Cleanup and Statistical Analysis of Sets of National Files

William E. Winkler[1], william.e.winkler@census.gov

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract
This paper provides background on the Fellegi-Sunter model of record linkage for identifying duplicates within and across files, background on the Fellegi-Holt model of statistical data editing for filling in missing data, and an overview of methods of adjusting statistical analyses for record linkage error. It also provides new examples of the severe errors that can occur in statistical analyses when there is record linkage error.

Keywords: record linkage, edit/imputation, measurement error, statistical model

1. Introduction

The national statistical agencies have long been at the forefront of methods of cleaning up data. The earliest methods were manual; later generalized systems based on the record linkage model of Fellegi and Sunter (1969) and the model of statistical data editing of Fellegi and Holt (1976) were developed independently in a number of agencies. More recently, the agencies have been interested in cleaning up national files which has required speed increases of the basic software on the order of 100 or more. With the faster software, a group of skilled individuals can do clean-up of a set of national files and preliminary analyses in 3-6 months; with software that is 100+ times slower (such as most commercial and experimental university software), it is not clear how long the clean-up would take.

In this paper, we do not go into detail about the speed increases of the generalized software that are covered in Winkler, Yancey, and Porter (2010) and Winkler (2008, 2010). We rather describe the methods which are primarily intended to produce quality analyses in sets of national files.

Record linkage (entity resolution) is the method of bringing together records associated with the same entity using quasi-identifiers such as name, address, date-of-birth, etc. While individual quasi-identifiers do not uniquely identify entities, a combination of quasi-identifiers may uniquely identify an entity such as a person. The quality of the information (quasi-identifiers) are crucial to the quality of record linkage (i.e., low false match error rates, possibly low false nonmatch error rates).

Because files (particularly large national files) can yield useful information in statistical analyses, various groups are interested in cleaning-up and merging individual files and (possibly) doing additional clean-up of merged files to correct for linkage error.

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 **B** = (b_1, …, b_m, x_1, …, x_k) , then analyses may have other serious errors in addition to the errors due to the linkage errors.

In this paper, we provide some background on the Fellegi-Sunter model of record linkage, methods of file clean-up and preparation prior to linkage, the Fellegi-Holt model of statistical data editing, current (very primitive) models to estimating record linkage error rates and models for adjusting statistical analyses for linkage error. A new example mimics larger situations that might typically be encountered in practice and illustrates how poorly existing methods perform.

2. Background, Methods, Current Research Problems

In the first subsection, we provide background on the Fellegi-Sunter model of record linkage. In the second, we provide an overview of preprocessing/standardization which can represent a 0.75 proportion of improved matching efficacy. Without the preprocessing/standardization, improving models and parameter estimation to yield reasonable improvements in methods of adjusting statistical analyses for linkage error would be impossible. In the third subsection, we give a brief overview of the Fellegi-Holt model of statistical data editing. In the fourth subsection, we provide a full likelihood development for the models of this paper that also holds (with a substantial change in notation) for the model of Chipperfield et al. (2011). The advantage of the subsection's model is that it is based on theory (Winkler 1993) generalizing the linear constraints of Meng and Rubin (1993) to convex constraints which may improve some of the methods of adjusting statistical analyses for linkage error as they do with general edit/imputation (Winkler 2008, 2010). The fourth subsection describes some of the previous work on adjusting statistical analyses for linkage error. The fifth subsection covers the methods of Chipperfield et al. (2011) and applies them to more realistic situations than have been used by prior authors of methods on statistical adjustment methods for linkage error.

2.1. The Fellegi-Sunter Model of 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 $T_\lambda \le R \le T_\mu$, then designate pair as a possible match
and hold for clerical review. (2)

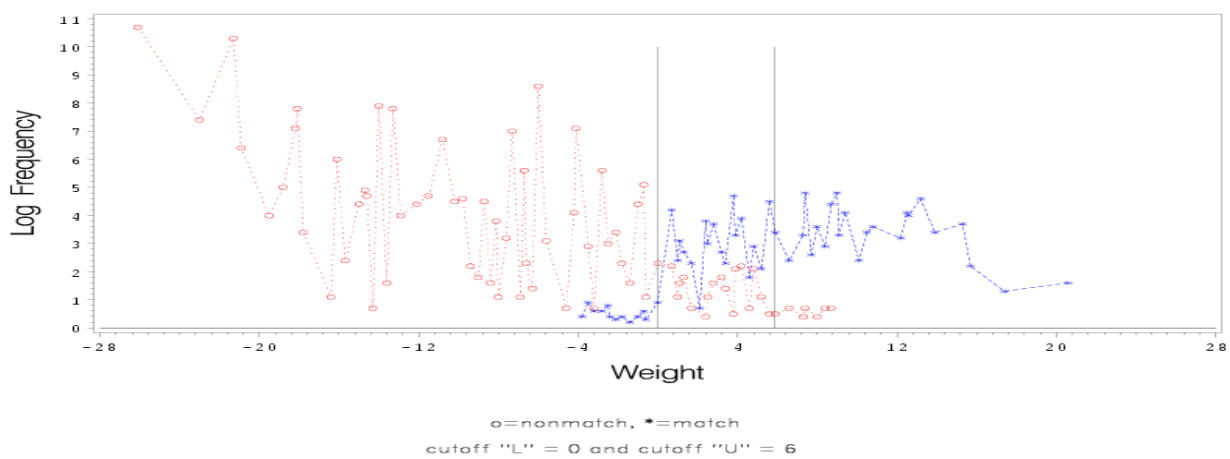
If $R < T_\lambda$, then designate pair as a nonmatch.

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.

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 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



For the 1990 Decennial Census, we had to estimate specific parameters in each of 457 regions. The clerical review region consisted almost entirely of individuals within the same household who were missing both first name and age (the only two fields for distinguishing individuals within a household).

2.2. Data Preparation

The methods for cleaning up data prior to the data being run through the matching routines account for between 50% and 75% of matching efficacy; improved parameter estimation account for the remaining

improvement in matching efficacy. To facilitate the preprocessing, individuals need to assure that names and addresses are broken into corresponding components for comparison, that dates and other fields are in forms that can be compared directly with appropriate algorithms, and that minor typographical error and its effect on the likelihoods is dealt with automatically.

Table 1. Examples of Name Parsing

	Standardized
1.	DR John J Smith MD
2.	Smith DRY FRM
3.	Smith & Son ENTP

	Parsed						
	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith		Son	ENTP	

The name standardization (Table 1) and the address standardization (Table 2) are absolutely crucial to accurate matching. Although there are much more advanced methods of standardization (Agichtein and Gani 2004; Cohen and Sarawagi 2004), we do not need the methods for most types of high quality person lists.

Table 2. Examples of Address Parsing

	Standardized
1.	16 W Main ST APT 16
2.	RR 2 BX 215
3.	Fuller BLDG SUITE 405
4.	14588 HWY 16 W

	Parsed								
	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
1.	W	16	Main			ST		16		
2.				2	215					
3.									405	Fuller
4.		14588	HWY 16				W			

Table 3. Different Date Formats

April 15, 1960
1960Apr15
04/15/1960

Table 4. Examples of Fields having Minor Typographical Error

F1a William
F1b Willam

F2a Roberta
F2b Rburta

F3a Jones
F3b Janes

2.3 Basic Parameter Estimation and Error-rate Estimation

The development in this section is due to that of Winkler (1988) and extended to semi-supervised learning in Winkler (2002). The notation is slightly more general because it deals with the representational framework of record linkage. The underlying computational algorithms are almost identical to those in Chipperfield et al. (2011). 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 in $\mathbf{A} \times \mathbf{B}$. Later, we will assume that some of the C_j will be subsets of M and the remaining C_j are subsets of U. 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, Cj is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or conditional independence (**CI**), we have

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

where the product is over the kth individual field agreement γ_{ik} in pair agreement pattern γ_i. In some situations, we use a Dirichlet prior

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

where the first product is over the classes Cj and the second product is over the fields. We use Du to denote unlabeled pairs and Dl to denote labeled pairs. Given the set D of all labeled and unlabeled pairs, the log likelihood is given by

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

where $0 \le \lambda \le 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs.

In the third terms equation (6), we sum over the observed z_{ij}. In the second term, we put in expected values for the z_{ij} based on the initial estimates $P(\gamma_i \mid C_j; \Theta)$ and $P(C_j; \Theta)$. 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**)), we put in new expected values and repeat the M-step. 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. We observe that if λ is 1, then we only use training data and our methods correspond to naïve or general Bayes methods in which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler (1988).

Belin and Rubin (1995) were the first to provide an (reasonably accurate) unsupervised method of estimating false match rates. Larsen and Rubin (2001) later used semi-supervised learning and MCMC methods to provide false match rate estimates. Winkler (2002) used EM methods to provide estimates of false match rates that were very slightly less accurate than those of Larsen and Rubin (2001) but for which computation was 100 times as fast (10 minutes per estimate). The additional speed was needed because the methods needed to be applied in ~500 regions into which Decennial Census matching had to be performed in 3-6 weeks. Winkler (2006a) provided unsupervised methods that improved over Belin and Rubin (1995) because he was able to create data structures that accounted for more of the relationships between records than had been available for Belin and Rubin. The Winkler (2006a) unsupervised methods were somewhat worse than the semi-supervised methods of Larsen and Rubin (2001) and Winkler (2002). Winkler (2004) provided methods of false nonmatch error rate estimation.

2.4. The Fellegi-Holt Model of Statistical Data Editing

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.

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 (2008), 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 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 (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 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 2008). 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 2006b, 2008; also 2010). The methods create records that *always* satisfy edits and preserve joint inclusion probabilities.

The generalized software (Winkler 2010) incorporates ideas from statistical matching software (Winkler 2006b) that can be compared to ideas and results of D'Orazio et al. (2006) and earlier discrete-data editing software (Winkler 2008) that could be used for synthetic-data generation (Winkler 2010). 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 2008, 2010). 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 2006b) 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 (2008) 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 (2010) 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.

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.

2.5. Current Models for Adjusting Statistical Analyses for Linkage Error

The first 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. All 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 $A \times B$. 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 $A \times B$ when no training data is available.

Lahiri and Larsen (2005) later extended the model of Scheuren and Winkler with a complete theoretical development. In situations where the true (not estimates) 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.

Chipperfield et al. (2011) provided methods of extending analyses on discrete data. The Chipperfield et al. methods are closely related to Winkler (2002) which contains a full likelihood development. Trancredi and Liseo (2011) applied Bayesian MCMC methods to discrete data. The Trancredi-Liseo methods are exceptionally 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 3 hours computation on each block (approximately 50-100 households). There are millions of blocks in the U.S.

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 $x \in A$ to $y \in B$ 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 $A \times B$. Although we have not encountered situations similar to Goldstein et al. (2012) where estimates of matching probabilities are very highly accurate and where we can obtain highly accurate estimates of relationships for (x, y) pairs on $A \times B$ (particularly from previous matching situations), the Goldstein et al. methods are highly promising, possibly in combination with another method.

The methods for adjusting regression analyses for linkage error 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.

Because the model of Chipperfield et al. (2011) is straightforward and deals with discrete data, we describe it prior to going to the examples that illustrate the extreme difficulties of having linkage-error-adjustment models that effectively deal with data that might be appropriately described as real-world.

2.6. Errors in Statistical Analyses due to Linkage Error of Discrete Data

The natural way of analyzing discrete data is with loglinear models on the pairs of records in $A \times B$. For consistency with Chipperfield et al. (2011) we follow their notation as consistently as possible. Rather than break out A and B as $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_m)$, we 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 $A \times B$ with x in A and y in B. If we knew the truth, we would know all the p_{xy}. We wish 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.

As preliminary notation, we describe contingency tables without the missing data. We assume that x takes G values and y takes C values. Then, the joint distribution of x and y is

$$p(x, y) = p_1(y \mid x, \Pi)\, p_2(x)$$

where $\Pi = (\pi'_1, \ldots, \pi'_)'$, $(\pi'_g = ((\pi'_1|g, \ldots, \pi'_C|g)'$, $\pi'_c|g$ is the probability the given x = g that y = c. The total number of probabilities p(x, y) is CG. Each p(x,y) is obtained by summing over all pairs $p_i(x,y)$ where x in A (first component) and y in B (second component). The standard estimate of

$\tilde{\pi}_{c|x} = n_{c|x}/n_x)$, where $n_x = \Sigma_c \Sigma_i\ w_{ic|x}$, where $w_{ic|x} = 1$ if $y_i = c$ and $x_i = x$ and $w_{ic|x} = 0$ otherwise.

When there is linkage error, we are concerned with methods that adjust for linkage error. If we can observe true matching status, then the underlying truth representation is.

$w^*_{ic|x} = 1$ if $y^*_i = c$, and $x_i = x$; else $w^*_{ic|x} = 0$.

Ordinarily, we may need to take a (possibly very large) sample to get at the truth and use the following semi-supervised learning procedure.

Take a (likely very large) sample s_c to get (possibly only somewhat) good estimates of $w^*_{ic|x}$. Use EM model to get estimates for all $\hat{p}_{xy}$. The sample s_c gives

$$\hat{p}_{xy*} = (\Sigma_{sc}\, w^*_{ic|x}\ \delta_i)/(\Sigma_{sc}\, w^*_{ic|x}). \qquad (7)$$

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x}/(\Sigma_c\, \tilde{n}_{c|x})^{-1}, \qquad (8)$$

where

$$\tilde{n}_{c|x} = \Sigma_i\, \tilde{w}_{ic|x}, \qquad (9)$$

$$\tilde{w}_{ic|x} = w^*_{ic|x}\ \hat{p}_{xy^*} + (1- \hat{p}_{xy^*})\ \tilde{\pi}_{c|x} \text{ if } i \notin s_c,$$

$$= w^*_{ic|x} \text{ if } i \in s_c \qquad (10)$$

$$= \tilde{\pi}_{c|x} \text{ if } i \in s_c \quad \text{and } \delta_i = 0 \ \ (\delta_i \text{ is indicator that true match}),$$

The (semi-supervised) EM procedure is

1. Calculate $\hat{p}_{xy^*}$ from (7),
2. Initialize $\tilde{\pi}_{c|x}^{(0)}$ and then calculate $\tilde{w}_{c|x}^{(0)}$ from (10) and then $\tilde{n}_{c|x}^{(0)}$from (9),
3. Calculate $\tilde{\pi}_{c|x}^{(t)}$ from (8) using $\tilde{n}_{c|x}^{(t-1)}$,
4. Calculate $\tilde{w}_{c|x}^{(t)}$ from (10) using $\tilde{\pi}_{c|x}^{(t)}$ and then calculate $\tilde{n}_{c|x}^{(t)}$from (9) using $\tilde{w}_{c|x}^{(t)}$,
5. Iterate between 3 and 4 until convergence.

A similar semi-supervised procedure (with full likelihood development) was used in Winkler (2002) and extended to an unsupervised procedure (Winkler 2006a) with a substantial decrease in accuracy for estimating false match rates. With very slight notational changes, the procedure given in steps 1-5 above is the same as the approach using the likelihood given by equation (6).

The procedure of Chipperfield et al. (2011) appears to work well in their simple empirical examples that have substantial similarity to Winkler (2002) but the methods of Chipperfield et al. are more directly generalizable.

Empirical example (Chipperfield et al. 2011):
Three values (employed, unemployed, not in labor force) are compared against same values in another file for a later time period. The total sample size 1000 which represents ~100 for each combination of cells across time periods. With Chipperfield et al., there is very little variation between the 3 labor-force values in one time period to another. There are only 3 × 3 possible patterns. With more realistic data, we might have thousands or millions of patterns. Each false match (x, y) might associate a completely unrelated y ∈ B that is chosen approximately randomly from thousands of B records .

3. Empirical Data

The empirical data consists of 55926 records from on State (1% sample) from a public-use file. In the following diagram, we have collapsed a number of the value-states of fields into a smaller number of value-states to make the analysis easier. There are approximately 1.5 million possible data patterns. Even if this relatively straightforward situation, it will be apparent that it is very difficult to extend the existing statistical-adjustment procedures to achieve high or moderate accuracy with complicated real data.

Table 5. Data (2000 PUMS data for one State)
(Number of values for each field)

-------- A data ----------					-------- B data ------		
Sex	age	race	marit	educ	occup	house	income
2	16	2	5	16	3	5	40

2560 data patterns 600 data patterns

Matching error (Table 6) was induced at the following rate in parts of the file at rates that might correspond to a 'good' matching situation with certain types of real data. We only consider the simplest situation where each $x \in A$ will either be matched with the correct $y \in B$ or not. With this simplification, each matching error represents a type of permutation of the records with $y \in B$. Different authors (Scheuren and Winkler 1993, Lahiri and Larsen 2005) have suggested methods for extending methods to the situations where some $x \in A$ do not have a corresponding $y \in B$ and vice versa. Chambers (2009) and Kim and Chambers (2012a,b) have given specific extensions along with empirical simulations with continuous but we will not consider any extensions in this paper.

The last column in Table 7 represents the counts after distortion due to matching error. The next-to-last column are the counts prior to matching error (i.e., truth). The first eight columns are associated with the values $(0, \ldots, n_j\text{-}1)$ associated with the n_j values associated with the jth field. Higher truth counts are usually reduced in the observed data due to matching error. When an initial value (9th column is blank) followed by 1 it is because a new matching pattern is created as a result of matching error. Approximately 16,000 (20% of the counts) have 1 in the 10th column of which 1/7 are false matches. There is no way to distinguish these false matches (presently) except via follow-up of a sample. The counts in the 10th column are such that the loglinear models associated with the initial (true) counts are quite different than the models associated with the final counts (that have no correction for matching error).

Table 6. Sampling Rates by Strata

Split records – induce matching error
(overall matching error 8-10%)

1. 8000 0.01 error
2. 8000 0.02 error
3. 8000 0.05 error
4 8000 0.08 error
5. 8000 0.12 error
6. 8000 0.15 error
7. 7926 0.20 error

Table 7. Sample Data Records
(Counts for true patterns followed by observed patterns)

	True	Observed
1 020 1 5 07 00 00 000	53	50
1 020 1 5 07 00 00 001	3	5
1 020 1 5 07 00 00 004	.	1
1 020 1 5 07 00 00 006	.	1
1 020 1 5 07 00 01 001	1	.
1 020 1 5 07 00 02 000	1	1
1 020 1 5 07 00 02 001	2	2
1 020 1 5 07 00 04 001	.	1
1 020 1 5 07 00 04 004	.	1
1 020 1 5 07 00 04 005	1	1
1 020 1 5 07 00 05 001	1	1
1 020 1 5 07 00 05 005	.	1
1 020 1 5 07 00 05 006	.	1
1 020 1 5 07 00 05 007	.	1
1 020 1 5 07 05 00 000	16	12
1 020 1 5 07 05 00 001	1	1
1 020 1 5 07 05 00 004	.	1
1 020 1 5 07 05 01 000	18	17
1 020 1 5 07 05 01 001	3	3
1 020 1 5 07 05 02 000	32	29

4. Discussion

Observations:

1. The empirical counts from the observed data *Obs* with the specified distortions vary significantly from the original data *Orig*. The loglinear models on *Obs* and *Orig* are very different.
2. The empirical example might have 2560 data patterns in one file and 600 data patterns in another file. This would correspond to a relatively small administrative-list example. If the sample size is 0.01 of A × B, then most small cells with counts 3 or less will be given $\hat{p}_{xy*} = 0$. It seems unlikely that this will yield suitable estimates of cell counts to improve loglinear modeling. Without correction for matching error, this data does not yield loglinear models that correspond to the loglinear models from the original 'truth' data. This means that, due to matching error (8-10%), we cannot reproduce analyses on the original data (even approximately) on the observed data.
3. If the sample size is 0.25-0.50 of the total number of pairs, then too many pairs will need to be reviewed for this procedure to work in practice. Even with a sample size with a proportion on the order to 0.25 pairs it is unlikely that there will be sufficient information to move the estimates of $\tilde{\pi}_{c|x}^{(t)}$ effectively away from the initial default values of 1/600. If there are informative priors for $\tilde{\pi}_{c|x}^{(0)}$, it is unlikely there is sufficient information to move away from the starting informative priors.
4. To drastically reduce the sample size, it is likely that the matching process must be modeled in detail as in Lahiri and Larsen (2005) or Scheuren and Winkler (1993). It seems that such modeling will need two additional layers of likelihood equations and estimation algorithms.
5. Without additional marginal information (from additional files), it is unlikely that is straightforward to pull apart (x, y) pairs that have been brought together erroneously. Subject matter specialists may be able to supply some edit rules that also allow us to pull apart erroneous (x, y) pairs.

6. Being able to use appropriate third-party data (an unusual situation) could reduce matching error (somewhat). In the reduced-matching-error situation, the statistical adjustments for matching error would work (somewhat) better.
7. Software for to implement steps 1-5 of the supervised EM procedure is a straightforward modification of Winkler (2002, 2006a) and also allows interactions and convex constraints. The basic procedures due to the application of the Chipperfield et al. (2011) are approximately equivalent to reducing the errors from 8% with the *Obs* data to 6% with the first set of processed data *Pr1*. This is not a sufficient reduction in error for valid analysis on *Pr1* because it would not approximate an analysis on *Orig*.
8. If we bring in edit/imputation restraints, this should improve error. We obtain additional restraints from subject matter specialists or certain external restraints (such as used in Winkler 2008, 2010 from administrative data). If we further bring in a few crude distributional restraints, then we hope to reduce the error in the resultant processed file *Pr2* to as low as 4% (which we also do not believe is suitably low for analytic purposes). The software, a somewhat straightforward modification/hybrid of the software from Winkler (2002) and Winkler (2010), has not yet been written.
9. Other procedures beyond the combination of Chipperfield et al. (2011) and edit/imputation are almost certainly needed. The most likely initial candidate for expanding methods is the type of statistical matching methods due to D'Orazio et al. (2006) and Winkler (2009).

5. Concluding Remarks

The procedures of Chipperfield et al. (2011) appear to be effective with very simple types of discrete data. The methods of previous authors were never intended for discrete data. In particular the methods of Goldstein et al. (2009, 2012) do not easily extend. Extending the methods for adjusting statistical analyses for linkage error is likely to necessitate much more detailed modeling of the matching process (to drastically reduce sample sizes in the 'truth' follow-up), additional modeling of distributions in contingency tables and the effects of linkage error, edit/imputation methods using additional knowledge from third-party files or subject-matter experts, and possibly ideas from statistical matching (D'Orazio et al. 2006).

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.

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.

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.

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.

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.

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.

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.

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

Winkler, W. E. (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. (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 Reduce Re-identification Risk that Preserve Analytic Properties, http://www.census.gov/srd/papers/pdf/rrs2010-02.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.