edit/imputation and administrative files may also require similar cleanup. 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., effective edit/imputation) of $\mathbf{A} = (a_i, \ldots, a_n, x_1, \ldots, x_k)$ and $\mathbf{B} = (b_1, \ldots, b_m, x_1, \ldots, x_k)$, then analyses may have other serious errors in addition to the errors due to the linkage errors.

The purpose of the paper is to describe the available newer theoretical ideas and new computational algorithms. If we have several administrative lists each with 100 million to one billion records, then the clean-up, merging, and analyses might be performed in 3-4 months with this software that is 40-100 times as fast. Without the faster software, the problem of extensive cleanup, merging, and analysis of sets of large administrative lists is computationally intractable.

In the next three sections, we provide background and insight into modeling/edit/imputation, record linkage, and adjustment of analyses for linkage error.

2. Modeling/edit/imputation

In this section we provide background on classical edit/imputation that uses hot-deck and provide a description of how hot-deck was assumed to work by practitioners. As far as we know, there has never been a rigorous development that may justify some of the assumed properties of hot-deck. We also provide background methods of creating loglinear models **Y** (Bishop, Fienberg and Holland 1975) that are straightforward to apply to general discrete data, background on general methods of imputation and editing for missing data under linear constraints that extend the basic methods and can also be straightforward to apply, and an elementary review of the EM algorithm. The application of the general methods and software is straightforward. The application can be done without any modifications that are specific to a particular data file or analytic use.

2.1 Classical data collection, edit rules, and hot-deck imputation

The intent of classical data collection and clean-up was to provide a data file that was free of logical errors and missing data. For a statistical agency, a survey form might be filled out by an interviewer 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 being placed in computer files. The purpose was to produce a 'complete' (i.e., no missing values) data file that had 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).

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 arose that, as a (computerized) record r_0 was changed to a different record r_1 by changing values in fields in which edits failed, then the new record r_1 would fail other edits that the original record r_0 had not failed.

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 required the computation of all implicit edits that could be logically derived from an originally specified set of 'explicit' edits. If the implicit edits were available, then it was 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). Further, Fellegi and Holt indicated how hot-deck might be used to provide the values for filling in missing values or replacing contradictory values. As shown in Winkler (2008b), hot-deck is not generally suitable for filling in missing values in a manner that yields records that satisfy edits and preserve joint distributions. Indeed, the imputation methods in use at a variety of statistical agencies and those that are also being investigated do not assure that aggregates of records satisfy joint distributions and that individual records satisfy edits.

The early set-covering algorithms necessary for the computation of 'implicit' edits required extremely large amounts of computer time (Garfinkel, Kunnathur, and Liepins 1986). A later algorithm (Winkler 1997), while as much as 100 times as fast, is not completely theoretically valid but works in most situations where skip patterns are not present in the survey form (see also Winkler and Chen 2002). Due to hardware-speed increases, the latter algorithm should work well in most day-to-day survey situations. Both Winkler (1997) and Boskovitz (2008) provided counterexamples to Theorem 1 in Garfinkel et. al (1986) which gave a method for greatly simplifying the set covering algorithms for implicit-edit generation. Boskovitz (2008) provided a complete theoretical development (including data with skip patterns), however software based on her algorithms has not yet been written and will likely be 10 times as slow due to the significantly greater amount of information that must be accounted for at different levels of the computational algorithms.

The intent of filling-in missing or contradictory values in edit-failing records r_0 is to obtain a records r_1 that can be used in computing the joint probabilities in a principled manner. The difficulty that had been 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.

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 (2008b), 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 that 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 2008b, also 2010d). Edit rules are in easily modified tables, and the logical consistency of the entire system is tested automatically according 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). Most modeling is very straightforward. It is based on variants of loglinear modeling and extensions of missing data methods that is contained in easily applied, extremely fast computational algorithms (Winkler 2006, 2008b; also 2010a). The methods create records that *always* satisfy edits and preserve joint inclusion probabilities.

2.2 How classical hot-deck is assumed to work

In this subsection we provide an explanation of some of the (possibly) subtle issues that significantly degrade the overall analytic characteristics of realistic data files (8 or more variables) that are subjected to *well-implemented* hot-deck. The reason that the issues may be subtle is that in many situations with hot-deck, the probabilistic model is not written down and the effects of the statistical evaluations (say logistic or ordinary regression) on hot-deck collapsing rules for matching are not evaluated. We will describe why it is effectively impossible in many practical survey situations to do the empirical testing and develop program logic necessary for a well-implemented hot-deck. Prior to this we provide some notation and background that will allow us to describe why hot-deck breaks down in terms of the basic modeling frameworks of Little and Rubin (2002) and Winkler (2003).

We assume $\mathbf{X} = X_i = (x_{ij})$, $1 \le i \le N$, $1 \le j \le M$ is a representation of the survey data with N rows (records) and M columns (variables). Record x_i has values x_{ij}, $1 \le j \le M$. The j^{th} variables X_j takes values x_{jk}, $1 \le k \le n_j$. The total number of patterns is $npat = n_1 \times \ldots \times n_M$. In most realistic survey situations (8 or more variables), the number of possible patterns *npat* is far greater than N (i.e., N << *npat*). Under classical hot-deck assumptions (that are essentially universally used in statistical agencies), the typical assumption is that we will be able to match a record $r_0 = (x_{01}, x_{02}, \ldots, x_{0M})$ having missing values of certain variables against a large number of donor records that have no missing variables and that agree with record r_0 on the non-missing values. If record r_0 has eight variables with the last three variables having missing values, then the intent of hot-deck (after it is implemented over an entire file) is to create a set of records that preserve the original probability structure of a hypothetical file **X** having no missing values.

We start with record $r_0 = (x_{01}, x_{02}, \ldots, x_{05}, b, b, b)$ where b represents a missing value for x_{06}, x_{07}, and x_{08}. Under the hot-deck assumptions, our matching would allow use to effectively draw from the distribution of $P(X_6, X_7, X_8 \mid X_1{=}x_{01}, \ldots, X_5{=}x_{05})$. In practice with real-world data, we typically have zero donors (rather than an exceptionally large number that would be needed to preserve joint distributions). Statistical agencies typically use ad hoc collapsing in which they attempt to match on a subset of the values x_{01}, x_{02}, …, x_{05}. For instance, there may be a matching hierarchy in which the first match attempt is on x_{01}, x_{02}, x_{03}. If a donor record is not found matching may be done on x_{01} and x_{02}. If no donor is found, then matching might be done on only x_{01} where it might be possible to always find a donor.

If we are able to match on x_{01}, x_{02} and x_{03}, we obtain a record $r_d = (x_{d1}, \ldots.\ x_{d8})$ that yields a hot-deck completed record $r_{0c} = (x_{01}, \ldots, x_{05}, x_{d6}, x_{d7}, x_{d8})$. There is no assurance that the substituted values will preserve joint distributions or create a record that satisfies edits. Indeed, elementary empirical work with exceptionally simple simulated data (that should preserve joint distributions under the hot-deck assumption) also demonstrate that joint distributions are not preserved. Although the elementary work uses data situations that are much nicer than many real-world situations, it still fails to yield hot-deck imputations that preserve joint distributions. To preserve joint distributions, it might be necessary to create some type of basic model for collapsing. A simplistic approach might be to use logistic regression to find what subsets of x_{01}, …, x_{05} are the best predictors of the remaining variables and choose the collapsing hierarchy based on a very large set of logistic regressions.

Even after such work (that is very specific to an individual data set), it is not clear why the joint distributions would be preserved. It would be much better to have a general modeling framework (possibly an extension of Little and Rubin (2002), chapter 13) and software that would work for arbitrary discrete data under mild assumptions. One mild assumption is the *missing-at-random* assumption (Little and Rubin 2002) that is effectively the hot-deck assumption in a framework in which it is possible to preserve joint inclusion probabilities. An effective model might be multinomial (or multinomial with weak Dirichlet prior) that all non-structural-zero cells are given a non-zero (but possibly very close to zero) values. In this situation (p_i), $1 \le i \le npat$ are the probabilities of the multinomial with the individual cells, and we have a suitable probability structure. With this extended hot-deck (effectively Little-Rubin ideas), we match against cells that agree with the non-missing part of a record r_0 and choose one cell (donor pattern or record) with probability proportional to size of the cell probability.

2.3. New Computational Algorithms for Modeling and Imputation

The generalized software (Winkler 2010b) incorporates ideas from statistical matching software (Winkler 2006) that can be compared to ideas and results of D'Orazio et al. (2006) and earlier discrete-data editing software (Winkler 2008b) that could be used for synthetic-data generation (Winkler 2010a). The basic methods are closely related to ideas suggested in Little and Rubin (2002, Chapter 13) in that they assume a missing-at-random assumption that can be slightly weakened in some situations (Winkler 2008b, 2010a). The original theory for the computational algorithms (Winkler 1993) uses convex constraints (Winkler 1990) to produce an EMH algorithm that generalizes the MCECM algorithm of Meng and Rubin (1993). The EMH algorithm was first applied to record linkage (Winkler 1993) and used by D'Orazio, Di Zio, and Scanu (2006) in statistical matching.

The current algorithms 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 computational speed for a contingency table of size 600,000 is 50 seconds and for a table of size 0.5 billion cells in approximately 1000 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.

Certain convex constraints can be incorporated in addition to the standard linear constraints of classic loglinear EM fitting. In statistical matching (Winkler 2006c) was able to incorporate closed form constraints P(Variable X1 = x11 > Variable X1= x12) with the same data as D'Orazio et al. (2006) that needed a much slower iterative fitting algorithm for the same data and constraints. The variable X1 took four values and the restraint is that one margin of X1 for one value is restricted to be greater than one margin of another value. For general edit/imputation, Winkler (2008b) was able to put marginal constraints on one variable to assure that the resultant micordata files and associated margins corresponded much more closely to observed margins from an auxiliary data source. For, instance one variable could a an income range and the produced microdata did not produce population proportions that corresponded closely to published IRS data until after appropriate convex constraints were additionally applied. Winkler (2010a) used convex constraints to place upper and lower bounds on cell probabilities to assure that any synthetic data generated from the models would have reduced/eliminated re-identification risk while still preserving the main analytic properties of the original confidential data.

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. An additional algorithm nearly instantaneously constructs an array that allows binary search to locate the cell for the imputation (for the two algorithms: total < 1.0 millisecond cpu time). For instance, if a record has 12 variables and 5 have missing, we might 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.

For imputation-variance estimation, other closely related algorithms allow direct variance estimation from the model. This is in contrast to after-the-fact variance approximations using linearization, jackknife or bootstrap. These latter three methods were developed for after-the-fact variance estimation (typically with possibly poorly implemented hot-deck imputation) that are unable to account effectively for the bias of hot-deck or that lack of model with hot-deck. Most of the methods for the after-the-fact imputation-variance estimation have only been developed for one-variable situations that do not account for the multivariate characteristics of the data and assume that hot-deck matching (when naively applied) is straightforward when most hot-deck matching is never straightforward.

3. Record Linkage

Fellegi and Sunter (1969) provided 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 **A** × **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:

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

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 1995). 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 $T_\lambda \leq R \leq T_\mu$, then designate pair as a possible match
and hold for clerical review. (2)

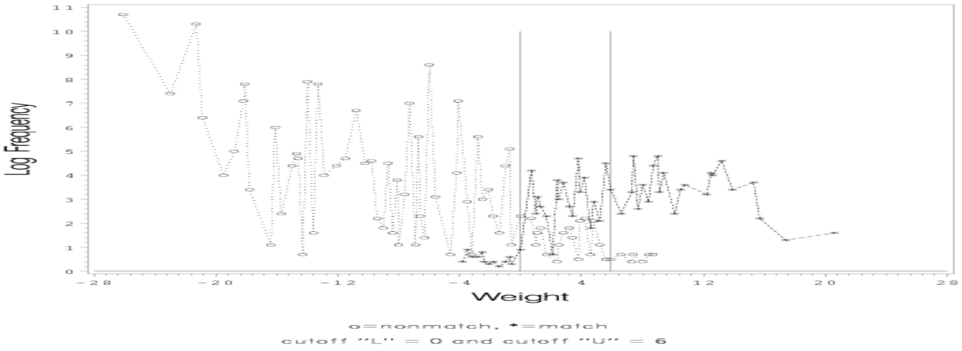
If $R < T_\lambda$, then designate pair as a nonmatch.

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 $T_\lambda \leq R \leq T_\mu$ is referred to as the *no-decision region* or *clerical review* region. In some situations, resources are available to review pairs clerically.

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 **A** × **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 and T_μ, 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 was 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).

Figure 1. Log Frequency vs Weight
Matches and Nonmatches Combined



In many situations with administrative lists, we need to process an enormous number of pairs. For instance, in the Decennial Census, we process 10^17 (300 million x 300 million). The way that we reduce computation is with blocking. Blocking consists of only considering pairs that agree on characteristics such as a Census block code plus first character of the surname. If we using multiple blocking passes, then we may additionally may consider pairs that only agree on telephone number, street address, or the first few characters of first name plus first few characters of surname. In traditional record linkage, two files are sorted according to a blocking criteria, matched, processed and then (possibly) successive residual files are processed according to subsequent blocking criteria. With a large billion-record file, each sort could require 12+ hours.

BigMatch technology (see e.g. Yancey 2007; Winkler, Yancey and Porter 2010) solves this issue by embedding the smaller file in memory, creating indices for each blocking criteria (in memory) and running through the larger file once. As each record from the larger file in read in, it is processed against each of the blocking criteria and separate scores associated with each pair along with other information are output. BigMatch is 50 times as fast as recent parallel software from Stanford (Kawai et al. 2006) and 40 times as fast as parallel software from Penn State (Kim and Lee 2007). In production matching during the 2010 Decennial Census, BigMatch did detailed computation on 10^12 pairs among 10^17 pairs in 30 hours using 40 cpus of an SGI Linux machine. In equivalently large situations with slower software, a project might require 80 machines and a whole crew of programmers to split up files and slowly put together all the matches coherently in 20 weeks. There would be substantial opportunity for error as the programmers broke up files into much smaller subsets, moved subsets to different machines, and then attempted to move (possibly hundreds) of outputs back to other machines.

4. Analysis Adjustment in Merged Files having Linkage Error

In this section, we describe research into methods for adjusting statistical analyses for linkage error. Unlike the much more mature methods in the previous two sections, there are substantial research problems. Scheuren and Winkler (1993) extended methods of Neter, Maynes, and

Ramanathgan (1965) to more realistic record linkage situations in the simple analyses of a regression of the form y = β x where y is taken from one file A and x is taken from another file B. Because the notation of Lahiri and Larsen (2005) is more useful in describing extensions and limitations, we use their notation.

Consider the regression model $\mathbf{y} = (y_1, \ldots, y_n)'$:

$$y_i = \mathbf{x_i'}\, \beta + \varepsilon_i,\ i = 1, \ldots, n \qquad (3)$$

where $x_i = (x_{i1}, \ldots, x_{ip})$ is a column vector of p known covariates $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2$, $\mathrm{covariance}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, $i, j = 1, \ldots, n$. Scheuren and Winkler (1993) considered the following model for $\mathbf{z} = (z_1, \ldots, z_n)$ given y:

$$z_i = \begin{cases} y_i \text{ with probability } q_{ii} \\ y_j \text{ with probability } q_{ij} \text{ for } i \neq j,\ i, j = 1, \ldots, n \end{cases} \qquad (4)$$

where $\sum_{j=1}^{n} q_{ij} = 1$ for $i = 1, \ldots, n$. Define $\mathbf{q}_i = (q_{i1}, \ldots, q_{in})'$, $j = 1, \ldots, n$, and $\mathbf{Q} = (\mathbf{q}_i, \ldots, \mathbf{q}_n)'$. The naïve least squares estimator of $\boldsymbol{\beta}$, which ignores mismatch errors, is given by

$$\boldsymbol{\beta_N}^{\wedge} = \mathbf{(X'\, X)^{-1}\, X'\, z},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is an $n \times p$ matrix.

Under the model described by (3) and (4)

$$E(z_i) = \mathbf{w}'\beta$$

where $\mathbf{w}_i = \mathbf{q}_i'\, X = \sum_{j=1}^{n} q_{ij}\, \mathbf{x}_j'$, $i = 1, \ldots, n$, is a p x 1 column matrix. The bias of the naïve estimator $\boldsymbol{\beta_N}^{\wedge}$ is given by

$$\mathrm{bias}(\boldsymbol{\beta_N}^{\wedge}) = E(\boldsymbol{\beta_N}^{\wedge} - \beta) = [(\mathbf{X'X})^{-1}\, \mathbf{X'W} - I]\, \beta = [(\mathbf{X'X})^{-1}\, \mathbf{X'Q\, X} - I]\, \boldsymbol{\beta}. \qquad (5)$$

If an estimator of $\mathbf{B}$ is available where $\mathbf{B} = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)\, y_i + \sum_{j\neq 1} q_{ij}\, y_j$. The Scheuren-Winkler estimator is given by

$$\boldsymbol{\beta_{SW}}^{\wedge} = \boldsymbol{\beta_N}^{\wedge} - \mathbf{X'X)}^{-1}\, \mathbf{X'\, B^{\wedge}} \qquad (6)$$

If q_{ij1} and q_{ij2} denote the first and second highest elements of the vector $\mathbf{q}_i$ and z_{j1} and z_{j2} denote the elements of the vector $\mathbf{z}$, then a truncated estimator of B is given by

$$\mathbf{B}_i'^{TR} = (q_{ij1} - 1)\, z_{j1} + q_{ij2}\, z_{j2}\ . \qquad (7)$$

Scheuren and Winkler (1993) used estimates of q_{ij1} and q_{ij2} based on software/methods from Belin and Rubin (1995). Lahiri and Larsen improve the estimator (7) (sometimes significantly) by using the unbiased estimator

$$\beta_U{}^{\wedge} = \mathbf{(W' W)^{-1} W' z}. \quad (8)$$

The issues are whether it is possible to obtain reasonable estimates of $\mathbf{q}_i$ or whether the crude approximation given by (7) is suitable in a number of situations.
Under a significantly simplified record linkage model where each q_{ij} for $i \neq j$, 1, …, n, Chambers (2009) provides an estimator approximately of the following form

$$\beta_U{}^{\wedge} = \mathbf{(W' Cov_z{}^{-1} W)^{-1} W' Cov_z{}^{-1} z} \quad (9)$$

that has lower bias than the estimator of Lahiri and Larsen. The matrix $\mathbf{Cov_z}$ is the variance-covariance matrix associated with $\mathbf{z}$. The estimator in (9) is the best linear unbiased estimator using standard methods that improve over the unbiased estimator (8). Chambers further provides an iterative method for obtaining an empirical BLUE using the observed data.
The issue with the Chambers' estimator is whether the drastically simplified record linkage model is a suitable approximation of the realistic model used by Lahiri and Larsen. The issue with both the models of Chambers (2009) and Lahiri and Larsen (2005) is that they need both a method of estimating q_{ij} for all i, j with all pairs of records and a method of designating which of the q_{ij} is associated with the true match. Scheuren and Winkler (1993) provided a much more ad hoc adjustment with the somewhat crude estimates of the q_{ij} obtained from the model of Belin and Rubin (1995). Lahiri and Larsen demonstrated that the Scheuren-Winkler procedure was inferior for adjustment purposes when the true q_{ij} were known. Winkler and Scheuren (1991), however, were able to determine that their adjustment worked well in a very large number of empirical scenarios (several hundred). Further, Winkler (2006) provided a 'generalization' of the Belin-Rubin estimation procedure that provides somewhat more accurate estimates of the q_{ij} and holds in a moderately larger number of situations.

5. Concluding Remarks

This paper describes methods of modeling/edit/imputation and record linkage that are reasonably mature methods in terms of improving the quality of administrative and that have been greatly enhanced by breakthroughs in computational speed. Newer methods for adjusting statistical analyses for linkage error (Lahiri and Larsen, 2005; Chambers 2009) are very much in their preliminary stages and need substantial additional research. A very new method due to Tancredi and Liseo (2011) shows great potential both theoretically and methodologically but must be extended to more practical computational situations.

1/ This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed on (statistical, methodological, technical, or operational) issues are those of the author(s) and not necessarily those of the U.S. Census Bureau.

References

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.

Boskovitz, A. (2008), “Data Editing and Logic: The covering set methods from the perspective of logic,” CS Ph.D. dissertation, Australia National University, http://thesis.anu.edu.au/public/adt-ANU20080314.163155/index.html .

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

D’Orazio, M., Di Zio, M., and Scanu, M. (2006), “Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints,” *Journal of Official Statistics*, 22 (1), 137-157.

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.

Garfinkel, R. S., Kunnathur, A. S., and Liepins, G. E., (1986), "Optimal Imputation of Erroneous Data: Categorical Data, General Edits," *Operations Research*, 34, 744-751.

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 .

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, H.-S., and Lee, D. (2007), “Parallel Linkage,” CIKM ’07.

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

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

Meng, X.-L., and Rubin, D. B. (1993), “Maximum Likelihood via the ECM Algorithm: A General Framework,” *Biometrika*, 80, 267-78.

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.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000), “Text Classification from Labeled and Unlabeled 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.

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, 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.

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

Winkler, W. E. (1991), "Error Model for Computer Linked Files," Pr*oceedings of the Section on Survey Research Methods*, *American Statistical Association*, 472-477.

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. (1995), "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 (also available at http://www.fcsm.gov/working-papers/wwinkler.pdf).

Winkler, W.E. (1997a), "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. (1999), "Issues with Linking Files and Performing Analyses on the Merged Files," *American Statistical Association, Proceedings of the Sections on Government Statistics and Social Statistics*, 262-265.

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 at http://www.census.gov/srd/papers/pdf/rr9801.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), "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. (2006b), "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. (2006c), "Statistical Matching Software for Discrete Data," computer software and documentation.

Winkler, W. E. (2008a), "Data Quality in Data Warehouses," in (J. Wang, ed.) *Encyclopedia of Data Warehousing and Data Mining (2nd Edition)*.

Winkler, W. E. (2008b), "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. (2010a), "General Discrete-data Modeling Methods for Creating Synthetic Data with Reduce Re-identification Risk that Preserve Analytic Properties," http://www.census.gov/srd/papers/pdf/rrs2010-02.pdf .

Winkler, W. E. (2010b), "Generalized Modeling/Edit/Imputation Software for Discrete Data," computer software and documentation.

Winkler, W. E. (2010c), "Record Linkage," Course notes from short course at the Institute of Education at the University of London in September 2010.

Winkler, W. E. (2010d), "Cleaning Administrative Data: Improving Quality using Edit and Imputation," Course notes from short course at the Institute of Education at the University of London in September 2010.

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 Scheuren, F. (1991), "How Computer Matching Error Affects Regression Analysis: Exploratory and Confirmatory Report ," unpublished technical report.

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.

Yancey, W.E. (2007), "BigMatch: A Program for Extracting Probable Matches from Large Files," Statistical Division Research Report, http://www.census.gov/srd/papers/pdf/RRC2007-01.pdf .