Methods and Analyses for Determining Data Quality

William E. Winkler (william.e.winkler@census.gov)

ACM Workshop on Information Quality in Information Systems
June 17, 2005

Outline

1. Background/Examples
2. Vision/Issues
3. Edit/Imputation
4. Record Linkage
5. Concluding Remarks

Health Researcher – Data $X = (x_{ij})$, $1 \le i \le m$, $1 \le j \le n$

m records, n columns (fields)

Fields

ID, name, address, doctor1, date-of-birth, weight, sex, treatment1, drug1, drug2, …, cost1a, cost1b, cost1c, …, StartDate, EndDate, bill1, bill2, …

Business or Edit Rule – male hysterectomy (error), drug does not coincide with listed treatment (error), bill1 is too large or too small given then treatment (error)

Male hysterectomy – change sex or change treatment (hysterectomy)?

If values are missing, how to fill them in (impute plausible values)?

Fellegi and Holt (JASA 1976). Edit/Impute – Change fields or impute fields in a manner that minimizes the difference between the changed record *r'* and the original record *r*. The changed record satisfies edit rules.

Issue: Who determines edit rules? If edit rules are failed, which fields to change?

Issue: If values are filled in (imputed), then is it possible to preserve probabilistic distributions (in some sense)?

The Ideal: Data *X* corresponds to some underlying reality.
If so, then the data can be used for *many analyses*.

A goal of edit/imputation: Allow one or two analyses with data *X*.

Available case methods of filling-in missing data do not work.
(only use cases with no missing data)

hot-deck will take a given missing-data record, match on the non-missing values, and select one of the matching records as a donor. It implicitly assumes that there are a very large number of donors so that a distribution $\hat{f}(x_1, x_2, ..., x_n)$ that estimates the true underlying distribution can be estimated.

e.g. data mining - association rules $P(X_1 = x_1, X_2 = x_2)$

Need to fill in missing data in a manner that preserves observed margins and margins that can be plausibly deduced using existing data.

If probabilities of form $P(X_1 = x_1, X_2 = x_2)$ are missing 20% of the time, then we need to fill-in the missing values using *observed data* and *a model*.

Fill-in data according to a hot-deck-like procedure (Little & Rubin 1987, Chapter 9, also 2nd edition 2002)

missing at random – missing values depend on observed values (often reasonable approximation with real data)

missing completely at random – missing values do not depend on observed (*almost never true with real data*)

non-ignorable nonresponse (e.g., high income individuals do not report) need model (or auxiliary information) for plausibly filling in data (in some situations can be approximated via missing at random)

Filled-in distribution changes dramatically from available case

Issue: If data are mined, then missing data must be imputed in a manner that reasonably approximates an underlying ‘reality’ (i.e., joint distributions).

Good imputation often requires subject-matter expertise (knowledge of data)

Imputation should be connected with editing (e.g., do not impute marital status of married to a child of less than 16).

Vision: Analyst with suitable software can do most edit/imputation and record linkage to ‘clean up’ a database or set of databases

Issue: Ignores quality of process (i.e., improve data capture software, survey form, method for getting information into computer). Redman (1996), English (1999), Loshin (2001).

If have ‘best’ process, will still have problems with missing/contradictory data and with duplication.

Issue: Need methods of modeling and estimating functional relationships that are reasonably robust to ‘messy’ data.

Edit/Imputation

Fellegi and Holt (*JASA* 1976) provided a method for identifying contradictory data and correcting it in one pass through the data.

If-then-else edit rules are converted to 'normal form' that reside in (easily maintained) tables. Main logic (set covering and integer programming) does not need to change.

Prior to Fellegi-Holt, as fields associated with a failing edit in an edit-failing record *r* were changed, the new record *r*' would fail edits that *r* had not failed. *Combinatorial optimization* (to 'correct' records).

Winkler (2003) connects edit with modern imputation for discrete data in a manner that 'preserves' probability distributions. Can be extended to continuous data (also Bruni 2003)

Continuous data (linear inequality edits)

SLICE Stat. Neth. (DeWaal, 2003) – suitable for up to 100,000 records.

SPEER (Draper & Winkler 1997) – less general, suitable for 10 million records. Still deals with 'known' edits.

Riera & Salazar (2004, 2005) – very fast, general. Still being tested.

Discrete Data

DISCRETE (Winkler 1997) – very fast ~10^30. Deal with largest survey situations. 100 times as fast as IBM system using Garfinkel, Kunnathor, & Liepins algorithms (*Operations Research* 1986).

DEISIS (Bruni 2001). Also Bruni (2002-2004). Faster, likely better than earlier methods. Still being tested.

Interesting newcomers (logic programming – satisfiability)

10^70, not yet tested with real data

Boskovitz, Goré, & Hegland (2003)

Franconi et al. (2001) – Census Data Repair

Issues:

1. Determining sets of edits.
2. How to do imputation (generally and in specific situations).
3. Determining 'bad' or 'unusual' data. Want to minimize amount of changes ('corrections').
4. Better metrics of quality.
5. How changes ('corrections') to data affect analyses. Will researchers still reach same conclusions with 'changed' data as with underlying 'true' data?

What is currently done.

1. Create truth deck and induce contradiction/missingness.
2. Perform edit/imputation and compare 'corrected' data with original data.

Aggregates (in varying forms) are needed for analyses (Moore & Lee *JAIR* 1998, DuMouchel et al *KDD* 1999, Owen *DMKD* 2003).

Data Quality Metrics (edit/imputation analysis)

Proportion of missing data in a field.
Proportion of imputed values that agree with original values (requires test decks).
How accurately aggregates needed for a particular analysis (from truth deck) are reproduced from imputed data.

Note: Most metrics are currently subjective. They are typically dependent on a particular analytic use.

Outlier detection methods are only suitable in certain contexts. Clean data will have always 'extreme' records.

Record Linkage

Let A and B be two files. We consider A $\times$ B. We wish to delineate matches M (likely duplicates), nonmatches U (likely non-duplicates), and potential matches (clerical review because insufficient information to make a decision).

Consider 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.

The classification rule is given by:

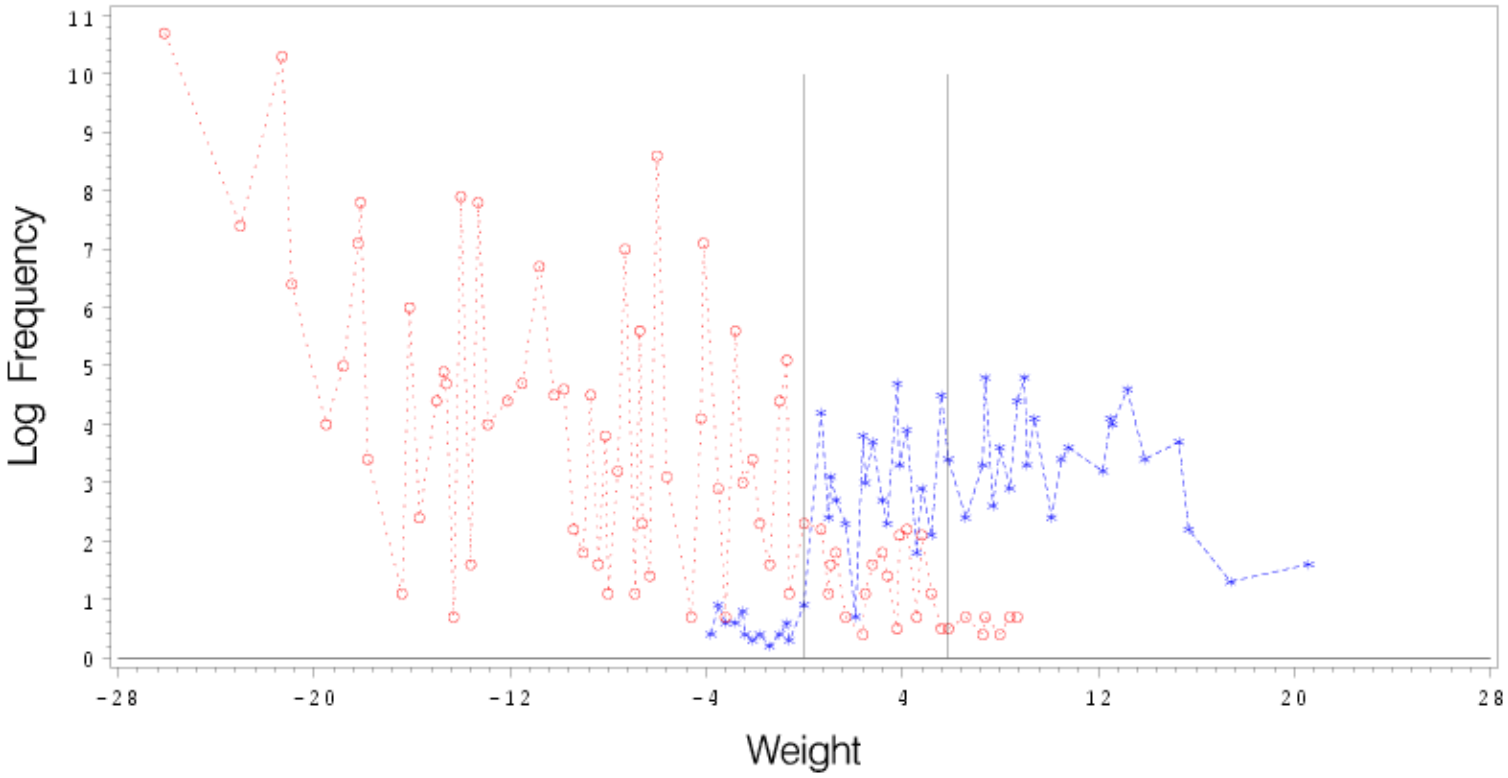
If $R > UPPER$, then designate pair as a link (match).

If $LOWER \leq R \leq UPPER$, then designate pair as a possible link and hold for clerical review. (2)

If $R < LOWER$, then designate pair as a nonlink (nonmatch).

Fellegi and Sunter (JASA 1969). Given fixed upper bounds on the false match and false nonmatch rates, the above rule is optimal in the sense that it minimizes the size of the clerical review region.

Figure 1. Log Frequency vs Weight
Matches and Nonmatches Combined



o=nonmatch, *=match

cutoff "L" = 0 and cutoff "U" = 6

No Training Data

$$P(\gamma) = P(\gamma \mid M)\ P(M) + P(\gamma \mid U)\ P(U)$$

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

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.

P(agree first | M), P(agree last | M) vary significantly
Typographical error rates differ in adjacent regions (suburban versus urban)

Metrics. False match rate, false nonmatch rate.

Do not need truth data set. Find optimal
parameters (nearly automatically)
Fellegi-Sunter (FS) – 3 variables, conditional independence
Winkler 1988 EM, conditional independence (naïve Bayes)
Winkler (1989a,b, 1993) general interaction accounting for
dependence, convex constraints to predispose probabilities
to appropriate regions, relative frequency (Smith vs Zabrinsky)
(Della Pietra et al. 1997 *IEEE PAMI*, Winkler 1990 *Ann Prob*)
Larsen 1994, 1996 MCMC
Belin & Rubin JASA 1995 EM – error rates
Larsen & Rubin JASA 2001 MCMC
Ravikumar & Cohen 2001 *UAI*

Bayesian network (naïve) classification often works well.
Lewis & Ringuette 1994
Ng & Jordan 2002 *NIPS*

Issue: Individual data fields have different representations that make comparisons difficult.

Winkler (1993 – agriculture)

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	

For addresses, use *commercial* software.

Using Hidden Markov (training data)
Borkar, Deshmukh, & Sarawagi 2001 *SIGMOD*
Christen, Churches, & Zhu 2002 *Aus. DM*
Churches, Christen, Lu, & Zhu 2002

Agichtein & Ganti *KDD* 2004 – no training data, need high quality reference file
Cohen & Sarawagi *KDD* 2004 – hidden Markov, tables

Dealing with typographical error

Bigram, Edit Distance (classic computer science methods)

Jaro and Winkler string comparators (Winkler ASA 1990)
Winkler extensions related to ideas of Pollock and Zamora, 1984 *Communications ACM*, modeling using test decks

Bigram easiest to compute, fastest – pairs of characters in common between two strings
Edit distance – slowest – dynamic programming, min number of insertions, deletions, substitutions to get from one string to another

In following, all string comparators scaled between 0.0 and 1.0, where 1.0 represents exact agreement

Table Proportional Agreement by String Comparator Values Among Matches, Key Fields by Geography

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

Φ_n (Smith, Smith) = 1.0 (character-by-character agreement)
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	Edit
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.925	0.818
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.917	0.889
NICHLESON	NICHULSON	0.926	0.956	0.906	0.889
JONES	JOHNSON	0.790	0.832	0.000	0.667
MASSEY	MASSIE	0.889	0.933	0.845	0.667
ABROMS	ABRAMS	0.889	0.922	0.906	0.833
HARDIN	MARTINEZ	0.000	0.000	0.000	0.143
ITMAN	SMITH	0.000	0.000	0.000	0.000

Adaptive string comparators (Hidden Markov)

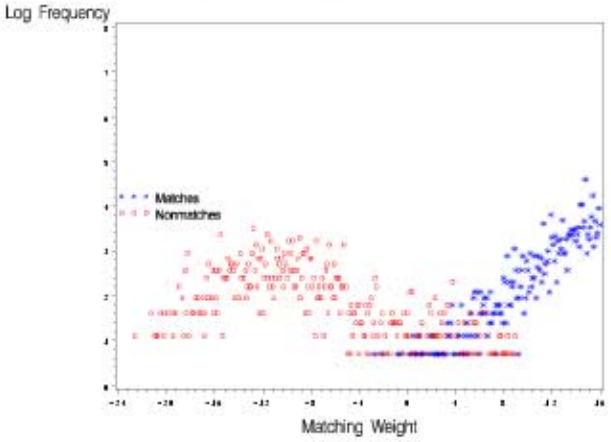
W. Cohen et al. (*IJCAI* 2003, *KDD* 2003) – general string comparator

Yancey (*ASA* 2003, 2005) – typically census-type data

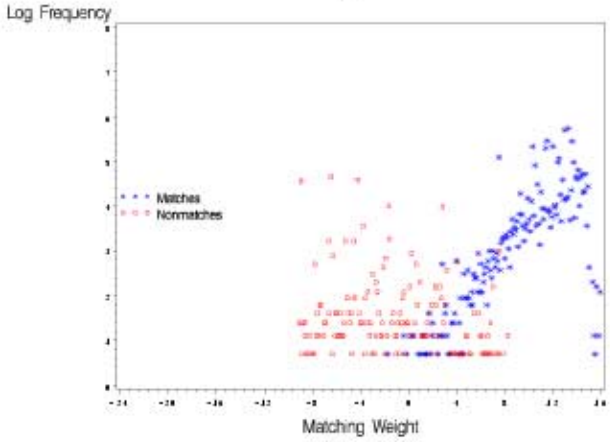
Bilenko & Mooney (*KDD* 2003) – entire free-form names, addr

Bilenko et al. (*IEEE Intel. Sys.* 2003)

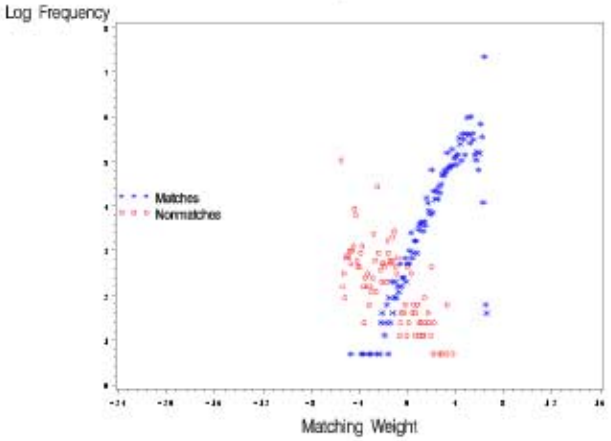
Good Matching Scenario



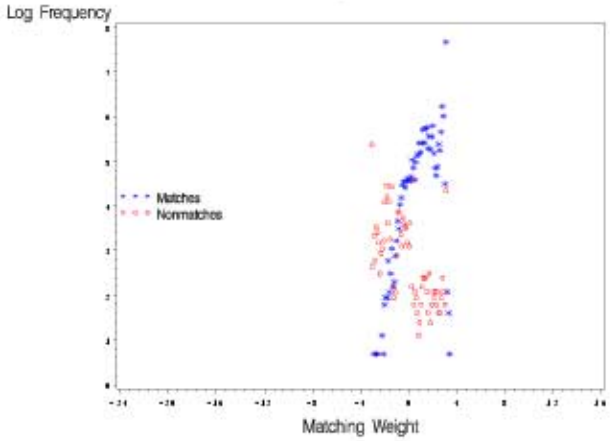
Mediocre Matching Scenario



1st Poor Matching Scenario



2nd Poor Matching Scenario



No Unique Identifiers to Connect Records

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}

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

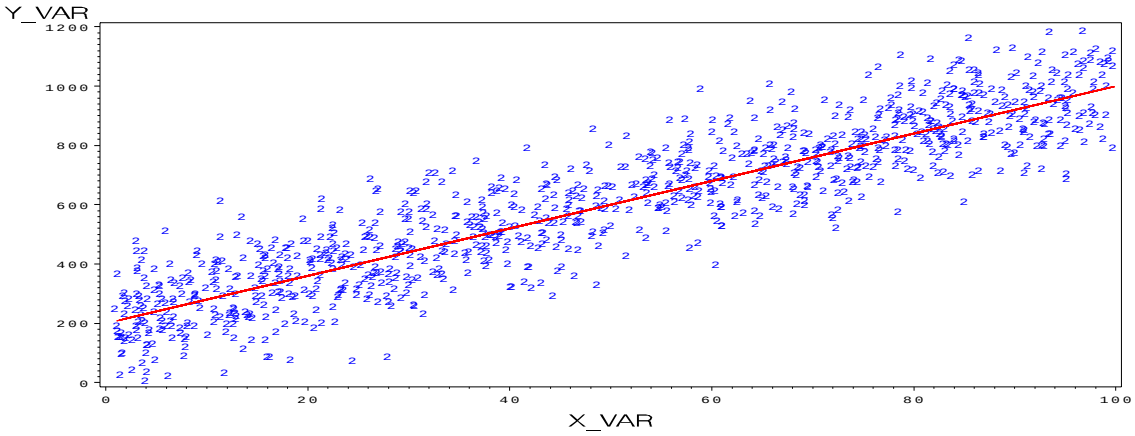


Figure 1b. 0.02 Matcher Error, Rsq = 0.81, beta = 7.9

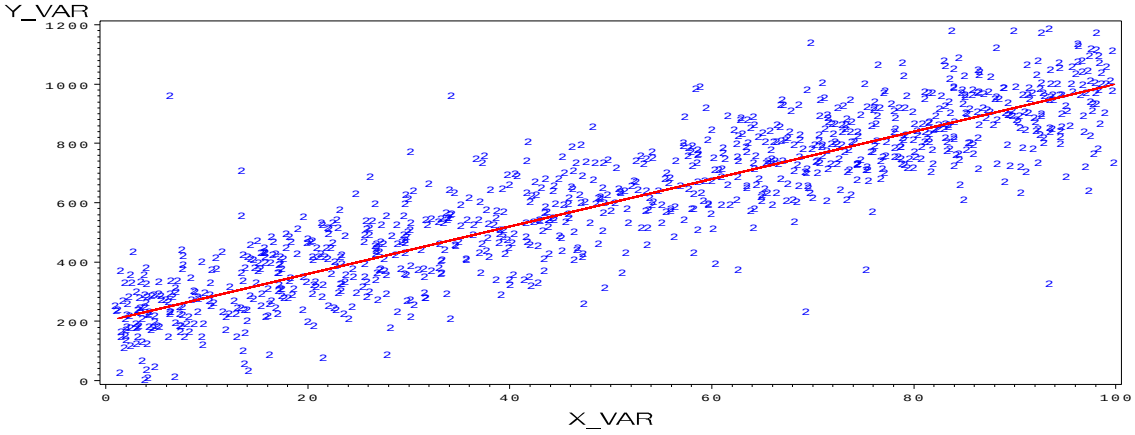


Figure 1c. 0.05 Matcher Error, Rsq = 0.74, beta = 7.8

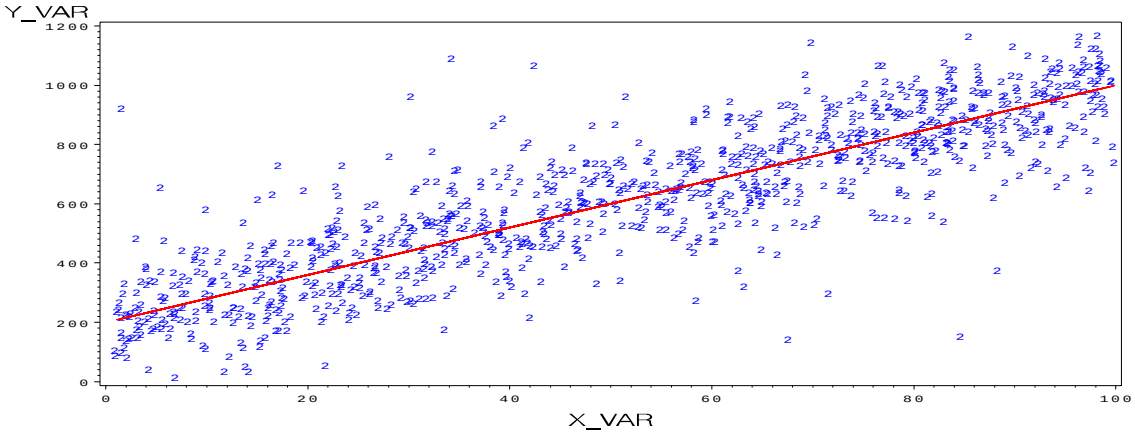


Figure 1d. 0.10 Matcher Error, Rsq = 0.63, beta = 7.0

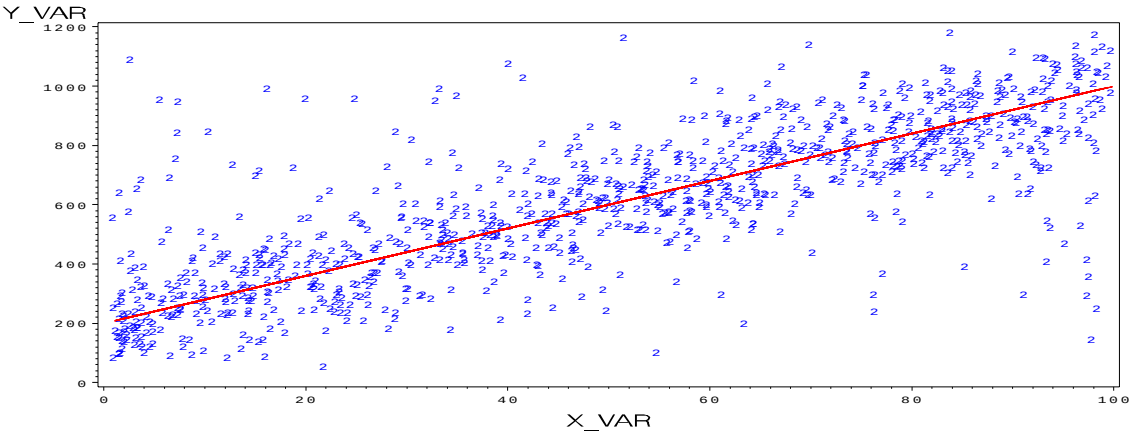


Figure 1e. 0.20 Matcher Error, Rsq = 0.50, beta = 6.2

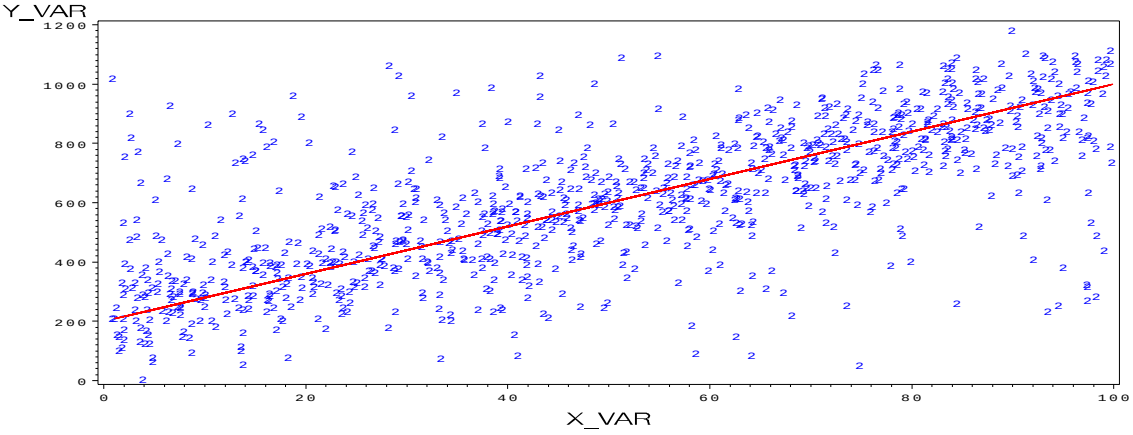
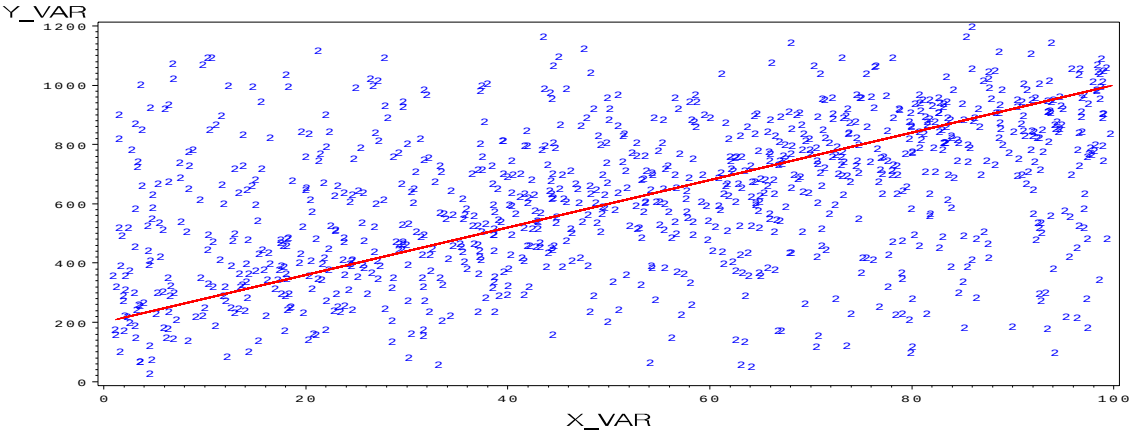


Figure 1f. 0.50 Matcher Error, Rsq = 0.19, beta = 3.8



Issues: Estimate false match rates automatically without training data? Belin & Rubin (1995 JASA) - holds for easiest 2-5%

With training data, regression problem (Vapnik 2000, Hastie, Tibshirani, & Friedman 2001) considered exceptionally difficult.

Small (0.2-0.5%) amounts of labeled combined with unlabeled (semi-supervised) - Larsen and Rubin (2001 *JASA*), Winkler (2002 *ASA*).

Cozman et al. 2003 *ICML* Workshop - Unlabeled data must have same model and come from same distribution as labeled data. Can these ideas be used to leverage small (or no training data)? Are there model classes that will work well (within ε) even though they are not exactly correct?

Estimation of false nonmatch rates without followup. Capture-recapture ideas of Sekar and Deming (1949) applied by Winkler (1989, also 1995).

Create ‘truth’ data for semi-supervised learning.
Naïve Bayes works well for classification. Kick out pairs that are a large subset of matches and nonmatches (with small error).

Training (‘Truth’) and Test Data (Semi-supervised learning)
1990 Census Data with truth and corrections
Only consider pairs agreeing on Census block and
first character of last name

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

	Files		Files		Files	
	A_1	A_2	B_1	B_2	C_1	C_2
Size	15048	12072	5022	5212	4539	4851
# pairs	116305		37327		38795	
# matches	10096		3623		3490	

Intuition (with high quality data):
Above a certain score, all pairs are matches with low error.
Below a certain score, all pairs are nonmatches with low error.

Size of 'truth' data (for training) with error proportions.

	Matches	Nonmatches
A pairs	8817 (.008)	98257 (.001)
B pair	2674 (.010)	27744 (.0004)
C pairs	2492 (.010)	31266 (.002)

Standard semi-supervised model (Nigam, McCallum, Thrun, & Mitchell *Machine Learning* 2000, Winkler 2000, 2002)

Need estimates of distributions of curves of matches, nonmatches (tails), false match rates, false nonmatch rates

$$P(\gamma_i \mid \Theta) = \sum_j^{|C|} P(\gamma_i \mid C_j ; \Theta)\, P(C_j ; \Theta) \qquad (4)$$

$$P(\gamma_i \mid C_j ; \Theta) = \prod_k P(\gamma_{i,k} \mid C_j ; \Theta) \qquad (5)$$

$$P(\Theta) = \prod_j (\Theta_{Cj})^{\alpha-1} \prod_k (\Theta \gamma_{i,k \mid Cj})^{\alpha-1} \qquad (6)$$

$$l_c(\Theta \mid D; z) = \log (P(\Theta)) + (1-\lambda) \sum_{i \in Du} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\, P(C_j ; \Theta)) + \lambda \sum_{i \in Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\, P(C_j ; \Theta)). \qquad (7)$$

where $0 \leq \lambda \leq 1$.

2 or 3 Classes C_i , Equation (5) conditional independence,
Equation (6) Dirichlet prior ($\alpha < 1.1$),
also general interaction (Winkler 1989, 1993, Larsen and Rubin 2001)

EM issues (3-class EM: C_1- match within household, C_2 - nonmatch within household, C_3 – nonmatch outside household

1. Models – CI – independent – *i1 (i0 – 1990 version) I,I,I*
   Larsen-Rubin CI in class 1, 4-way person,
   4-way household in classes 2 and 3, *g1 I,HP,HP*
   Winkler 4+ way interactions in all classes, *g3 (g0 1990 version)*
2. lambda – how much to emphasize training data
3. delta – 0.000001 to 0.001 – smooth out peaks ($\delta = \alpha - 1$)
4. how many iterations (Friedman 2001, numerous ICML 2003)
5. number of degrees of partial agreement
   a. agree, disagree (and/or blank) [small base = 2]
   b. very close agree, moderately close agree, somewhat agree, blank, disagree [large base = 5]

metaparameters – Hastie, Thibshirani, Friedman 2001, Friedman 2001

Figure 1a. Estimates vs Truth, File A
Cumulative Matches, Tail of Distribution
Independent EM, Lambda=0.2

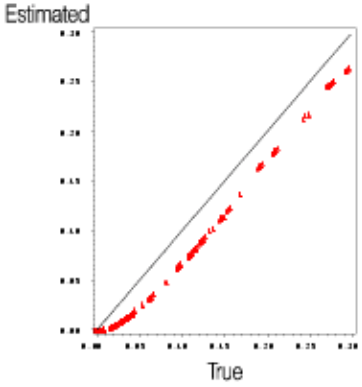


Figure 1b. Estimates vs Truth, File A
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda=0.2

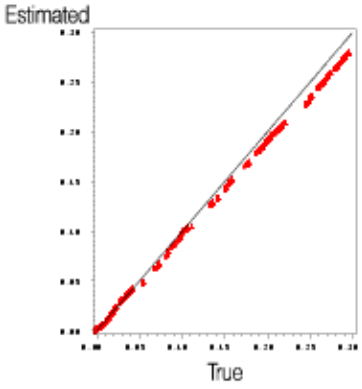


Figure 1c. Estimates vs Truth, File B
Cumulative Matches, Tail of Distribution
Independent EM, Lambda=0.2

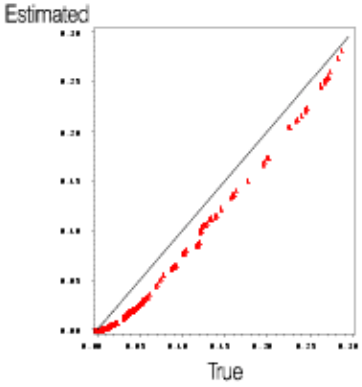


Figure 1d. Estimates vs Truth, File B
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda=0.2

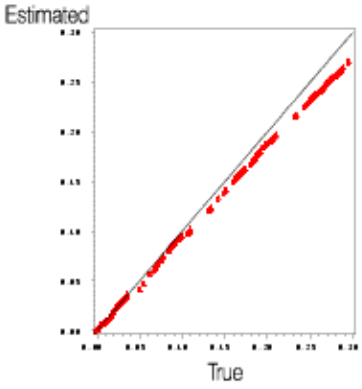


Figure 1e. Estimates vs Truth, File C
Cumulative Matches, Tail of Distribution
Independent EM, Lambda=0.2

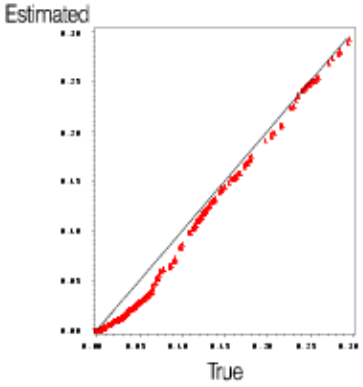


Figure 1f. Estimates vs Truth, File C
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda=0.2

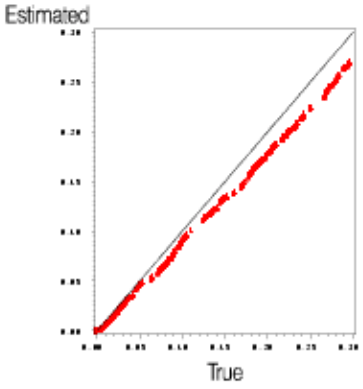


Figure 2a. Estimates vs Truth, File A
Cumulative False Match Rates by Weight
Independent EM, Lambda=0.2

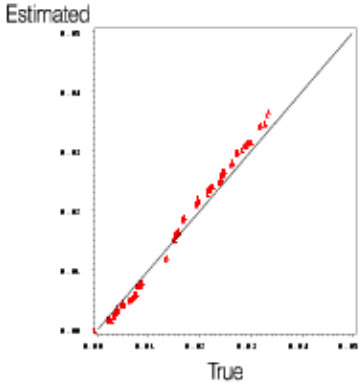


Figure 2b. Estimates vs Truth, File A
Cumulative False Nonmatches by Weight
Independent EM, Lambda=0.2

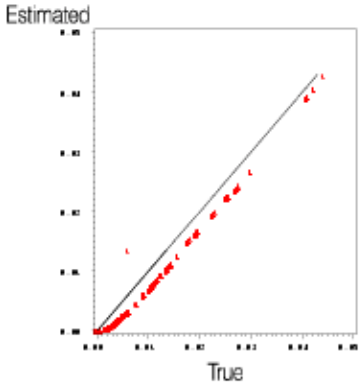


Figure 2c. Estimates vs Truth, File B
Cumulative False Match Rates by Weight
Independent EM, Lambda=0.2

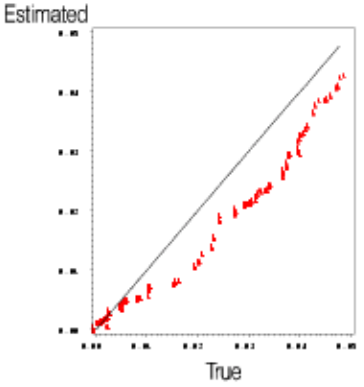


Figure 2d. Estimates vs Truth, File B
Cumulative False Nonmatches by Weight
Independent EM, Lambda=0.2

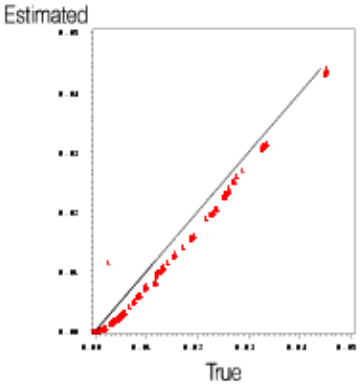


Figure 2e. Estimates vs Truth, File C
Cumulative False Match Rates by Weight
Independent EM, Lambda=0.2

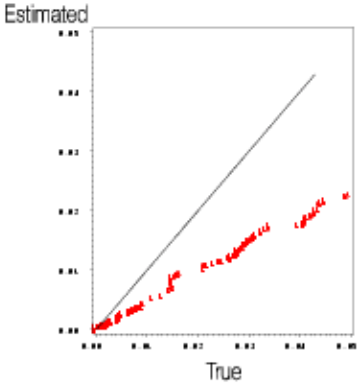
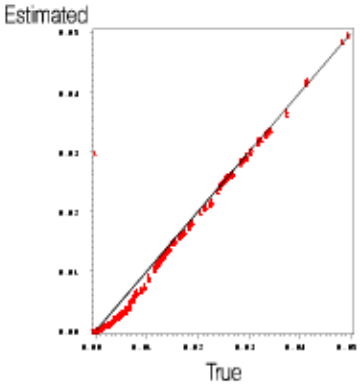


Figure 2f. Estimates vs Truth, File C
Cumulative False Nonmatches by Weight
Independent EM, Lambda=0.2



Issues:

1. Need situation where matches M can be separated with very small error.
2. False nonmatch rates may be susceptible to whether data can be standardized (represented in homogenous manner where fields can be compared).

In many match situations, a subset of the matches (pairs) cannot be brought together because (weakly) identifying information does not bring pairs together.

Ex1: Susan K. Smith 1985Jan07
Karen Jones 1964Apr10
Usually uses middle name, recently changed last name, one date-of-birth completely wrong (business lists worse than person lists.)

BigMatch (Yancey & Winkler 2004)

Tradition matching – sort on criteria, bring together pairs agreeing on sort fields, decide matches, repeat – called *blocking* passes

A – moderate size file ~300 million records
B - administrative list ~4 billion records

Hold A file in memory (100 million records plus indices ~ 4 gigabytes of RAM)
Perform one pass on each file. Never sort large file.

Significant savings due to not sorting large file (0 CPU replaces 5+ days CPU), reduced usage of disk space, much less skilled programmer intervention.

Winkler (2004) – Ten blocking passes with Census data.
300 million $\times$ 300 million = 10^{17} pairs
10 blocking passes = 10^{11} pairs – 3 days CPU time
100,000 pairs per second, easily parallelizable
Obtain 99.5+% of matches in 10^{11} pairs

Truth deck

Methodology for estimating 'missed' matches (see Winkler 2004).
Capture-recapture (Sekar & Deming 1949), loglinear models

Issue: In all sets of files, there will be duplicates that cannot be delineated because (weakly) identifying information is too poor.

Weakly identifying information such as name, address, and date-of-birth.

606,411 matches (truth)

Best 11 – 1350 missed matches, no 3-grams in common
Best 4 {1, 3, 11, 9} – 2766 missed matches
Best 5 {1, 3, 11, 9, 8} – 1966 missed matches

Hard-to-find missed matches (artificial data)
(date-of-birth missing in file B, address missing in file A)

	Household 1		Household 2	
	First	Last	First	Last
HeadH	Julia	Smoth	Julia	Smith
Child1	Jerome	Jones	Gerone	Smlth
Child2	Shyline	Jones	Shayleene	Smith
Child3	Chrstal	Jcnes	Magret	Smith

Issues:

1. We have matched two files A and B. Is the intersection A $\cap$ B a representative subset of A or B? (Zadrozny ICML 2004).
2. We have matched two files A and B. Two metrics are the proportion of A matched and the proportion of B matched. If we know something about the populations (files) A and B, can we determine if the matching rates are reasonable?
3. We bring A with data X together with file B containing data Y. How do we adjust for matching error in any joint analysis of (X, Y)?
4. We make an implicit assumption that, if we have a set that contains all of the pairs that are close according to comparison metrics, then we will have all matches. How do we deal with 'true' matches that cannot be brought together according to the basic comparison metrics?

Precision is the proportion of designed matches that are truly matches. *Recall* is the proportion of true matches that were designated as matches. The first metric is somewhat comparable to false match rate and the second metric is somewhat comparable to false nonmatch rate.

Precision-recall breakeven and F-statistic do not make much sense because the number of nonmatches so greatly exceeds the number of matches. The false nonmatch rate is not easily determined or combined with other metrics.

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}

Adjust (simple) regression analysis for linkage error (Scheuren & Winkler 1993, Lahiri & Larsen 2005 *JASA*)

Microdata confidentiality – Construct metrics using distributional properties of *X* and *Y* data to link files A and B (Winkler 2002, Evfimievski 2004, http://www.cs.cornell.edu/aevf/).

Valid Analytic Relationships ←→ Metrics for Comparing Records in Files (simple situation – business with higher receipts (x-variable) will pay corresponding high taxes (y-variable))

Observation: X and Y data may be better for matching some business databases than name and address information. Alternative is having bridging file (information, Winkler 1999, economists prior to 1999) that contains alias names and alias addressees.

Businesses from two files can be clustered using ZIP code and NAICS industrial-category code. Each business (A-file) associated with 15 or less other business (B-file). Use X, Y data to make match determination.

Example: Steel & Konschnik (1999) – empirical example

Methods: Scheuren & Winkler (1997), also Evfimievski (2004), Lambert (1993), use analytic properties of files to create metrics for linking records.

Issue: How to create (crude) models and systematically create (crude) metrics for bringing together corresponding records using x and y data?

Figure 6b. 1st Pass Matching, Observed Data

1104 Points, beta = 2.47, R – square = 0.07

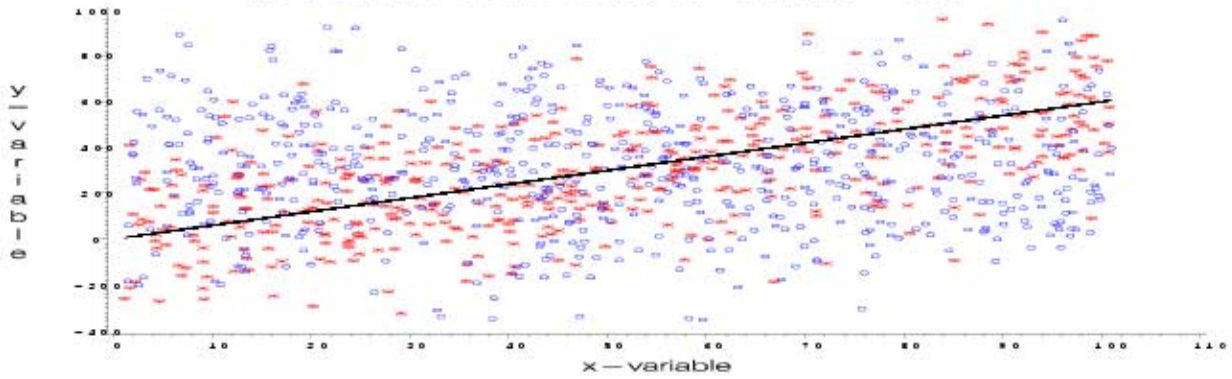
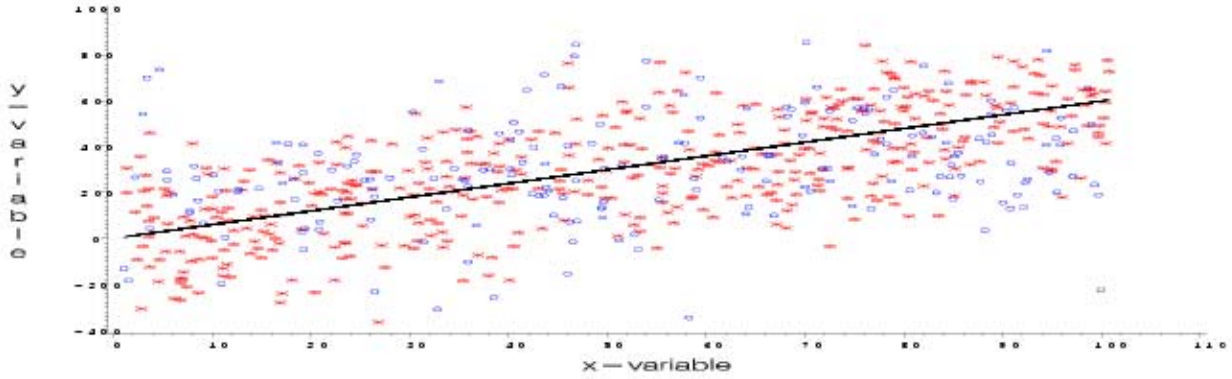


Figure 7b. 2nd Pass Matching, Observed Data

650 Points, beta = 4.75, R – square = 0.33



Suggested Procedure.

1. Clean-up individual files first. Duplicate detection can be done somewhat in parallel with edit/imputation. Try to estimate how many duplicates are being missed (false nonmatches, etc.). Determine whether auxiliary information might help the duplicate detection. If edit rules are available, edit/impute according to the rules. Use outlier detection (with suitable aggregates) if necessary.
2. Merge files according to weakly identifying matching information in files. Evaluate quality of matching. If quality too low, determine if new metrics can be constructed using (x,y) functional relationships or if auxiliary files may be helpful.

Note: It is very difficult to detect errors in data.

Aggregates (in varying forms) are needed for analyses (Moore & Lee *JAIR* 1998, DuMouchel et al *KDD* 1999, Owen *DMKD* 2003).

Concluding Remarks

Almost never are measures of quality (aggregate or otherwise) attached to files.

The standard quality measures (e.g., completeness, accuracy, consistency, etc.) are too general. New specific measures are needed to reflect specific analytic uses of files.

If files are linked then error rates should be calculated (false match, false nonmatch, precision, recall). If precision is high (~ 90%), then any (*x,y*)-analyses (*x* from one file, *y* from another), are likely to be seriously compromised. If precision is low (~50%), then most (*x,y*)-analyses may not be possible.
The *quality of record linkage* (entity resolution) is highly dependent on the quasi-identifying information used in the linkages.