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

william.e.winkler@census.gov

ESSnet Data Integration, 24 November 2011

Outline

1. Background
2. Modeling/edit/imputation
3. Record Linkage
4. Adjusting Analyses for Linkage Error
5. Concluding Remarks

A = (a_i, …, a_n, x_1, …, x_k) linked with
B = (b_1, …, b_m, x_1, …, x_k) using common identifying information (x_1, …, x_k)

to produce **A** × **B** = (a_i, …, a_n, b_1, …, b_m) for analyses

Some Issues:

1. Are **A** and **B** nearly complete subsets (and representative of some underlying population)?
2. What about records from A and B that cannot be linked with the correct corresponding records in the other files?
3. What about false matches?
4. What if there are missing items a_i or b_j in files **A** and **B,** respectively? What if there are 'edit' failures?

Fayad, U. & Uthurusamy, R. (2002). Ninety percent of the cost of making a data warehouse is in cleaning up the respective files.

Administrative lists offer great potential but ...

1. They must be cleaned (modeling/edit/imputation)
2a. They must be unduplicated (record linkage)
2b. They must be linked across files (record linkage)

3. Analyses must be adjusted for linkage error

Background on Modeling/Edit/Imputation for Discrete Data

Generalized, parameter-driven methods suitable for use in many different surveys

Based on model of Fellegi and Holt (*JASA* 1976)

Principles

1. The minimum number of fields in each edit-failing record r_0 should be changed to create an edit-passing record r_1 (*error localization*).
2. Imputation rules should be derived automatically from the edit rules.
3. When imputation is necessary, it should maintain marginal and joint distributions of fields.

Current systems do not impute according to any principled methods/models.

Winkler (2003) connected FH editing with imputation as in Little and Rubin (2002, Chapter 13)

Winkler (2008) created fast generalized software for modeling/edit/imputation and *production*. Software suitably fast for all surveys. Demonstrated how methods are much easier to apply and how exceptionally poorly *well-implemented* hot-deck-based methods were.

Winkler (2008) also showed how to scale microdata to external benchmark constraints using convex constraints.

Generalized software modules

1. module *GEN* to find all edits (structural zeros)
2. modeling module *GIFP* (iterative fitting)
3. error localization and imputation module *EL*

Hot-Deck does not preserve joint probabilities

Hot-Deck does not create records that satisfy edits

Imputation Using Winkler (2008, 2010b) *Always* Works

The set covering algorithms (Winkler 1997) are ~100 times as fast those developed by IBM and the modeling algorithms (Winkler 2008, 2010b) are ~100 times as those in commercial software. They are suitable for situations with 10-100 million records.

Record linkage finds duplicates within a list or across lists.

Table 1. Elementary examples of matching Pairs of records (dependent on context)

Name	Address	Age
John A Smith	16 Main Street	16
J H Smith	16 Main St	17
Javier Martinez	49 E Applecross Road	33
Haveir Marteenez	49 Aplecross Raod	36
Gillian Jones	645 Reading Aev	24
Jilliam Brown	123 Norcross Blvd	43

Example of how record linkage is used

Matching two files

Measure a population by capture-recapture

Two independent listings of a set of small geographic regions

Population estimate = $s_A\, s_B / s_{AB}$ where s_A is size of first population, s_B is size of second population, and s_{AB} is size of overlap

Reducing error in measuring s_{AB} is crucial

Clerical matching is too error prone and too slow

Resources for US Decennial Census Matching

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

Figure 1. Log Frequency vs Weight Links

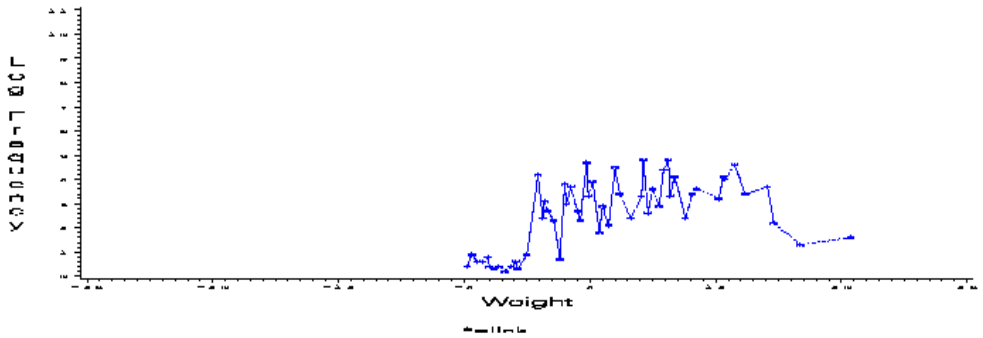


Figure 2. Log Frequency vs Weight Nonlinks

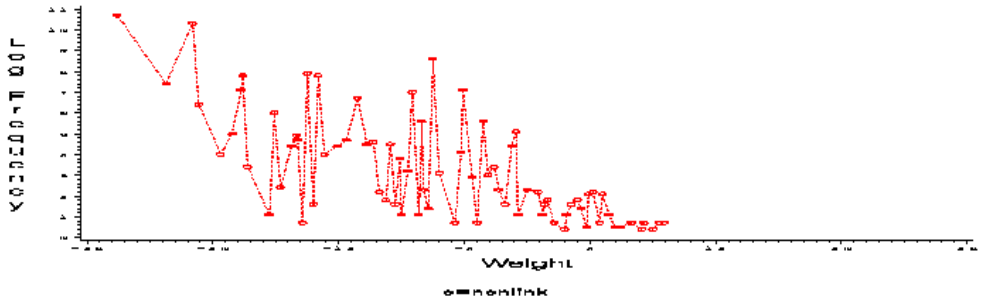


Figure 3. Log Frequency vs Weight Links and Nonlinks Combined

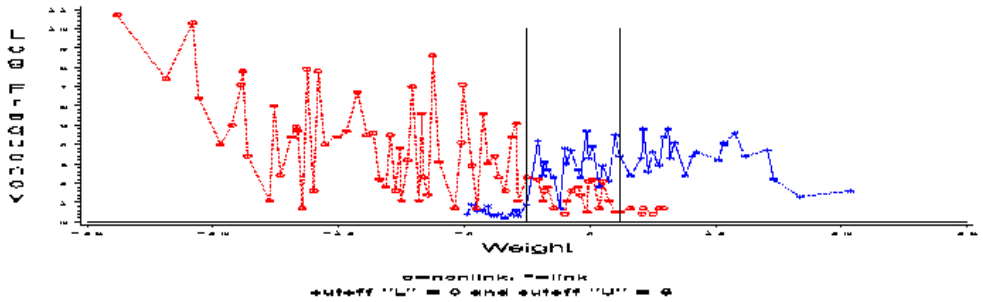


Figure 1. Log Frequency vs Weight Links

Figure 1a. Good Matching Scenario

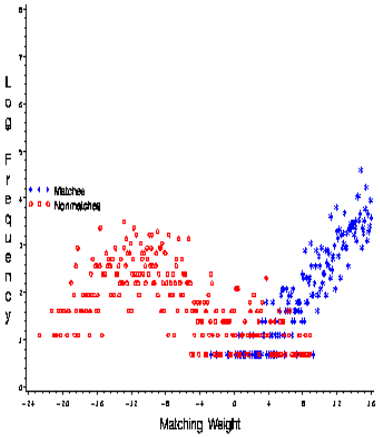


Figure 1b. Mediocre Matching Scenario

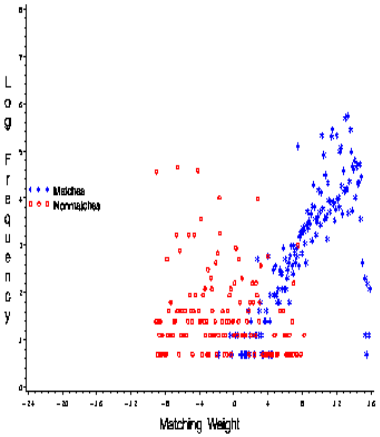


Figure 1c. 1st Poor Matching Scenario

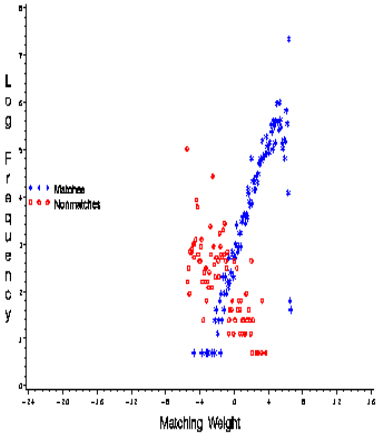
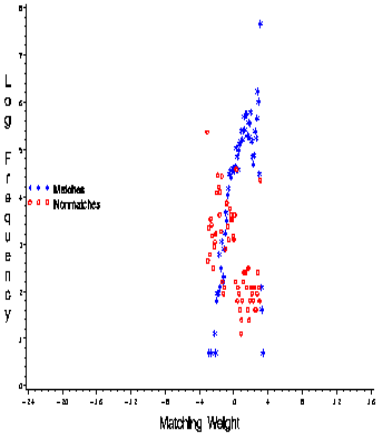


Figure 1d. 2nd Poor Matching Scenario



For 1990 and subsequently, the new methods gave the possibility of high accuracy matching in relatively small geographical areas (blocks, ZIP+4s, contiguous groups of blocks).
With the Decennial Census (300+ million) and administrative lists (billions of records), need to search across very large numbers of records using strings (name, date-of-birth) that might have typographical error.

Represents the same entity

File A	File B
Rcbert Smith	Bob Smlth, Ph.D.
7771 Broacl Street	7711 E. Broad St
March 17, 1977	May 27, 1987
571-222-6666	703-666-2222

Table 5. Final 2010 Decennial Census Blocking Strategy

1. Phone Number
2. State, County, BlockID, first initial of first name, first initial of last name
3. First Initial, Last Initial, Month-of-birth, Year-of-birth
4. State, County, LocalCensusOfficeID, first two letters of first name, first two letters of last name, sex, AgeGroup (0s, 10s, 20s, …)

40+ times as fast as recent prototype parallel software from ANU, Stanford, PSU.

Detailed computation on 10^12 pairs among 10^17 pairs (300 million × 300 million) in 15 hours using *40 cpus* on an SGI Linux machine.

Likely finds 97.5+% of true matches with false match rate <0.5%.

Any software 35% slower is unsuitable for Decennial matching.

Adjusting Analyses for Linkage Error

Two files **A** = (y, m_1, …, m_k) and **B** = (x, m_1, …, m_k)
where (m_1, …, m_k) are quasi-identifying link variables.

Want to do regression of $y = \beta x + \varepsilon$.

1980s: Clean up files, re-do matching with very extensive clerical review and follow-up to (nearly) eliminate matching error prior to doing regression. Issue: very large amounts of individuals for months.

Scheuren and Winkler (1991, 1993) – Develop a procedure that might adjust the regression for linkage error. At a minimum, get an idea of how much improvement in the regression would be possible if (exceptionally) large amounts of clerical review done to improve the matching.

Simulation – Two files A, B where $A \cap B = A, B$.

$y = 200 + 8x + \varepsilon$ where $\varepsilon \sim N(0, 1200)$.

False match error rates - 0.00, 0.02, 0.05, 0.10, 0.20, and 0.50

Table 1. Effect of Matching Error on the Beta Coefficient and the R-square Value

Matching Error	Beta	R-Sq
0.00	7.8	0.82
0.02	7.7	0.80
0.05	7.2	0.69
0.10	6.8	0.68
0.20	6.2	0.52
0.50	4.0	0.21

Figure 1a. 0.00 Matcher Error, Rsq = 0.83, beta = 8.1

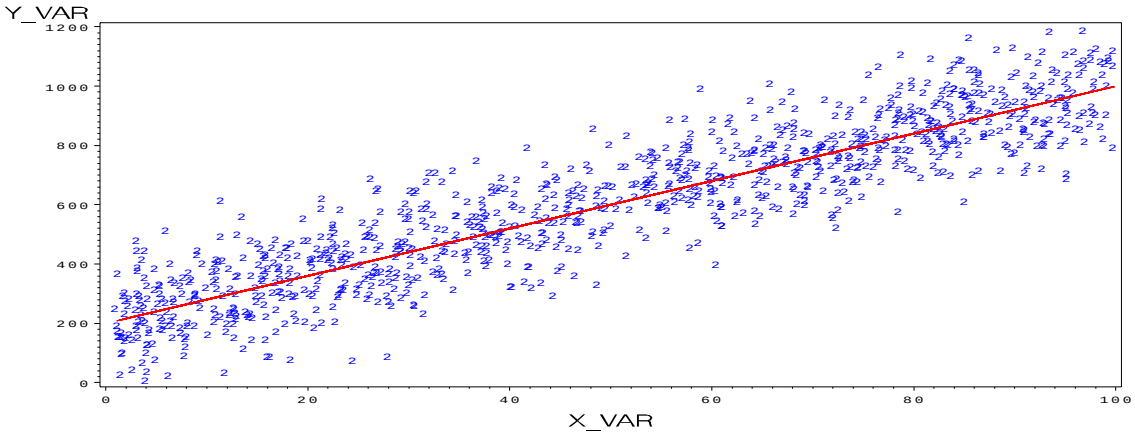


Figure 1b. 0.02 Matcher Error, Rsq = 0.81, beta = 7.9

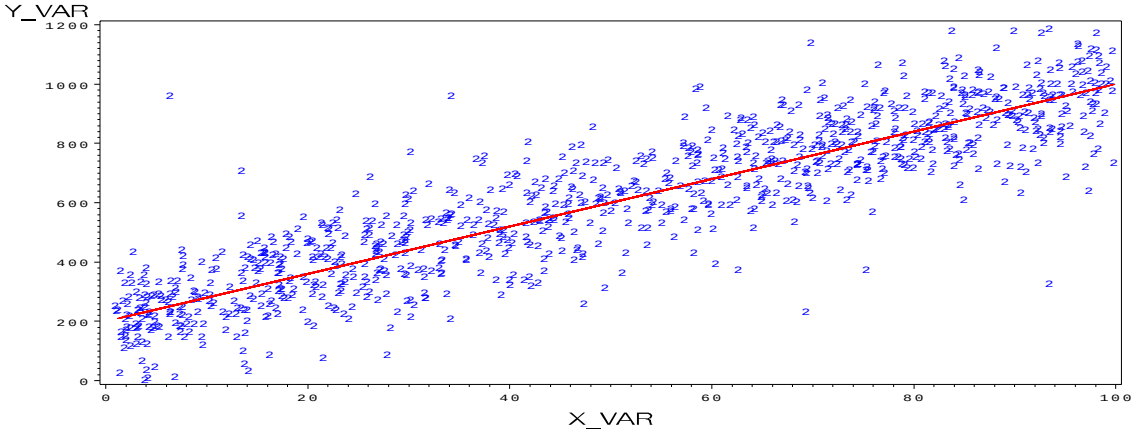


Figure 1c. 0.05 Matcher Error, Rsq = 0.74, beta = 7.8

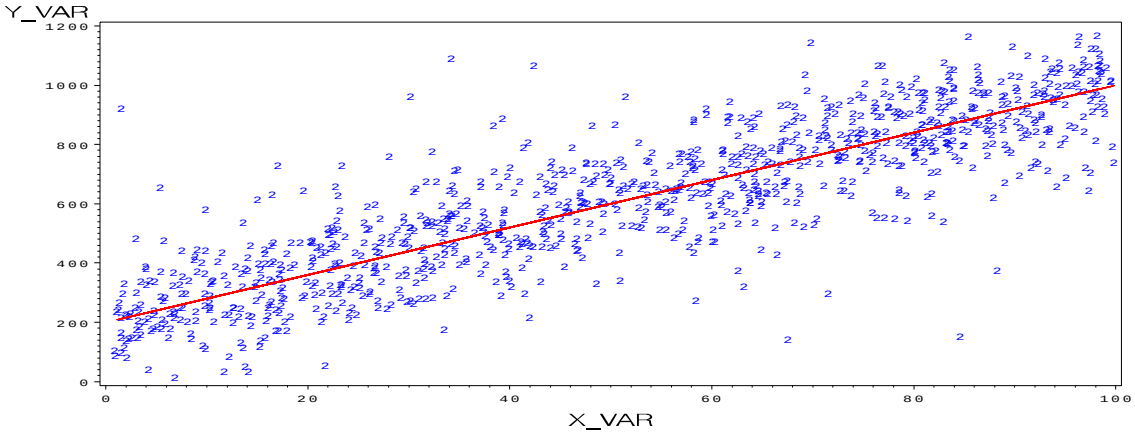


Figure 1d. 0.10 Matcher Error, Rsq = 0.63, beta = 7.0

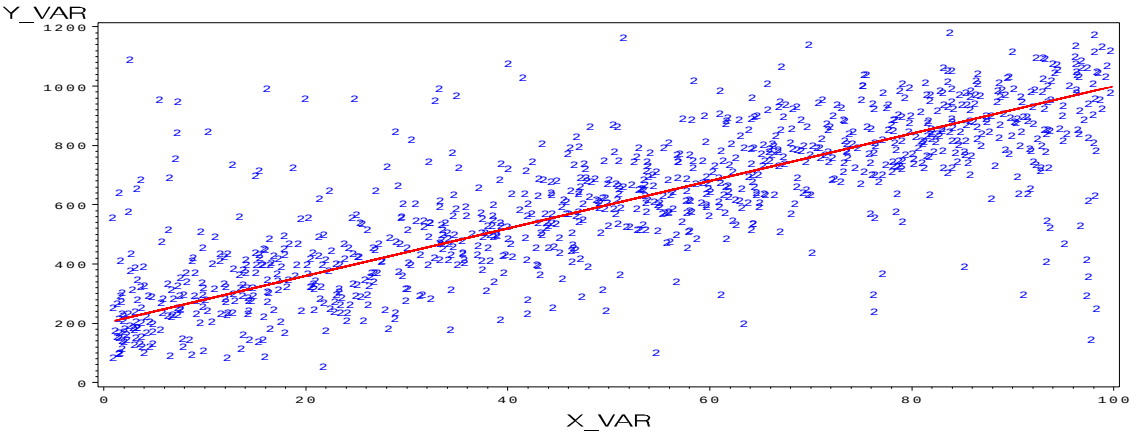


Figure 1e. 0.20 Matcher Error, Rsq = 0.50, beta = 6.2

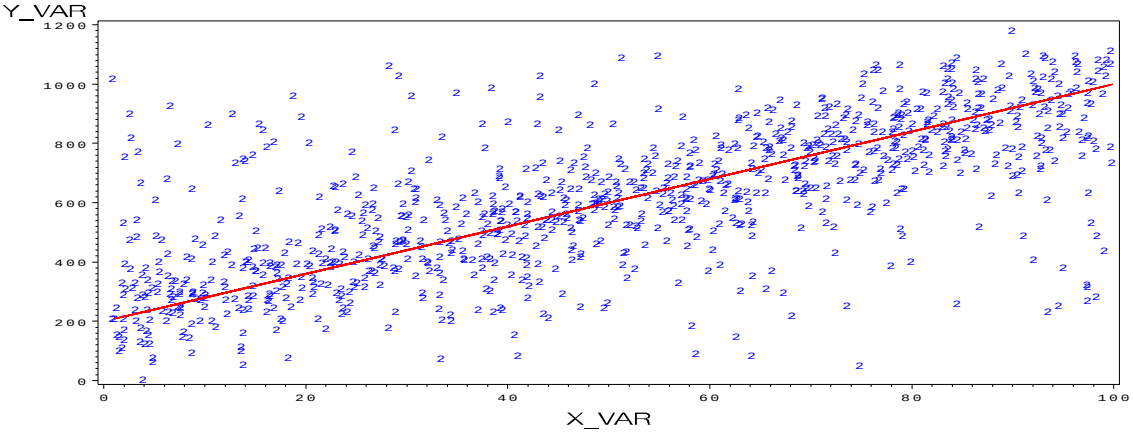
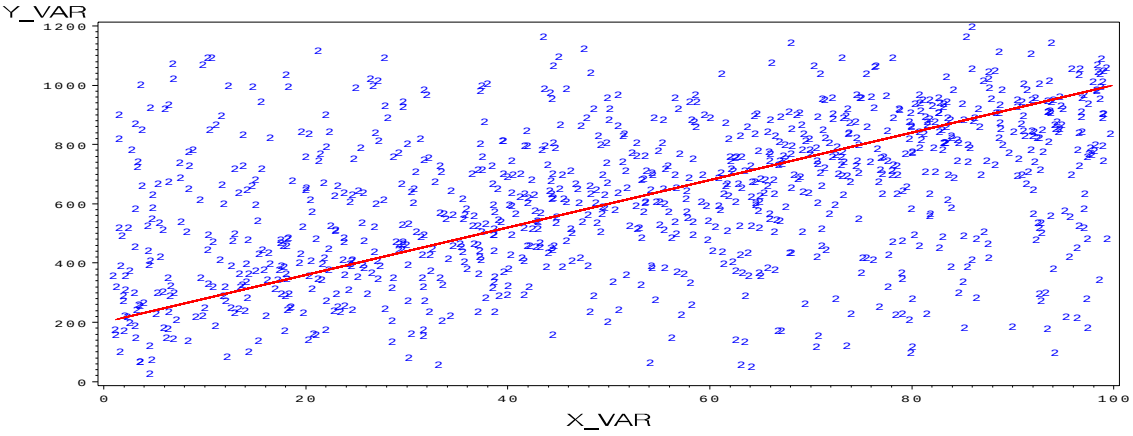


Figure 1f. 0.50 Matcher Error, Rsq = 0.19, beta = 3.8



Scheuren-Winkler (1993) **(**also Lahiri-Larsen 2000, 2005**)**

Files A and B are matched.

$Y = X\beta + \varepsilon$.

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

$p_i + \sum_j q_{ij} = 1$.

$E(Z) = (1/n) \sum_i \mathrm{E}(Z|i) =$

$(1/n) \sum_i (Y_i\, p_i + \sum_j Y_j\, q_{ij}) =$

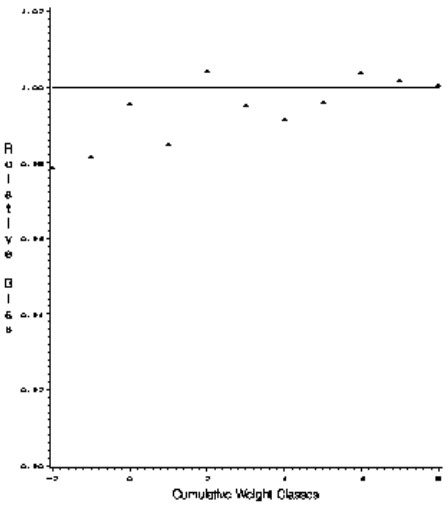
$(1/n) \sum_i Y_i + (1/n) \sum_i [Y_i\, (-h_i) + Y_{\varphi(i)}\, h_i] =$

$\bar{Y} + B,$

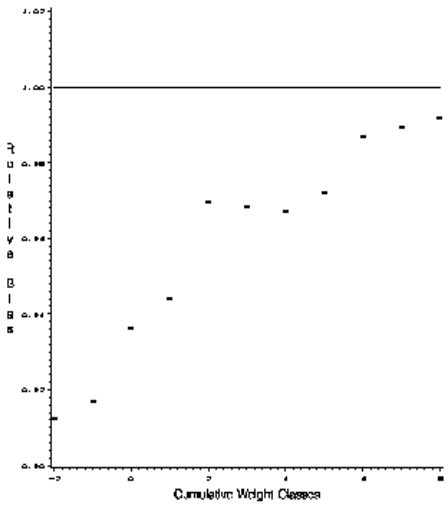
where $h_i = 1 - p_i$.

Under an assumption of 1-1 matching, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let φ be defined by $\varphi(i) = j$.

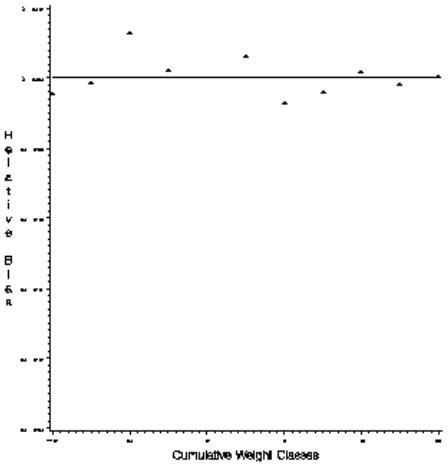
True Probabilities, Adjusted



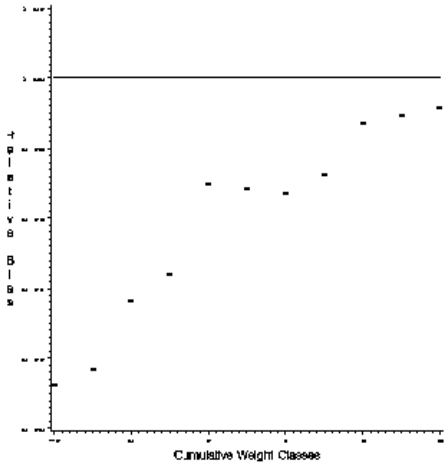
True Probabilities, Unadjusted



Estimated Probabilities, Adjusted



Estimated Probabilities, Unadjusted



Lahiri and Larsen (2005) improve significantly over Scheuren and Winkler but need to the true matching probabilities p_i, q_{ij}, i, j = 1, …, n. Also, they did a much more complete treatment of the multi-variate situation.

Chambers (2009) generalizes Lahiri and Larsen under a drastically simplified record linkage model where $q_{ij1} = q_{ij2}$, $i \neq j1$, $i \neq j2$.

One issue: Nobody has suitable methods for estimating p_i, q_{ij}. Current methods are due to Belin and Rubin (1995) and Winkler (2006).

Concluding Remarks

Modeling/edit/imputation and record linkage are relatively mature technologies in terms of quality. Speed breakthroughs yielding ~100-fold and ~40-fold improvements in speed make the methods promising for administrative lists.

The methods of adjusting analyses for linkage are very much in their infancy and need considerable research.