Probabilistic Methodology for Genealogical Record Linkage: Determining Weight Robustness

Krista Jensen
John S Lawson

Brigham Young University
Statistics Department

Record Linkage

- What is record linkage?
  - Process that joins two records of information for a particular individual or family
- Applications of Record Linkage
  - Genealogical research
    - Census Records
    - Ecclesiastical Records
  - Medical research
  - Data storage
  - Government

Census Data

- Benefits of census data
  - Information
  - Completeness
  - Starting point for genealogical research
- Collection methods
  - Training
  - Instruction given to enumerators
- Concerns with census data
  - Correctness of data
  - Age
  - Place of origin

Census Indexes

- What is a census index
  - Head of Household
  - Individuals with different last names
  - Subset of questions
  - Availability of census records. Census record access limited from 1930 to present for privacy
- Fields available in census record indexes
  - Surname, given name, age, gender, race, place of origin, state, county, census page information

Probabilistic Methodology

Overview of Theory

- 3 decisions possible (e_i) *where i=1,2,3*
  - Definitions of Events e_i where $i=1,2,3$
    - e_1 two fields are a match (positive link)
    - e_2 two fields are a of undetermined status
    - e_3 two fields are a non-match (positive non-link)

Probabilistic Methodology

- A weight is calculated for each field based on conditional and unconditional probabilities
  - Definitions of Probabilities
    - $P\ (e_i|M)$ can be calculated from a known set of matches
    - $P\ (e_i)$ can be estimated using sample pairs
    - $P\ (M)$ is constant for all comparisons

- A score for each comparison is calculated (sum of the weights)
- Threshold Values are used to determine the classification of each record comparison

Probabilistic Methodology

Calculating the Weights

$$w_k = \ln[P(M \mid e_i)]$$

Using Bayes Rule:

$$P(M \mid e_i) = \frac{P(e_i \mid M)P(M)}{P(e_i)}$$

Probabilistic Methodology

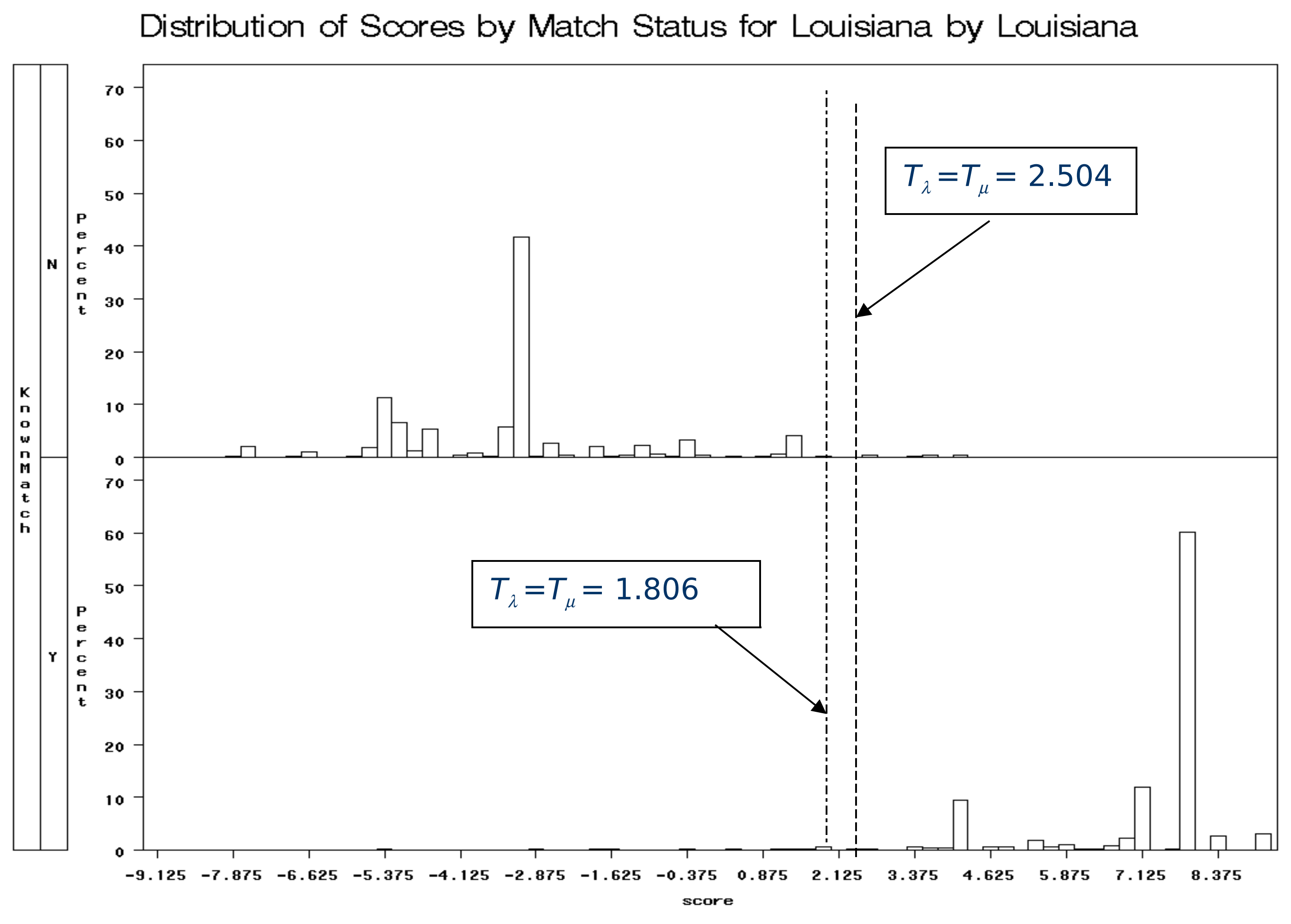
The Scores

$$W = \sum w_k = \sum \ln[P(M \mid e_i)]$$

$$= \sum \ln[P(M)] + \sum \ln\left[\frac{P(e_i \mid M)}{P(e_i)}\right]$$

A Weight is calculated for k fields, the score is the sum of those weights

Probabilistic Methodology



$T_{\lambda} = T_{\mu} = 2.504$

$T_{\lambda} = T_{\mu} = 1.806$

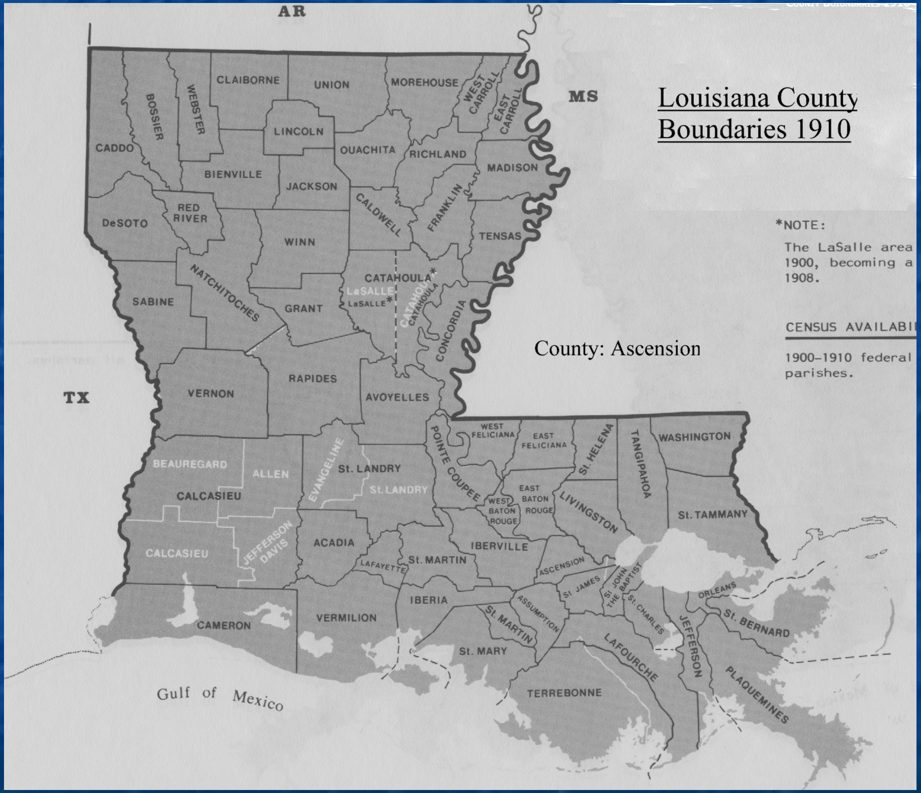
Project Data

- Census Record availability
- Geographical areas sampled
  - California
  - Connecticut
  - Illinois
  - Michigan
  - Louisiana

WA
OR
ID
MT
ND
MN
SD
WY
NV
UT
CA
CO
NE
KS
IA
WI
MI
IL
IN
OH
MO
KY
TN
AZ
NM
OK
AR
TX
LA
MS
AL
GA
SC
NC
VA
WV
PA
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
DC
FL
AK
HI

Project Data

- Sampled counties from 1910 and 1920.
  - County boundaries that changed were eliminated from selection
  - Records were extracted for the counties of interest



Louisiana County Boundaries 1910

County: Ascension

Project Data

STATE	Record Size	Matches
Connecticut	18,799	2,405
Illinois	32,211	4,984
Louisiana	18,233	596
Michigan	31,497	4,539
Southern California	32,684	2,779
Northern California	21,436	1,943

Algorithm Adaptations

- Place of Origin Index
  - Prussia in 1920 matches Germany in 1910
  - Hungary and Austria match for either year
- Enumeration Locality Index
- Considerations for Age
  - Range of 8-12 years classified as “same”
  - Range of 7 and 13 years classified as “close”
  - Range greater than 13 years and less then 7 years classified as “different”

Results

Averaged: Fields	Weight for "Same"	Weight for "Close"	Weight for "Different"
Given Name	4.18009	-1.2599	-4.76084
First 3 letters of Given Name		3.3928	
First letter of Given Name		0.357	
Last 3 letters of Given Name		-0.2251	
Age	2.45507	-0.1078	-2.63094
Race	0.18305	0.84357	-1.58802
Place or Origin	1.49957	-0.9575	-2.66818
Locale of Census	2.02468	1.52134	-1.35869
County	0.50254		-3.16472

Score Calculation

1920-8780	DRECHSKER	OTTO C	D622	48	M	W	SAXO	CT	TOLLAND	4-WD ROCKVILLE VERNON	T625	198	1	272	B
1910-2334	DRECHSLER	OTTO	D622	38	M	W	GERM	CT	TOLLAND	4-WD ROCKVILLE	T624	143	3	71	A

Given name - match, Age - match, Gender - match, Race - match, Origin - match, State match, County - match and District - match. Provides a score of

4.18 + 2.45 + 0.18 - 2.67 + 2.02 + 0.50 + 2.02 = 6.498

Results

Error Rates obtained using Averaged Weights				
Census Record Set	$T_{\mu} = 2.405$	$T_{\mu} = 1.806$	$T_{\lambda} = 2.405$	$T_{\lambda} = 1.806$
Connecticut	0.04075	0.01414	0.01701	0.02542
Illinois	0.01083	0.00522	0.0141	0.01233
Louisiana	0.02687	0.0198	0.01346	0.02359
Michigan	0.02600	0.01763	0.00902	0.01316
Southern California	0.04282	0.03203	0.01293	0.0169
Northern California	0.02162	0.01338	0.00842	0.01302

Results

- Problems encountered with blocking deal mainly with surname
- Misspellings cause problems with matching first name.
  - Highest weight: record pair not identified as a potential match because the negative weight for the classification of “different” is given to the score.

Results

- Recommendations for using the averaged weights:
  - Averaged weights obtained in this project can be used when linking indexed census records from 1910 and 1920
  - When linking census records between other decades new weights need to be calculated. (this will take into account the population fluctuations of the time period.)

Discussion

- Pros for Averaged Weights
  - Time saving
    - Do not need a set of known matches to calculate conditional and unconditional probabilities
  - Low error rates
  - Robustness
- Cons for Averaged Weights
  - Better results were obtained using other weights for some data sets

Future Research

- Linkage Problems
  - Not using a compression code
  - Misspellings in given name
- Solution: Use a secondary algorithm that counts the number of letters that match and take the corresponding percentage of the weight and apply that to the score
  - Briggs and Briggo apply 83% of the weight
  - Take off all ‘s’ at the end of a surname
  - Apply secondary algorithm to given name