Record Linkage Methods

William E. Winkler

NISS/Telcordia Workshop on Data Quality

Outline

1. Introduction
2. Fellegi-Sunter Theory
3. Parameter Estimation
4. Name and Address Standardization
5. String Comparators, 1-1 Matching
6. Error-Rate Estimation
7. Analytic Linking
8. Micro-Data Confidentiality
9. Statistical Data Editing
10. Research Problems

Use of Text for Classification

Machine Learning
- Classification of Newspaper Articles into Subject Categories
- Classification into Disease Categories
- Industry and Occupation Coding

Information Retrieval
- Library Document Search
- Web Search

Record Linkage
- Identification of Duplicates Given Name, Address, Age

Model of Fellegi and Sunter (*JASA* 1969)
Newcombe et al. (1959), Tepping (*JASA* 1968)

Files **A** and **B** are matched

Classify pairs from $\mathbf{A} \times \mathbf{B}$ into
Matches M and nonmatches U

$R = P(\gamma \, \varepsilon \, \Gamma \mid M) \, / \, P(\gamma \, \varepsilon \, \Gamma \mid U)$

γ is an agreement pattern

Each pair is a record to be classified

agree/disagree – yes/no

relative frequency (smith vs. zabrinsky)

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

If $T_\lambda < R < T_\mu$, then designate pair as a potential match and hold for clerical review.

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

μ - bound on false match rate

λ - bound on false nonmatch rate.

Theorem FS (1969). Above decision rule is optimal in the sense that, for fixed bounds on the rate of false matches and nonmatches, it minimizes the clerical review region.

Conditional Independence

P(agree first, agree last, agree age | M) =
P(agree first | M) P(agree last |M) P(agree age| M)

P(agree first, agree last, agree age | U) =
P(agree first | U) P(agree last | U) P(agree age | U)

No Training Data

Optimal parameters vary significantly from one region to the next in the 1990 U.S. Census (Winkler *ARC* 1989b)

Software (Winkler and Thibaudeau 1991) finds optimal yes/no parameters automatically, builds frequency tables automatically that are scaled to yes/no parameters. Entire U.S. (450 regions in 1990) matched in three weeks.

Do not need truth data set. Find optimal parameters
(nearly automatically)
Fellegi-Sunter (FS) – 3 variables, independence
Winkler 1988 EM, independence
Winkler (1989a,b, 1993) general interaction accounting for dependence, convex constraints to predispose probabilities to appropriate regions, relative frequency (Smith vs Zabrinsky)
Larsen 1994, 1996 MCMC
Belin and Rubin JASA 1995 EM
Larsen and Rubin 2000 MCMC

Some papers (e.g. Winkler rr94/05) available at
http://www.census.gov/srd/www/byyear.html.

Figure 1. Log Frequency vs Weight
Links

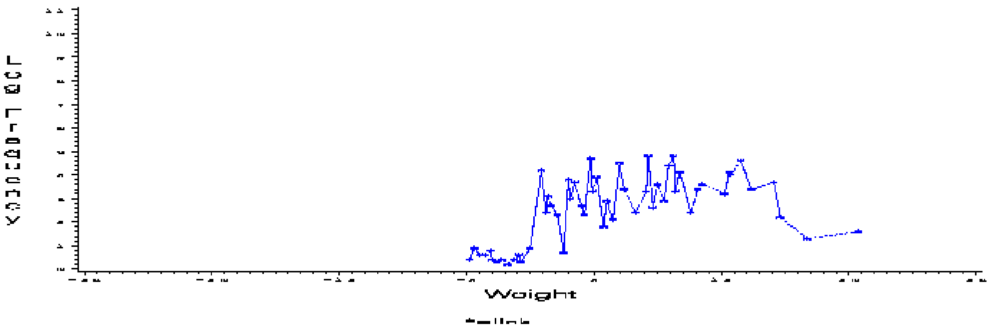


Figure 2. Log Frequency vs Weight
Nonlinks

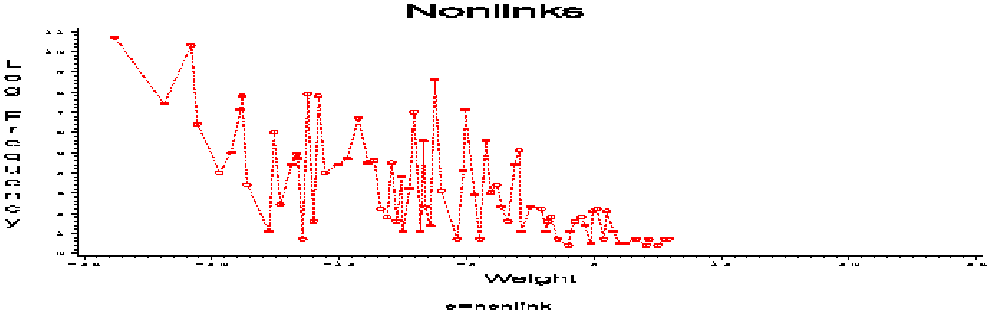
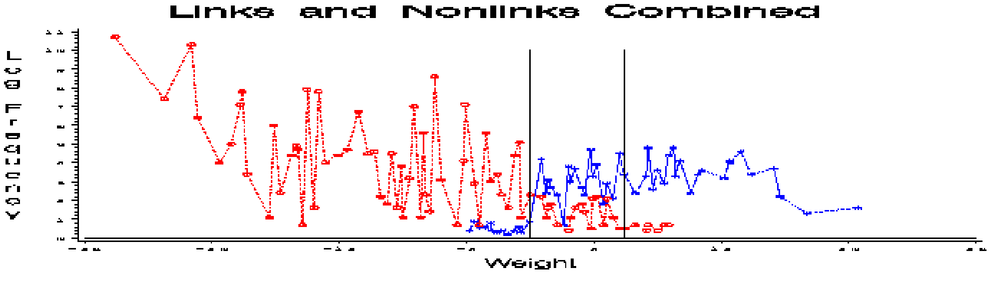


Figure 3. Log Frequency vs Weight
Links and Nonlinks Combined



Bayesian Networks as Special Case of the
Fellegi-Sunter Model of Record Linkage
(FS *JASA* 1969, Winkler ASA 2000)

Strong Statistical Basis for FS Model and Bayesian Networks

Naïve Bayes – Conditional Independence
Nigam, McCallum, Thrun, & Mitchell
Machine Learning 2000

General Bayesian Networks (Interactions)
Winkler ASA 2000

Matching Information

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
Bobbie Sabin	645 Reading Aev	22
Roberta Sabin	123 Norcross Blvd	
Gillian Jones	645 Reading Aev	22
Jilliam Brown	123 Norcross Blvd	43

Need to see the overall lists and the context in which they are used and matched.

Name Parsing and Standardization

Table 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	

Table 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 (1)

	Pre2	Hsnm	Stnm	RR	Box
1.	W	16	Main		
2.				2	215
3.					
4.		14588	HWY 16		

Parsed (2)

	Post1	Post2	Unit1	Unit2	Bldg
1.	ST		16		
2.					
3.				405	Fuller
4.		W			

String Comparators

Bigrams -

Jaro JASA 1989 – insertions, deletions, transpositions

Winkler 1994 – adjustments for agreements on first few characters (Pollock and Zamora *CACM* 1984).

Table Proportional Agreement by
String Comparator Values
Among Matches

	StL	Col	Wash
First			
Φ_n=1.0	0.75	0.82	0.75
$\Phi_n \geq$ 0.6	0.93	0.94	0.93
Last			
Φ_n=1.0	0.85	0.88	0.86
$\Phi_n \geq$ 0.6	0.95	0.96	0.96

Φ_n(Smith, Smith) = 1.0
Φ_n(Dixon, Dickson) = 0.8533.

Table Comparison of String Comparators Using Last Names, First Names, and Street Names

Two strings		String comparator values		
		Jaro	Winkler	Bigram
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.700
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.889
NICHLESON	NICHULSON	0.926	0.956	0.625
JONES	JOHNSON	0.790	0.832	0.204
MASSEY	MASSIE	0.889	0.933	0.600
ABROMS	ABRAMS	0.889	0.922	0.600
HARDIN	MARTINEZ	0.000	0.000	0.365
ITMAN	SMITH	0.000	0.000	0.250
JERALDINE	GERALDINE	0.926	0.926	0.875
MARHTA	MARTHA	0.944	0.961	0.400
MICHELLE	MICHAEL	0.869	0.921	0.617
JULIES	JULIUS	0.889	0.933	0.600
TANYA	TONYA	0.867	0.880	0.500
DWAYNE	DUANE	0.822	0.840	0.200
SEAN	SUSAN	0.783	0.805	0.289
JON	JOHN	0.917	0.933	0.408
JON	JAN	0.000	0.000	0.000

string comparator – model adjustment to likelihood ratios with piecewise linear functions

Truth data set

E.g., for each match know the string comparator values associated with comparisons of first name, last name, etc.

$$P(1 - ((j+1)/50) \leq \Phi_n < 1 - (j/50) \mid M) = m_j$$

$$P(1 - ((j+1)/50) \leq \Phi_n < 1 - (j/50) \mid U) = u_j$$

for j = 1, 2, ..., 50.

1-1 Matching

```
HouseH1     HouseH2

husband
wife        wife
daughter    daughter
son         son
```

$$\begin{matrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \\ c_{41} & c_{42} & c_{43} \end{matrix}$$

4 rows, 3 columns
Take at most one in each row and column

c_{ij} is the (total agreement) weight from matching the ith person from the first file with the jth person in the second file.

Stat. Can. 1987 – greedy algorithm
Jaro 1989 – lsap of Burkard & Derigs 1980
Winkler 1994 – mlsap – same speed as Burkard-Derigs, storage reduced by factor of 500 (100 mB to 0.2 mB), less error

Figure 4a. 1—1 Matching

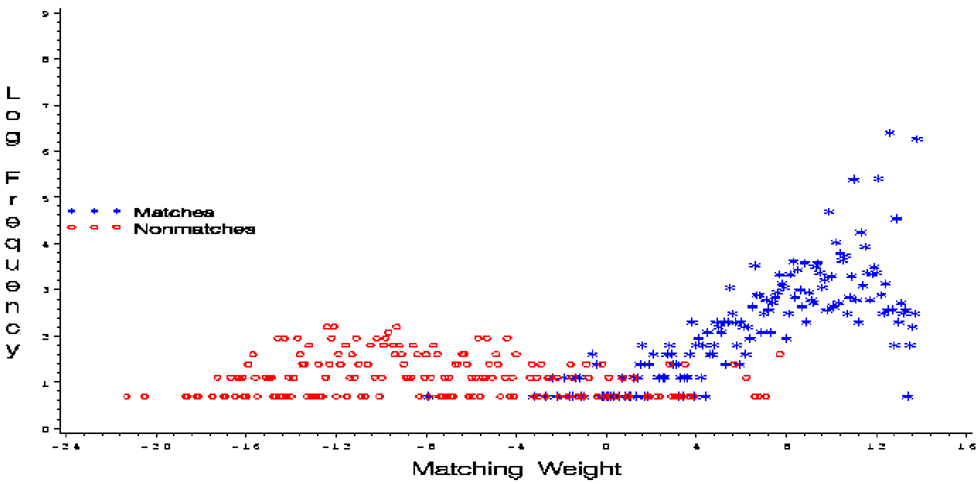
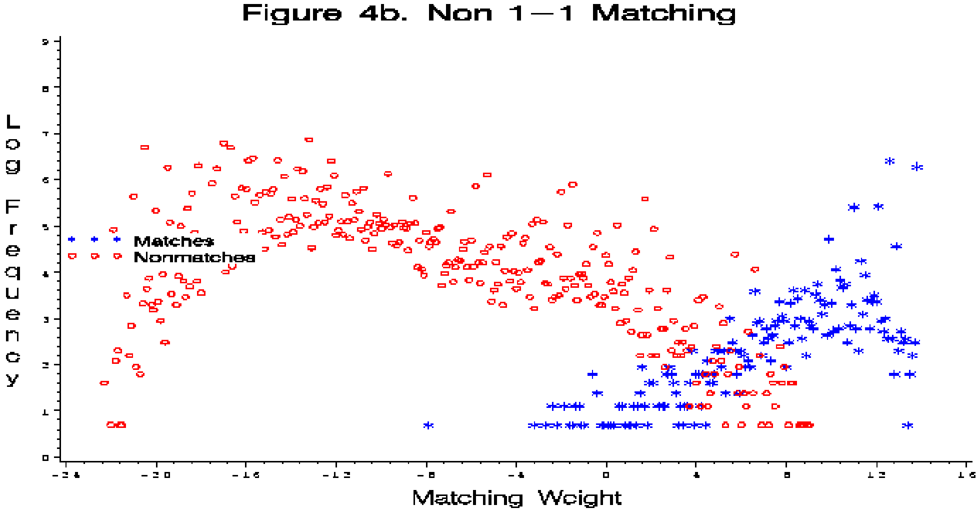


Figure 4b. Non 1—1 Matching



Error Rate Estimation

Belin-Rubin *JASA* 1995 – Training data to get crude idea of shape of curves. X^{α} ($\alpha > 0$) Box-Cox transform. EM to get parameters. Works well in some situations (Scheuren and Winkler 1993). 1-1 matching.

No Training Data – Non-1-1 Matching
Winkler 1993 – Fit interactions. Estimated error rates are accurate

No Training Data – 1-1 Matching
Winkler ASA 1994 EM + ad hoc
Larsen ASA 1996 MCMC + ad hoc
Both less accurate than BR, applicable in more situations

Figure 1. Log Frequency vs Weight
Links

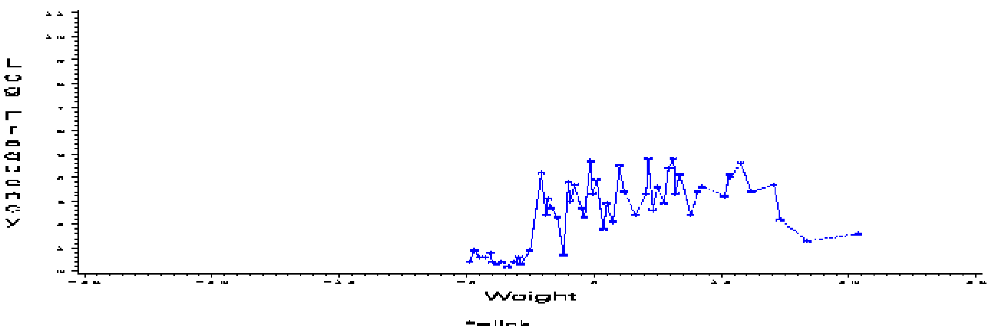


Figure 2. Log Frequency vs Weight
Nonlinks

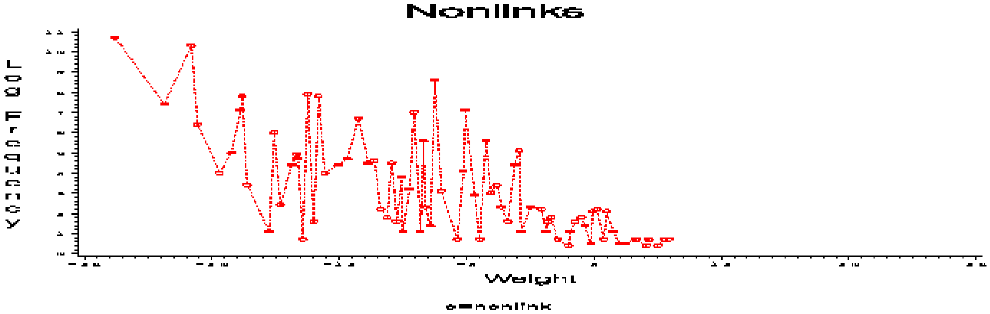
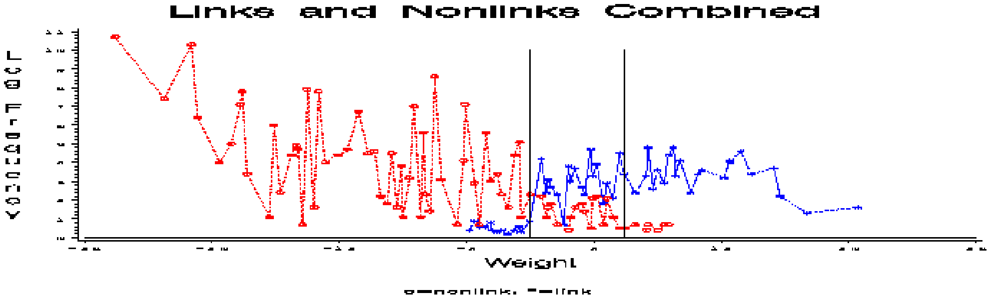


Figure 3. Log Frequency vs Weight
Links and Nonlinks Combined



Blocking

Soundex and NYSIIS encoding of names

The methods are intended to account for very minor typographical variations. NYSIIS is used in many variants of record linkage systems. NYSIIS provides substantially more codes than Soundex. Soundex is easier to describe.

Soundex consists of 4 characters. The first character agrees with the first character of the string such as surname that is being encoded. The next three characters are digits. The end of the Soundex code is '0' filled.

Examples:

Anderson, Andersen	->	A536
Bergmans, Brighma	->	B625
Birk, Berque, Birck	->	B620
Fisher, Fischer	->	F260
Lavoie, Levoy	->	L100
Llwellyn	->	L450

First Pass – ZIP Code + Soundex
Second Pass – ZIP Code + HouseNum
Third Pass – ZIP Code + Age

Estimation of False Non-Match Rates (Missed Matches)

False Non-matches

S_{11}	S_{12}
S_{21}	S_{22}

S_{11} - captured by both blocking criteria
S_{12} - captured by 1^{st} & not 2^{nd}
S_{21} - captured by 2^{nd} & not 1^{st}
S_{22} - captured by neither

$S_{22} = S_{12}\, S_{21} / S_{11}$.

With 3 lists, estimate S_{222}.
With 4 lists, estimate S_{2222}.

Loglinear Models (Bishop, Fienberg, & Holland 1975, Chapter 6). Example in Winkler (1989b)

Adjustment of Analyses for Matching Error

$y = \beta x$

where y from File A, x from File B

Matching variables (name and address) uncorrelated with x and y.

Methods evaluated for varying R-square values, varying amounts of overlap of Files A and B, varying amounts of matching error

Scheuren and Winkler *Surv. Meth.* 1993
  use best 2 matches
Lahiri and Larsen ASA 2000
  use best n matches

Scheuren and Winkler *Surv. Meth.* 1997

Files A and B are matched.

$Y = X\beta + \varepsilon.$

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

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

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

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

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

$$\bar{Y} + B,$$

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 $\varphi(i) = j$.

Figure 1a. Good Matching Scenario

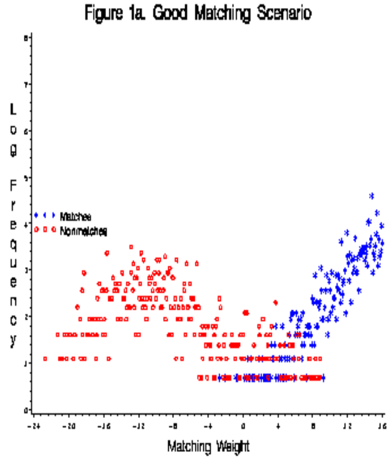


Figure 1b. Mediocre Matching Scenario

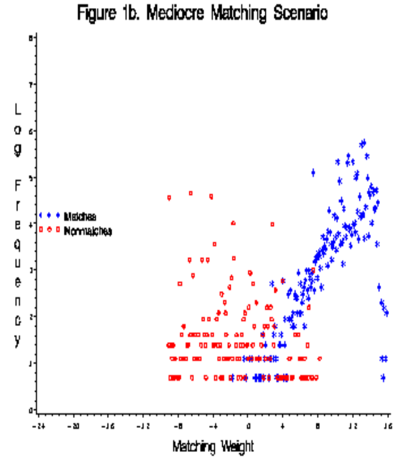


Figure 1c. 1st Poor Matching Scenario

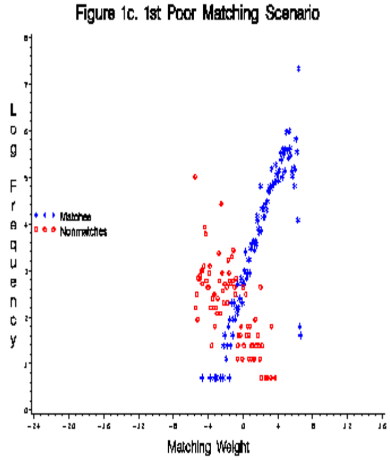
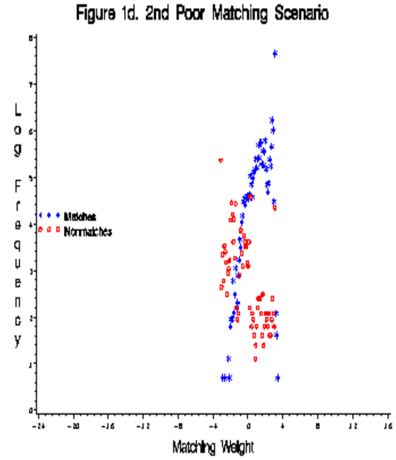
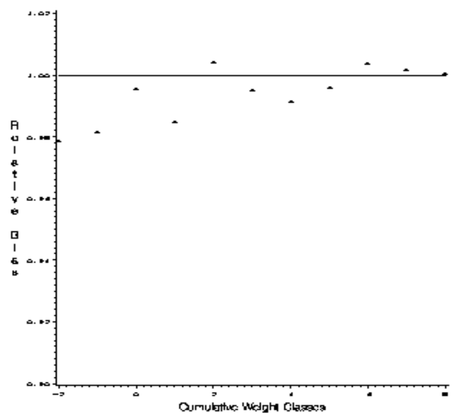


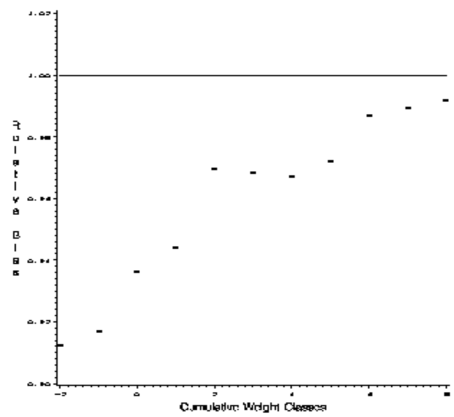
Figure 1d. 2nd Poor Matching Scenario



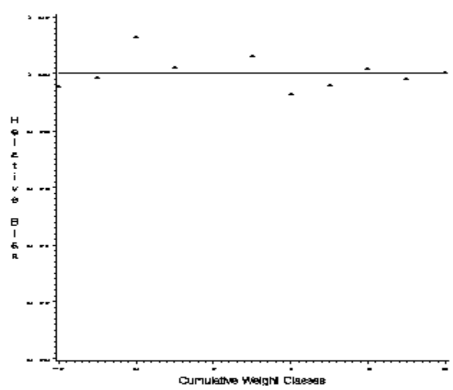
True Probabilities, Adjusted



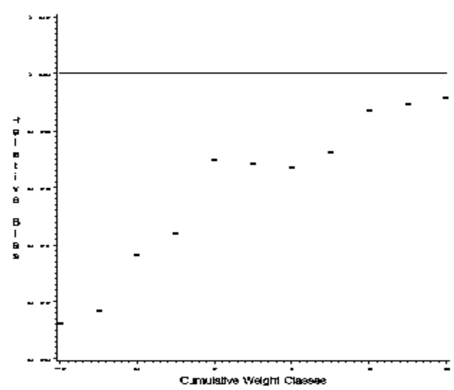
True Probabilities, Unadjusted



Estimated Probabilities, Adjusted



Estimated Probabilities, Unadjusted



Scheuren-Winkler (1997)

$y = \quad \beta x$

where y from File A, x from File B

analytic linking methods take the form

↗RA↘
RL← RA← EI

Linking Files for New Analyses

Economics- Companies

Agency A		Agency B
fuel	------>	outputs
feedstocks	------>	produced

Health- Individuals

Receiving Social Benefits	Agencies B1, B2, B3
Incomes	Agency I
Use of Health Services	Agencies H1, H2

File A	Common	File B
A_{11} , ... A_{1n}	Name1, Addr1	B_{11},...B_{1m}
A_{21} , ... A_{2n}	Name2, Addr2	B_{21},...B_{2m}
.		.
.		.
.		.
A_{N1} , ... A_{Nn}	NameN, AddrN	B_{N1},...B_{Nm}

$$\mathrm{Pred}(A_{Ni}) = B_{Nj}$$

Figure 2a. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, True Data, High Overlap

1104 Points, beta=5.85, R−square=0.43

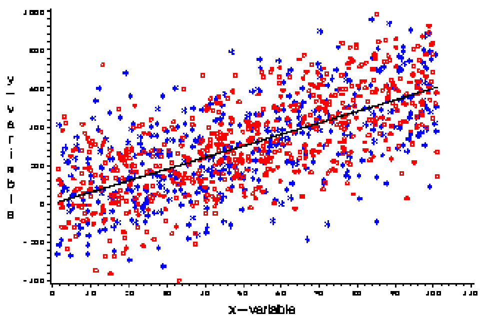


Figure 2b. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, Observed Data, High Overlap

1104 Points, beta=2.47, R−square=0.07

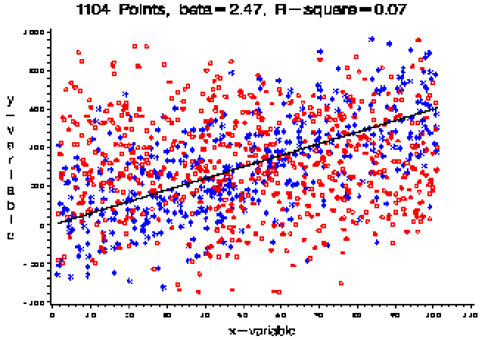


Figure 2c. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, Outlier−Adjusted Data

1104 Points, beta=4.76, R−square=0.40

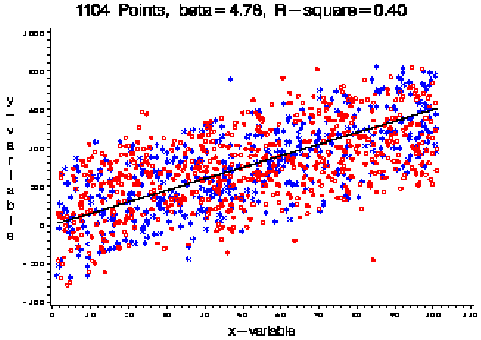


Figure 3a. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, True Data, High Overlap

650 Points, beta=5.91, R−square=0.48

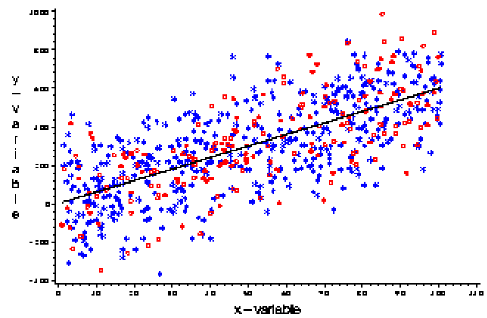


Figure 3b. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, Observed Data, High Overlap

650 Points, beta=4.75, R−square=0.33

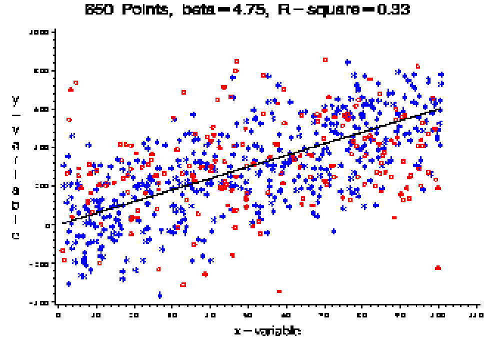
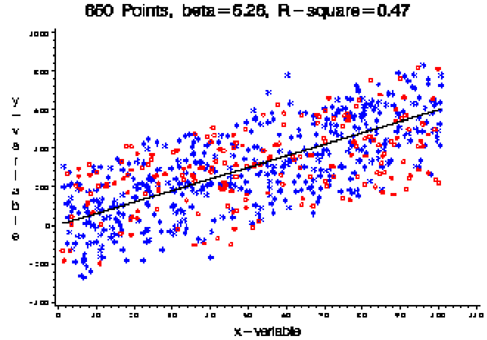


Figure 3c. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, Outlier−Adjusted Data

650 Points, beta=6.26, R−square=0.47



Micro-data Confidentiality

Kim 1986, 1989
Fuller *J. Official Stat.* 1993
Kim and Winkler 1995
Winkler *Res. Official Stat.* 1998
Roque 2000

Additive Noise

$$Y = X + \varepsilon$$

Preserve means and covariances, even on subdomains

Table Increase in re-identification rates using modern record linkage versus simpler methods

Distribution of match probabilities for known vectors of different dimensions in a modified masked released data set of size 150. (Entries are percentages).

	Dimension of known vector					
Match	- Fuller -		--- Winkler ---			
Probability	Four	Eight	Four	Six	Six* ?	Eight
0.0-0.1	51	2	42	4	12	0
0.1-0.2	21	5	0	0	8	0
0.2-0.3	13	2	0	0	10	0
0.3-0.4	4	3	0	0	6	0
0.4-0.5	1	7	0	0	0	0
0.5-0.6	2	20	0	0	0	0
0.6-0.7	1	23	0	0	0	0
0.7-0.8	3	27	0	0	0	0
0.8-0.9	3	11	0	0	0	0
0.9-1.0	1	0	58	96	64	100

*/ Match against 1500 instead of 150.

Statistical Data Editing of Files

Consistency- values do not contradict each other
Completeness – values not missing

Sets of Linear and Discrete Constraints

An edit is a set of points satisfying constraints. A record fails an edit if it is in the set of points defined by the edit.

For continuous x's,

$\Sigma_i a_{ij} x_j \leq C_j$ for j=1,2,…,n.

For discrete,

{Age ≤ 15, marital status = Married}

Fellegi and Holt *JASA* 1976

FH – Check logical consistency of set of edit rules prior to receipt of data (no training data). All edits reside in easily maintained tables. With one pass through a record, it is possible to automatically fill-in contradictory and missing values so that resultant record satisfies all edits. Integer program finds minimum number of fields to change (impute).

Garfinkel, Kunnathur, and Liepins, *Oper. Res*. 1986
Set Covering, Integer Programming
Winkler 1995 – heuristic gets same answer as branch/bound
99+%, up to 1000 times as fast
Winkler 1997 – new set covering up to 100 times as fast as
IBM-ISTAT that uses variant of GKL
Chen 1998
Chen, Winkler, and Hemmig 2000
DeWaal 2000 – Discrete & Continuous

Research Problems

Match Two Administrative Lists Efficiently

550 million records
300 million records
Multiple blocking – 600 trillion pairs

Sophisticated Blocking
Gill 1999, 2001 – UK National Health System.
For residuals not found by conventional blocking, use all words in name, dump components to multiple PCs that grind away.

Winkler and Yancey 2000, 2001 – Small file of 10 million against large file of 500 million.
No formatting or sorting passes of large file.
Matches according to multiple blocking criteria.

Possible Generalizations - Clustering

McCallum, Nigam, and Ungar - KDD 2000

Jagadish, Koudas, Srivastava – SIGMOD 2000
Ferragina, Grossi *JACM* 99

Appropriate Creation and Use of Training Data

Use of labeled (training) and unlabeled data.
Nigam, McCallum, Thrun, Mitchell – *Machine Learning* 2000 – EM methods
Winkler (2000) – general EM methods
Larsen and Rubin (2000) - MCMC

Unsupervised learning – no labeled training data
Winkler (1989a, 1993) EM
Larsen (1996) MCMC

Background: For matching problems, characteristics (agreement patterns) associated with pairs can vary significantly. For large matching problems, too much clerical review (indeterminate regions).

Problem: Find small subset of patterns that can be sampled to yield labeled training data. Based on overall model of patterns and labeled training data, get new estimates of parameters and matching rules to reduce (drastically?) size of clerical review regions.