SURV622/SURVMETH622: Record Linkage Applications and Methods

February 3, 2025

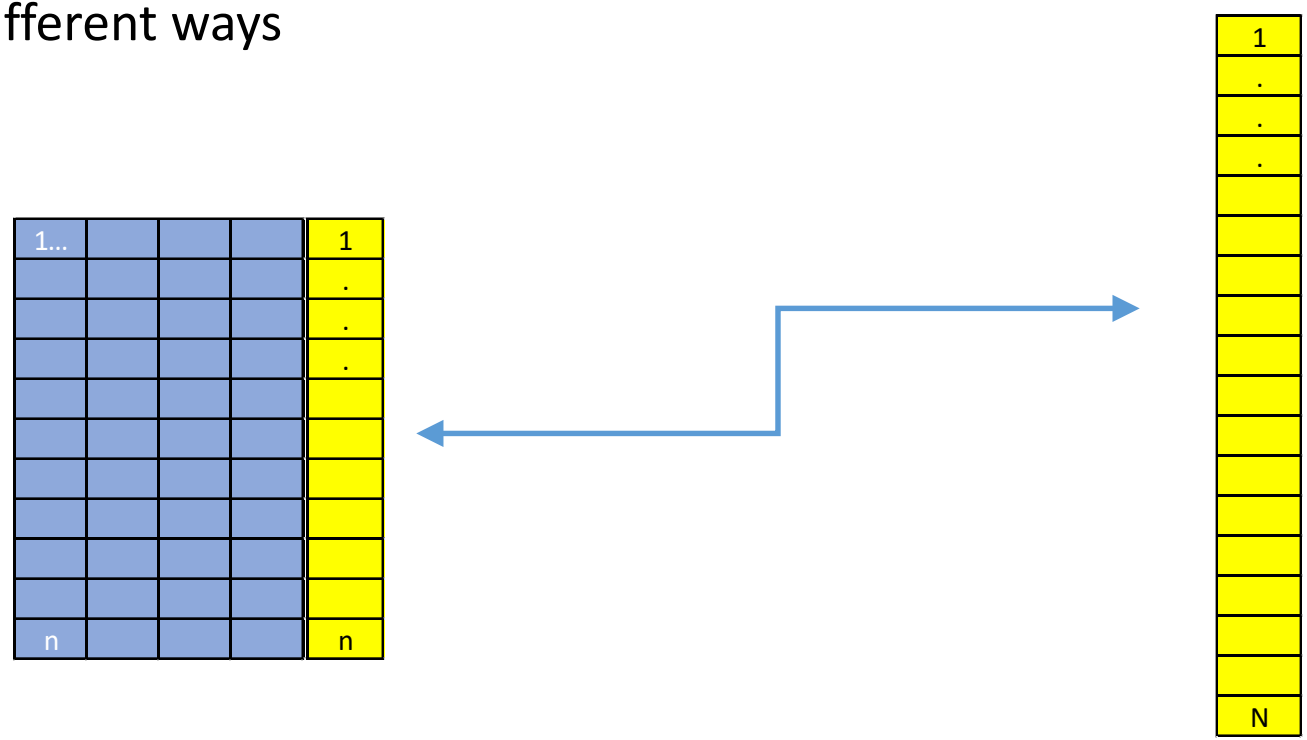
James wagner

Applications of record linkage to improve estimates – either from survey data or administrative data

- **Sample frame** development
  - Remove duplicate records from an existing sample frame
  - Augment an existing sample frame with records from a new data source
  - Conduct multiple frame surveys
- Estimates of **undercoverage** in a census count
- **Augmentation** of information contained in a survey data file or in administrative records
  - Validate survey responses / admin data
  - Impute missing survey values
  - Replace survey questions with information from administrative records
  - Add variables not collected on a survey or, alternatively, in the administrative records

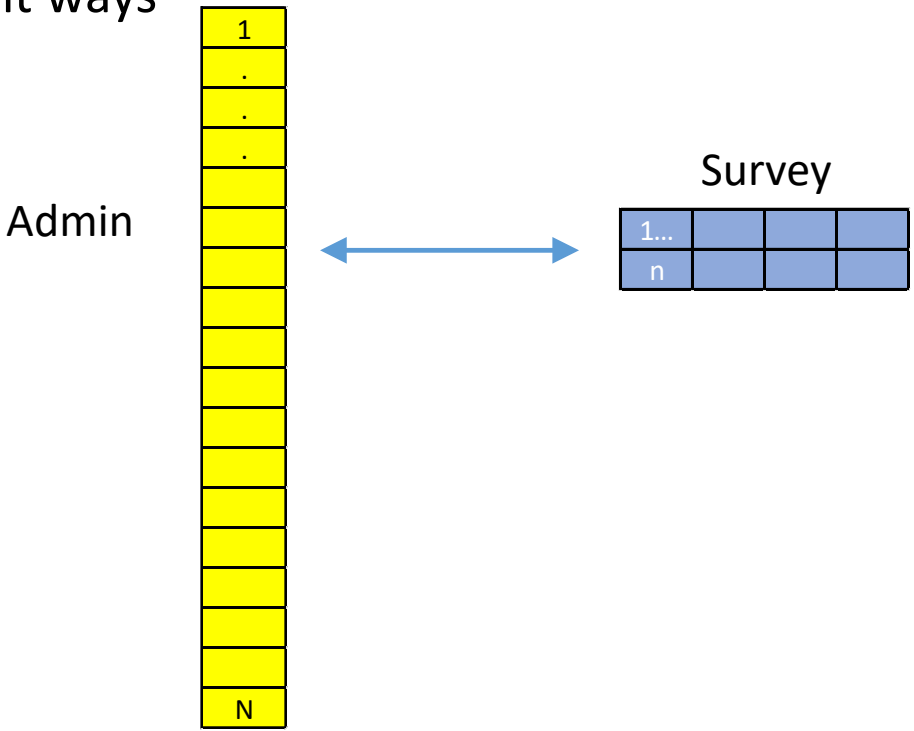
Combined Survey and Administrative Data--Design

- Designs may combine survey data and administrative records in different ways



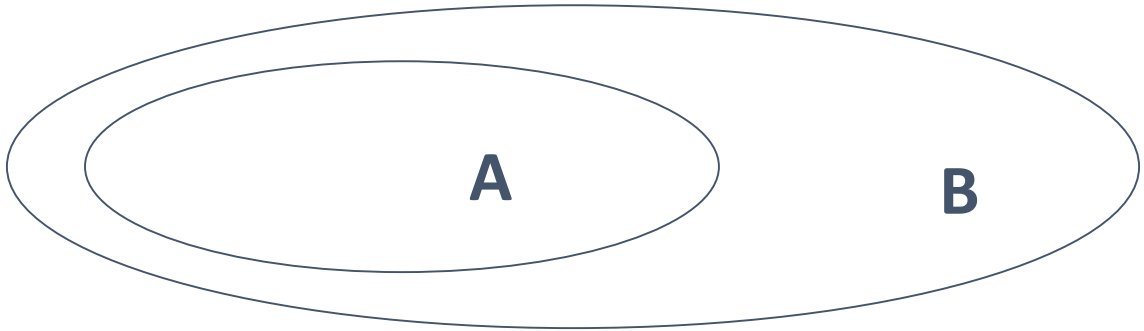
Combined Survey and Administrative Data--Design

- Designs may combine survey data and administrative records in different ways



Linking example: Building a sample frame with data from multiple sources

- National Agricultural Statistics Service (NASS) frequently conducts multiple frame surveys
  - NASS list of farm operations–call this frame A
  - Sample of land segments chosen to form an area frame, with farming operations in those segments listed on the ground–call this frame B



Linking example: Building a sample frame with data from multiple sources

- All units in frame A are included in frame B
- On its own, frame B would produce noisy estimates, but covers farm operations missing from frame A
- Estimator for dual frame survey typically of form:

$$\hat{Y}_{(p)} = pY_{ab}^{A} + (1-p)Y_{ab}^{B} + \hat{Y}_{b}$$

- Record linkage necessary to determine which units are in both frame A and frame B

Linking Example: Population Size Estimates with Capture-Recapture

- Idea originally from the animal abundance literature
  - Capture certain number of animals
  - Tag and release them
  - Re-capture and count proportion of tagged animals
- Similar concept applied to estimating size of human populations

Linking Example: Capture-Recapture Estimates of Syrian Civil War Deaths

- Suppose you have two lists of deaths in the Syrian Civil War
  - Neither is complete, but the two lists are independently compiled
- Record linkage allows analysts to determine which deaths are recorded only on list A, which deaths are only on list B, and which deaths are on both list A and list B
- Assuming that the two lists are independent, this information is sufficient to estimate the total number of deaths
  - If sample sizes are small, the estimate may be noisy, but it should be unbiased

Linking Example: Capture-Recapture Estimates of Syrian Civil War Deaths

- Consider the following matrix:
  - Deaths recorded on list A are $x_{11} + x_{12}$
  - Deaths recorded on list B are $x_{11} + x_{21}$
  - Overlap on the two lists is x_{11}
- Under independence:

		Sample B	
		Present	Absent
Sample A	Present	X_{11}	X_{12}
	Absent	X_{21}	X_{22}

$$\frac{X_{11}}{X_{11}+X_{12}} = \frac{X_{21}}{X_{21}+X_{22}}$$

Linking Example: Capture-Recapture Estimates of Syrian Civil War Deaths

- X_{22} is not observed, but under independence:

$$X_{22} = x_{12}x_{21}/x_{11}$$

- Estimated total number of deaths is: $x = x_{11} + X_{12} + x_{21} + x_{22}$
- Similar methods used to estimate census undercounts (and potentially to adjust census data)
  - Approach rests on records from census data collection linked to records from a post-enumeration study

Linking Example: Capture-Recapture Estimates of Syrian Civil War Deaths

Numerical example #1

		Sample B	
		Present	Absent
Sample A	Present	300	200
	Absent	300	X_{22}

What is the estimated total number of deaths?

Numerical example #2

		Sample B	
		Present	Absent
Sample A	Present	50	450
	Absent	550	X_{22}

What is the estimated total number of deaths?

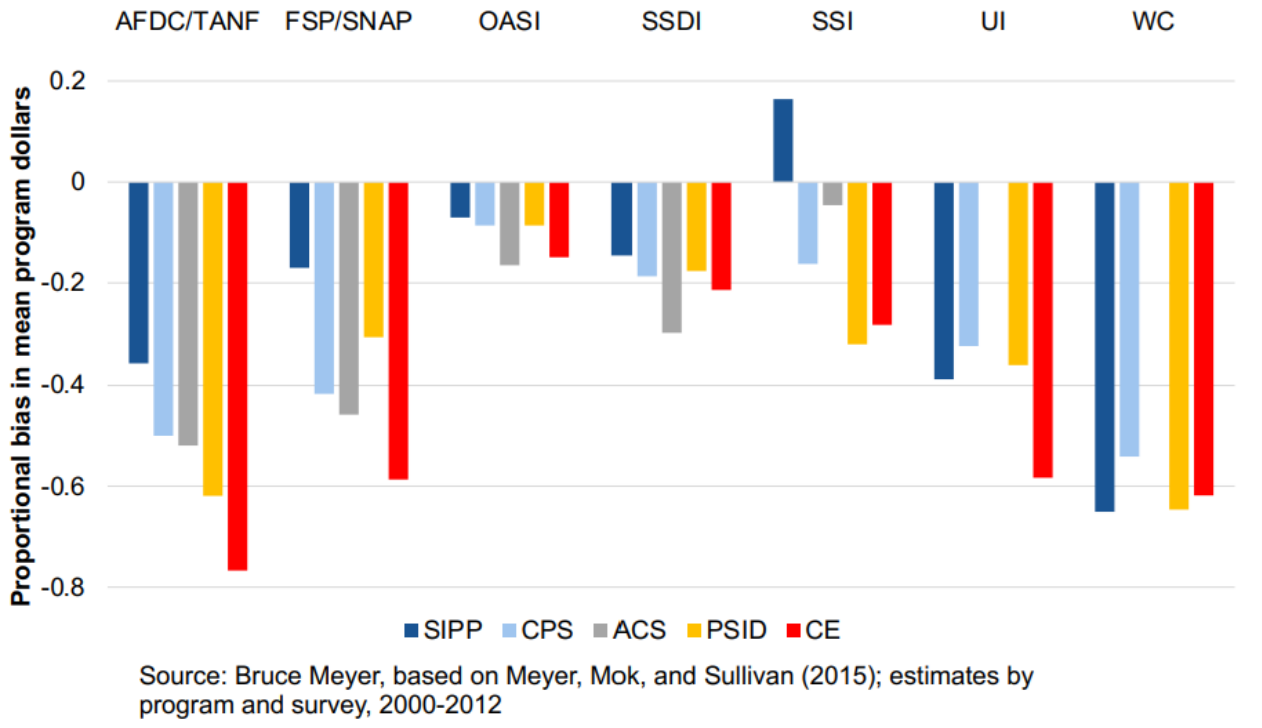
Linking Example: Capture-Recapture Extensions

- Using more than two lists
  - Similar idea, except using idea of “capture history” to estimate size
  - Example: An individual in the first and third lists but not the second would have a capture history of (1,0,1). Goal is to estimate (0,0,0).
  - Requires linking all lists
- Capture-Recapture with Respondent-Driven Sampling
  - Match RDS data to other administrative/survey records

Linking example: Improving Survey Data on Income

- Work by Bruce Meyer and collaborators (e.g., Meyer, Mok, and Sullivan 2015) finds significant discrepancies in income captured between survey and administrative data
  - Income from social insurance programs significantly understated in survey data
  - Primary source of under-reporting found to be inaccurate respondent reports, rather than coverage error or nonresponse error
- Linking household survey records to administrative records could provide better measures of income from programs such as TANF, SNAP, Workers Compensation, Unemployment Insurance, and others
  - Goal of Comprehensive Income Dataset project underway at Census Bureau

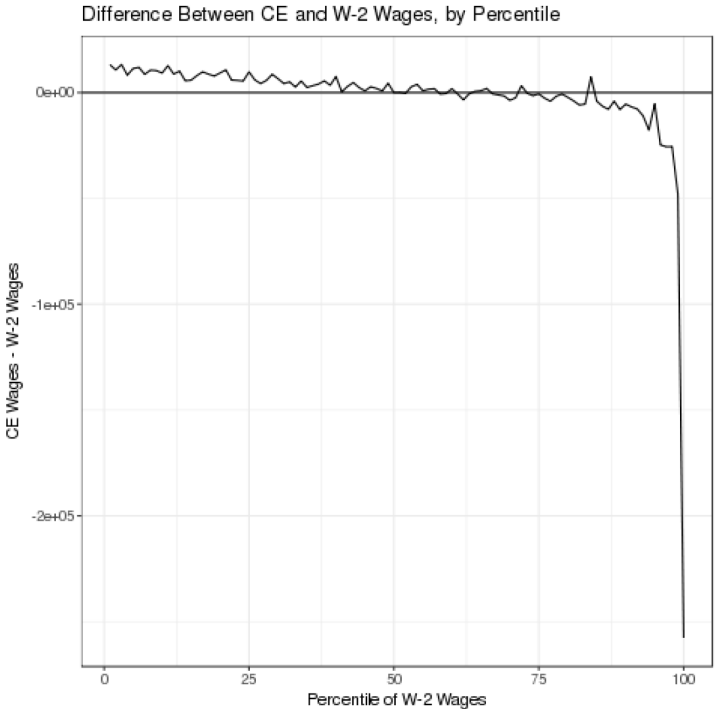
Survey Data Understates Income from Government Programs



Source: Bruce Meyer, based on Meyer, Mok, and Sullivan (2015); estimates by program and survey, 2000-2012

Self-reported income vs administrative records

- Comparison of CE Survey reports of wages to W-2
- Survey reports include other sources of income

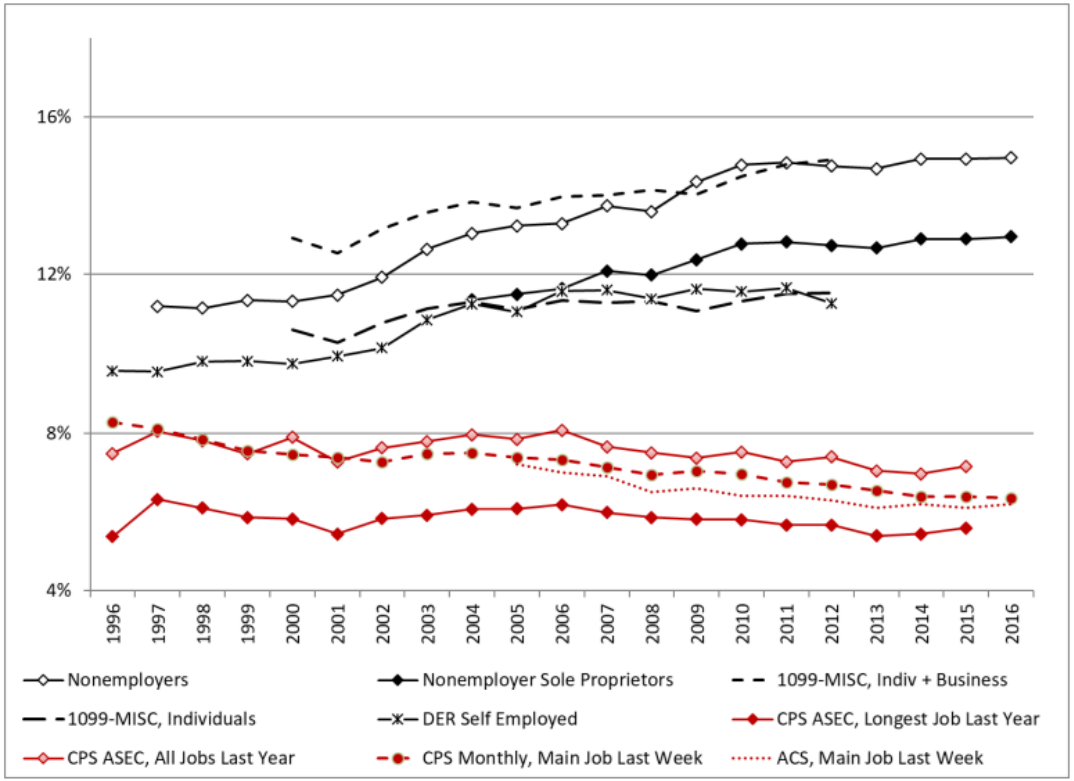


Brummet, et al., 2018

Linking Example: Improving Survey Data on Employment Status

- Abraham, Haltiwanger, Sandusky, and Speltzer (2017) find significant discrepancies between self-employment as measured in survey data and tax records
  - Share of people with self-employment income during the year higher in tax data than in household survey data
  - Between 1997 and 2012, incidence of self-employment was flat in household survey data, but rose notably in tax data
  - Significant disagreement in linked records between survey versus tax data self-employment status
- Linking tax information to survey records could improve published statistics

Self-Employment Levels and Trends



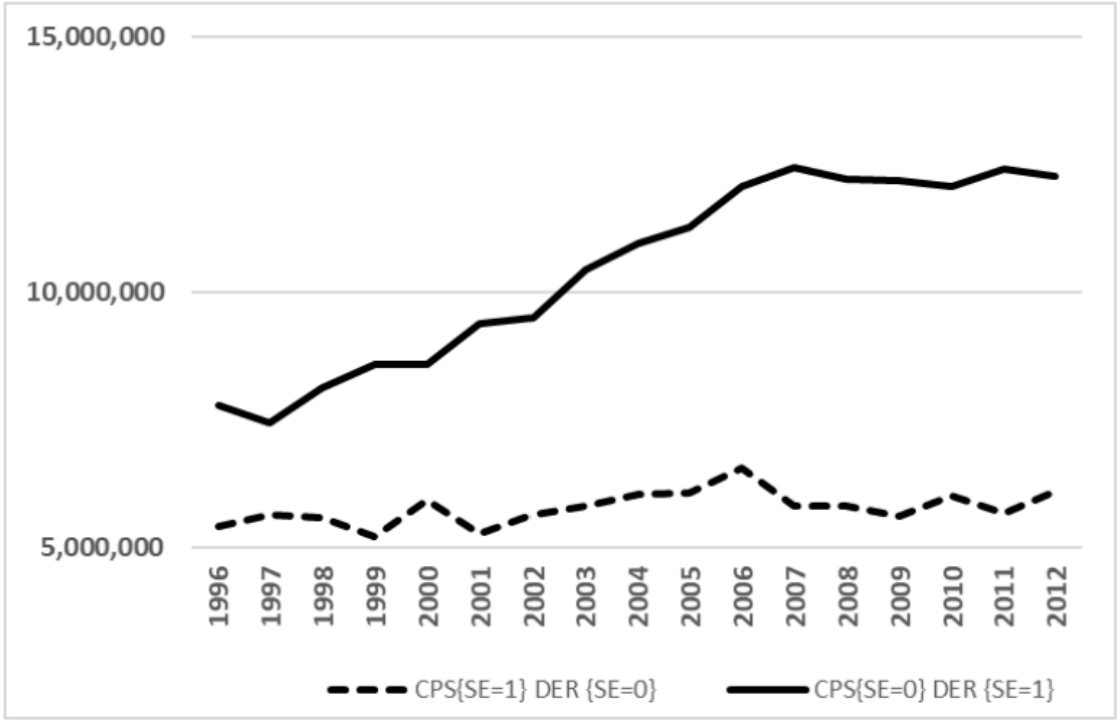
Source: Abraham, Haltiwanger, Sandusky and Spletzer (2018)

CPS-ASEC vs. DER Self-Employment Status Crosswalk (1996-2012)

		Not self-employed in DER	Self-employed in DER	Total
Not self-employed in CPS				
	Number	202,311,037	10,459,170	212,770,208
	Row Share	95.1%	4.9%	100.0%
	Column Share	97.2%	65.4%	95.0%
Self-employed in CPS				
	Number	5,776,887	5,531,764	11,308,651
	Row Share	51.1%	48.9%	100.0%
	Column Share	2.8%	34.6%	5.0%
Total				
	Number	208,087,924	15,990,935	224,078,859
	Row Share	92.9%	7.1%	100.0%
	Column share	100.0%	100.0%	100.0%

Source: Abraham, Haltiwanger, Sandusky and Spletzer (2018)

CPS vs. DER Off-Diagonals (1996-2012)



Source: Abraham, Haltiwanger, Sandusky and Spletzer (2018)

Linking Examples: Augmenting Survey Records with Additional Information

- Panel Study of Income Dynamics and Health and Retirement Study at the University of Michigan have augmented survey records with administrative data records
  - Administrative information has included residence in assisted housing, school identifiers and associated school information, Medicare claims, and Social Security records
  - Addition of linked information to data files has opened up whole new areas of research
- HUD is studying the possibility of using tax assessment records to replace and augment questions on the American Housing Survey (Bucholz, 2015)
  - Linking success rates vary by type of housing unit–from 70-80% for single family homes to 30% for manufactured homes and 13% for condominiums in multi-unit buildings
  - Availability and definition of variables on tax assessment files varies across jurisdictions
    - For example, what counts as a floor? How is square footage measured?

Example: Data from NCSES and IRIS

- Survey of Earned Doctorates (SED): Survey at the end of doctorate
  - Demographics, sources of support, graduate school experience, post-graduate plans
- Survey of Doctorate Recipients (SDR): Survey of science, engineering, health graduates several years after graduation
  - Sample of science, engineering, health graduates
  - Employment/career information
- UMETRICS: Transaction-level grant data for IRIS member universities
  - Information about federal funding, team size, etc.

Datasets

- SED/SDR contains information about graduates and the jobs they get after receiving their PhDs.
  - Research question: Do graduates work in the same field as their degree? Do graduates work in the same location as their PhD institution?
- UMETRICS contains information about funding.
  - Research questions: How does research happen in terms of team size and connections? What are the funding patterns in different universities?

Doctorate Recipients and Funding

- Research questions:
  - How does the type of funding affect career outcomes?
  - How does federal funding affect career pathways of doctorate recipients?
- We need record linkage to link SED/SDR data to UMETRICS data!

Linking Benefits with SED/SDR and UMETRICS

- We are able to answer more (and arguably more interesting) research questions
- Fill in missing values
  - UMETRICS imputes gender
- We can check/validate survey responses
  - SED includes questions on whether the student received funding, but students many times don't know the source of funding
- We can potentially improve surveys
  - If linkage is possible, we might be able to remove questions that are redundant and reduce respondent burden

Getting Started with Record Linkage: Pre-processing

Pre-Processing

- The very first step to doing record linkage is pre-processing the data
- The part might seem boring…but it's where most of the work is
- "In situations of reasonably high-quality data, preprocessing can yield a greater improvement in matching efficiency than string comparators and 'optimized' parameters. In some situations, 90% of the improvement in matching efficiency may be due to preprocessing" (Winkler, 2009)

Developing Pre-Processing Intuition: Which are matches?

Name	Age	Address
Bob Smith	24	1241 Main St.
Javier Martinez	47	232 Reading Dr.
Gillian Brown	31	1834 West 12th St

Name	Age	Address
Robert Smith	24	1231 Main St.
Haveir Marteenez	26	232 Redding Drive
Jilliam Jones	32	471 Orchard Lane

Pre-Processing Records for Matching

- Possible issue: We might reject a match when there's a very small difference or there may be different standards
- Put free-form names and addresses into a common format
  - Example: Separate fields for title, given name, middle name, and surname
  - Example: Create separate fields for street number, street name, and street descriptor (e.g., Street, Place, or Road)
  - Example: Separate dates into day, month, year

Standardize Common Terms

- Possible issue: Different ways of denoting the same common term might prevent exact matching
- Titles such as Doctor/Dr., etc.
- Name suffixes such as Junior/Jr., the Third/III
- Street prefix/suffix, such as Street/St., West/W, etc.
- Months (January/Jan) or Days of Week (Monday/Mon)
- Field of study (Chemistry/Chem)

Pre-Processing

Name	Age	Address
Bob Smith	24	1241 Main St.
Javier Martinez	47	232 Reading Dr.
Gillian Brown	31	1834 West 12th St

Address Number	Address Prefix	Address Name	Address Suffix
1241		main	street
232		reading	drive
1834	w	12th	street

Standardizing Names

Names are a common source of problems:

- Methods have been developed to standardize names before comparing
- Example: A nickname dictionary is sometimes used to help with standardizing names (Bob, Rob, Robert)
- Example: Soundex groups letters other than a, e, i, o, u, y, w, and h into categories; creates a code containing the first letter of the name and a numerical code based on the next three hard consonant (but with only one entry for consonants with the same code or that are separated only by an h or w); adds 0s as needed for a code with three numbers
    - Names Smith, Smythe, and Snide all would be coded as S530, since i, h, y, and e ard dropped; m and n are in the same coding group; and t and d are in the same coding group

Blocking

- With data files of even moderate size, not feasible to compare all possible pairs
  - Example: if matching a survey file with 100,000 records to an administrative file with 100 million records, 10 trillion possible pairs!
- Solution is to *block* the records–divide them into groups based on one or more variables and make comparisons only within groups
  - Blocking variables could be based on geography (e.g., state or zip code), name (e.g., first three letters of last name), a combination (e.g., zip code and first letter of last name), or something else
  - Implicit assumption in simple methods that blocking variables are not affected by data entry errors, through there are more sophisticated methods that allow for the possibility of such errors may occur

Blocking

- Blocking will increase share of matches among pairs that are compared and reduce number of false matches, at the potential cost of increasing the number of true matches that are missed
- Example: Linkage to be performed on two files, a smaller file containing 2,000 records drawn from 100 geographic areas (20 per area) and a larger file containing records for all households known to live in those areas (100,000 records or 1,000 per area)
  - How many pairs need to be compared without blocking?
  - How many pairs need to be compared with blocking on area?
  - What is a scenario in which blocking might reduce the number of false matches?
  - What is a scenario in which blocking might prevent true matches from being identified?

Pre-processing Notes

- Pre-processing is very context dependent!
- It’s an iterative process
- Don’t neglect it–if you try to do record linkage without pre-processing, you will NOT get good results
- The fancy algorithms will do what they’re supposed to do–it’s up to you to give it good pre-processed data

Record Linkage Methods

Types of Record Linkage

- Deterministic linkage
  - Compare single identifier or group of identifiers contained on two records
  - Link occurs if identifiers match exactly
- Probabilistic linkage
  - Most common approach uses multiple identifiers that may contain errors, calculates the relative likelihood that two records represent versus do not represent the same unit
  - Link occurs if relative likelihood that records are a true match exceeds a defined threshold
- Statistical matching (data fusion)
  - Using information from two non-overlapping data sets–data set A containing information on X and Y and data sets B containing information on X and Z (but perhaps not Y)--to impute values of Z to the records on data set A

Deterministic Linkage

- In a deterministic linkage, two records are said to be a match if and only if they agree *exactly* on a set of identifiers (the *match key*)
  - Potential match keys might be the SSN or a combination of first name, last name, age or date of birth and address
- Any disagreement in the match key will lead to records not being matched. Causes of error might include:
  - Key error in entering SSN
  - Misspelling of name or address
  - Difference in timing of when information collected (could affect last name, age, and address)
- Deterministic linkage will work well *only* when working with edited files such that the analyst can be confident that match key values are accurate and comparable

Probabilistic Linkage

- Consider two data sets A and B containing records for members of a common population
  - AxB the set of all potential pairs of records from A and B
  - AxB can be partitioned into the set of pairs that are matches (*M*) and the set of paris that are not matches (*U*)
  - Whether a pair belongs to M or U is not observed
- In most common approach, linkage decisions based on estimates of the probability that a pair is a match relative to the probability that the pair is not a match
  - Linkage occurs when the relative probability that a pair is a match exceeds a defined threshold
- Basic theory underlying this approach developed in classic paper by Felligi and Sunter (1969)

Statistical Matching

- Statistical matching (data fusion) occurs when variable Z is added to data set A containing X and Y based on information contained in some disjoint data set B
  - At a minimum, data set B must contain information on X and Z
  - Typical assumption is that Y and Z are conditionally independent
- Common methods for determining value of Z to be imputed into records in A based on information contained in B
  - Nearest neighbor hot deck (choose value from donor record in B that is most similar to recipient record in A, based on values of X)
  - Fit regression model $Z=X\beta+\varepsilon$, use estimated coefficients to impute values for records in A
- Obvious limitation of standard approach: Assumption of conditional independence often not tenable
  - Example: Let X be education, Y be occupation, and Z be annual earnings. Unlikely that Y and Z are independent, conditional on X.

Statistical Matching

- May be able to do better if have access to a data set B that contains information on X, Y, and Z
  - Distance function used in hot deck procedure to identify donor units in data set B can be modified to incorporate X and Y
  - Regression model can be modified to included both X and Y as regressors: $Z=X\beta+Y\alpha+\varepsilon$
- Some additional quality considerations in data fusion applications
  - Are X's defined the same way on both data files?
  - Are the data files contemporaneous?
  - How can uncertainty in the imputation be propagated in the fused data file?
  - Are the results sensitive to the assumptions about relationships among the variables?

Probabilistic Matching

Fellegi-Sunter Intuition

If two records match on month of birth, how much evidence does that provide for them being matches? What if they matched on day of birth? Year of birth?

Day of Birth	Month of Birth	Year of Birth
21	2	1992
7	4	1973
4	3	1982

Fellegi-Sunter Linkage Model

- Consider all possible pairs *r* in a set of records that contain multiple potential linking variables
  - For example, variables could be (1) first name, (2) last name, (3) day of birth, (4) month of birth, (5) year of birth, (6) street address number, (7) street address name, and (8) zip code
- Define a vector γ_r that contains a 1 in the position of each linking variable that agrees between the two records in the pair, else a 0. For example: γ_r = {1,1,1,1,0,1,0,1}
- Analyst determines the probability of observing γ_r given $r \in M$, $Pr(\gamma_r | r \in M)$, and the probability of observing γ_r given $r \in U$, $Pr(\gamma_r | r \in U)$
- Record linkage decision based on

$$R = \frac{Pr[\gamma_r | r \in M]}{Pr[\gamma_r | r \in U]}$$

Fellegi-Sunter Linkage Model

- Demine m_i as the probability of observing a match for element *i* in γ_r when $r \in M$ and u_i as the probability of observing a matching for element *i* in γ_r when $r \in U$.
- Analysis can be greatly simplified if willing to assume that these probabilities are independent across the different variables.
  - Assumption not likely to be realistic. People with the same zip code, for example, will be more likely than people selected at random to live on a street with the same name
  - Algorithm may perform well even if assumption of independence not satisfied
- Under assumption of independence, likelihood of γ_r for any given record pair $r \in M$ can be written as:

$$Pr[\gamma_r | r \in M] = \prod_{i=1}^{K} m_i^{\gamma_i} (1 - m_i)^{1-\gamma_i}$$

Fellegi-Sunter Linkage Model

- Similarly, the likelihood of γ_r for any given record pair $r \in U$ can be written as:

$$Pr[\gamma_r \mid r \in U] = \prod_{i=1}^{K} u_i^{\gamma_i} (1-u_i)^{1-\gamma_i}$$

- This means that the likelihood ratio R can be written as:

$$R = \frac{Pr[\gamma_r \mid r \in M]}{Pr[\gamma_r \mid r \in U]} = \frac{\prod_{i=1}^{K} m_i^{\gamma_i} (1-m_i)^{1-\gamma_i}}{\prod_{i=1}^{K} u_i^{\gamma_i} (1-u_i)^{1-\gamma_i}} = \prod_{i=1}^{K} \left[\frac{m_i^{\gamma_i} (1-m_i)^{1-\gamma_i}}{u_i^{\gamma_i} (1-u_i)^{1-\gamma_i}} \right]$$

Fellegi-Sunter Linkage Model

- Taking natural logarithm of both sides, obtain:

$$\ln R = \sum_{i=1}^{K} \ln \left[\frac{m_i^{\gamma_i} (1 - m_i)^{1-\gamma_i}}{u_i^{\gamma_i} (1 - u_i)^{1-\gamma_i}} \right]$$

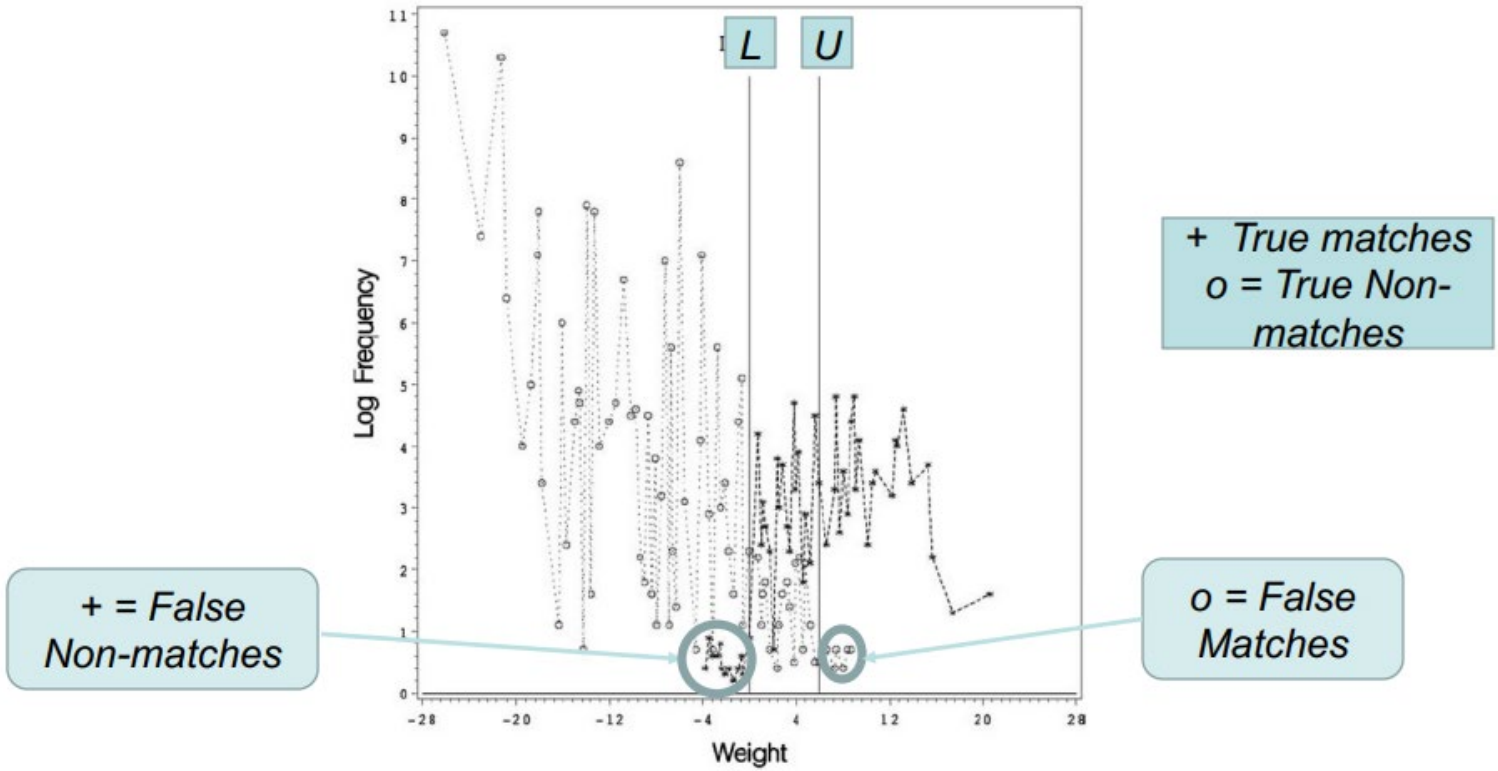
  - Term in this sum are the agreement weights for the individual variables
  - ln(R) is match score for the record pair
- Analyst sets thresholds for ln R such that
  - Ln R > Upper → link (L)
  - Lower < ln R ≤ Upper → possible link (P)
  - Ln R ≤ Lower → non-link (N)

Estimating Match Probabilities in the Fellegi-Sunter Model

- One method for estimating m_i and u_i: Use external information of error rates to estimate m_i and external information on relative frequencies of u_i.
- Example: suppose that population is half male and half female, but that sex is coded incorrectly on 10 percent of records
  - How can external information be used to calculate m_i?
    - Sex is correct on both A and B in 81% of cases (0.9*0.9=0.81)
    - Sex is incorrect on both A and B in 1% of cases (0.1*0.1=0.01)
    - This means that $m_i = 0.82$
  - How can external information be used to calculate u_i?
    - If pairs of elements for non-matching cases selected randomly from A and B, 25% of pairs will be M-M and 25% of cases will be F-F
    - This means that $u_i = 0.50$

Estimating Match Probabilities in the Fellegi-Sunter Model

- Another method is to use the information contained in A and B, together with a set of assumptions, to estimate the maximum likelihood values of m_i and u_i
  - Herzog, Scheuren, and Winkler (2007) discuss the Expectation-Maximization (EM) algorithm that is commonly used in record linkage applications
  - Note that it may note be critical for m_i and u_i to be estimated precisely if model nonetheless assigns high match scores to pairs that are true matches



Source: Herzog, Scheuren, and Winkler (2007)

String Comparators for Text Variables

- Many fields commonly contain minor typographical errors
  - Example: Katharine or Katherine
- Researchers have developed methods for comparing the “distance” between one string and another. Examples:
  - Levenshtein distance: minimum number of single-character edits required to change one string into the other; can be reformulated to lie between zero (no similarity) and one (exact match)
  - Jaro distance: based on number of characters in two strings that match and number of transpositions required to put those characters in the same order; values between zero (no similarity) and one (exact match)
  - Jaro-Winkler distance: modified Jaro distance to give greater weight to matching on initial characters; also lies between zero and one
- When used in record linkage context, typical approach to accept variable comparisons with a score above some threshold
  - Note that Felligi-Sunter likelihood ratios need to be adjusted to account for accepting variables with partial agreement

Example: Levenshtein Distance (Edit Distance)

- Number of edit operations (insertion, deletion, or substitution of a single character) needed to convert one string into another
  - Edit distance 0 if two strings identical
- Edit distance between “sitting” and “kittens”
  - Starting with “sitting”, replace s with k; replace i with e; remove g; add s (4 operations)
  - Starting with “kitten”, replace k with s; replace e with i; remove s; add g (4 operations)
- Range can be normalized to lie between 0 (no similarity) and 1 (exact match):

$$1 - \left[\text{edit distance} \Big/ \text{maximum edit distance given string lengths} \right]$$

Example: Jaro Distance

- Formula for Jaro distance is

$$sim_j = 0 \text{ if c=0, else} = W_1 \frac{c}{L_1} + W_2 \frac{c}{L_2} + W_3 \frac{(c-\tau)}{c},$$

  - $W_1 + W_2 + W_3 = 1$ (commonly each ⅓)
  - L_1, L_2 length of two strings
  - c number of characters that match
  - τ number of characters in common transposed
- What is the Jaro distance between Martha and Marhta (using W=⅓)?

Example: Jaro-Winkler Distance

- Jaro-Winkler modifies Jaro formula to give more weight to initial characters in a string:
  - $sim_W = sim_J + lp(1-sim_J)$
  - l is min(k, 4), where k is the length of the common prefix in the two strings (number of characters starting with the first that are identical)
  - $0 < p < 0.25$ (0.1 is a typical value)
- What is the Jaro-Winkler distance between Martha and Marhta?

Quality Metrics for Record Linkage

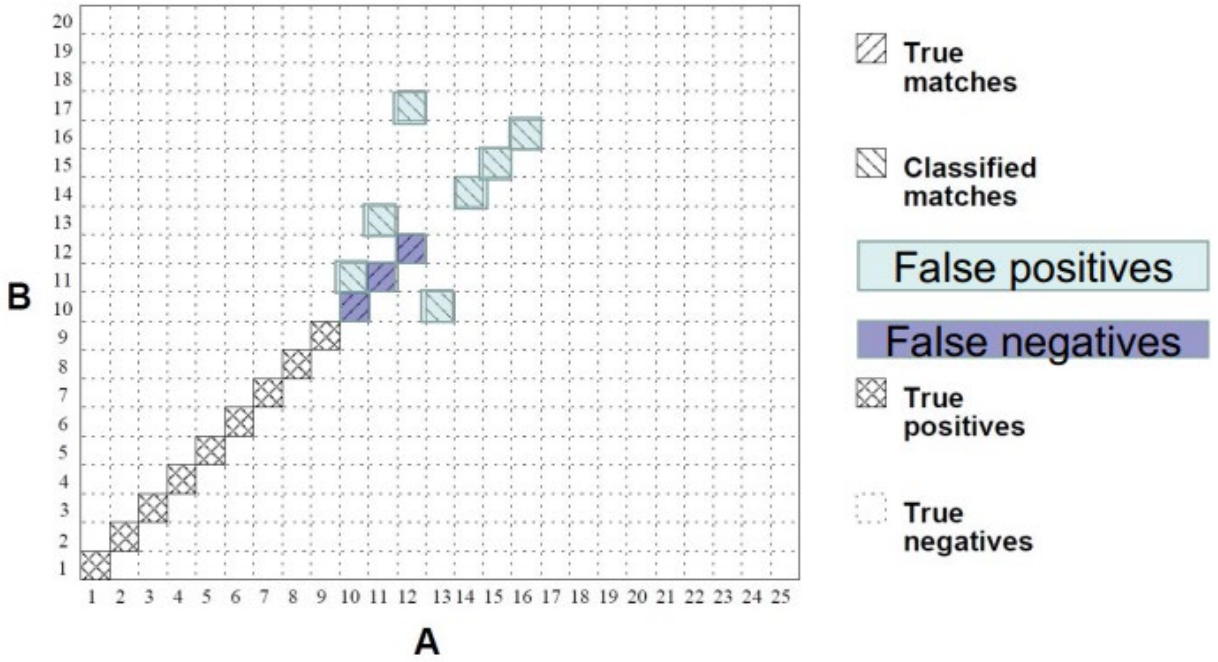
Table 1. Confusion matrix of record pair classification		
Actual	Classification	
	Match ($\widetilde{M}$)	Non-match ($\widetilde{U}$)
Match (M)	True match	False non-match
	True positive (TP)	False negative (FN)
Non-match (U)	False match	True non-match
	False positive (FP)	True negative (TN)
Note: Defined on the comparison space (all pairs).		

Adapted from: Christen and Goiser (2007)

Quality Metrics for Record Linkage

Measure	Definition	Comment
Alpha error	FN/(TP+FN)	Share of true matches that are incorrectly classified
Beta error	FP/(TN+FP)	Share of true non-matches that are incorrectly classified
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Dominated by TN
Precision	TP/(TP+FP)	True matches as a share of classified matches
Sensitivity (TPR)	TP/(TP+FN)	Share of true matches that are classified correctly; also referred to as recall or true positive rate
Specificity (TNR)	TN/(TN+FP)	Also referred to as true negative rate; driven by TN
Precision-Sensitivity Breakeven	Precision = Sensitivity	
F-measure	2(Precision*Sensitivity)/ (Precision+Sensitivity)	Harmonic mean of precision and sensitivity measures

Quality Metrics for Record Linkage



Quality Metrics for Record Linkage

- In this example, what are the TP, TN, FP, and FN?
- What is the accuracy rate: (TP+TN)/(TP+TN+FP+FN)?
- What is the precision: TP/(TP+FP) ?
- What is the sensitivity: TP/(TP+FN)?

Quality Metrics for Record Linkage

- Also can think about metrics for choice of blocking variables
  - Ideally would like blocking to reduce the number of pairs as much as possible, while also retaining as many true matches as possible
- Possible to construct metrics to describe this

$$\text{Reduction rate} = 1 - \frac{N_b}{|A||B|}$$

$$\text{Pairs completeness} = \frac{N_m}{|M|}$$