Probabilistic Record Linkage in Genealogical Research

John Lawson, Dave White, Brenda Price and Ryan Yamagata

Agenda

• Introduction

• Description of Probabilistic Record Linkage

• Applications to Quaker Records in N.C.

• Future Directions
Introduction

More Complete Information about an Individual

• Census Records
• Birth Records
• Death Records
• Marriage Records
• Church Records
• Immigration Records
• Wills
• Deeds
Information Age

Credit Records

Medical Records

Stored Electronically, for Quick Recall and Search
Introduction

Genealogical Records

• No Identifier Field such as SSN
• Different Spellings or nicknames
• Misreported Dates or day, month, year interchanges
• Missing information
• Other Errors
Probabilistic Record Linkage

• Adapted by Church of Jesus Christ of Latter Day Saints Family History Department in TempleReady™

• We Will Describe the Approach and show its application to Genealogical Research
Probabilistic Record Linkage

History

- 1946 - Dunn Introduces Concept
- 1959 – Newcomb et. al. – linked vital records
- 1960’s – Development Theoretical Foundations
  Du Boise
  Nathan
  Tepping
  Fellegi and Sunter
- Recently Computer Software
  CAMLINK, CAMLIS, LinkPro
Probabilistic Record Linkage

Methodology

• Record Consists of Fields

• When Comparing Two Records each compared field receives a weight
   
   + if fields agree
   
   - if fields are different
   
   0 if field from one or both record is missing

• Decision on whether two fields should be linked is based on the sum of the weights “Score” over all fields compared

• Link, Do not Link, Undetermined
Probabilistic Record Linkage

Methodology

Calculating the Weights:

\[ w_i = \ln[P(M \mid e_i)] \]

Using Bayes Rule

\[ P(M \mid e_i) = \frac{P(e_i \mid M)P(M)}{P(e_i)} \]
Probabilistic Record Linkage

Methodology

• $P(e_i)$ can be estimated using sample pairs
• $P(e_i|M)$ can be calculated from a known set of matches
• $P(M)$ is constant for all comparisons
Probabilistic Record Linkage

The Weights

\[ w_i = \ln[P(M \mid e_i)] \]

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

\[ = \ln[P(M)] + \ln \left[ \frac{P(e_i \mid M)}{P(e_i)} \right] \]
Probabilistic Record Linkage

• The Scores

\[ W = \sum w_i = \sum \ln[P(M | e_i)] \]

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

• Blocking
Probabilistic Record Linkage

Histogram of Matches and Non-Matches

Score = Sum of Weights
Application to Genealogical Research

The Data:

• Church (Quaker Congregation) and County Records
• Perquimans and Pasquotank Counties, NC
• 1600 to 1900
• Births, Deaths, Marriages, and minutes of town meeting
• 9279 Individual records
Application to Genealogical Research

Records from Town Meeting Minutes:

Benjamin C. Winslow, s. William & Julian, b. 3-5-1837, Chowan Co.
Esther P. Winslow. (dt. Silas & Elizabeth Chappell, b. 2-10-1840, Chowan Co.)
Ch: Harriett Ann  b. 6-23-1862.
       William W.  “  11-8-1864.
       James Claudius  “  9-21-1873.
       Ora
       Henry

Laden.
1880, 8, 7. Sarah (form Winslow) rpd m. (not m in mtg).

Birth Record:
George Durant son of George & Ann Durant was borne the 24th December 1659
Application to Genealogical Research

• Records entered manually into PAF

• GEDCOM file created from PAF

  - RIN’s
  - MRIN’s

• Visual Basic Program: GEDCOM → Flat File

  - Flat File
  - 9279 records

• SAS (Statistical Analysis System)
Application to Genealogical Research

9279 Total Records = 43,045,281 pairwise comparisons

Blocking by Surname and Sex:
1875 Records with no Surname
7404 Records remaining = 220,931 pairwise comparisons
2118 matches
218,813 non-matches

Blocking by Surname only

treated no surname together in one block
9279 total records 1,961,004 pairwise comparisons
3692 matches
1,957,312 non-matches
<table>
<thead>
<tr>
<th>Field Number ($i$)</th>
<th>Variable</th>
<th>$w_i (S)$</th>
<th>$w_i (D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Given Name</td>
<td>3.47715</td>
<td>-2.81401</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
<td>0.69078</td>
<td>-8.1628</td>
</tr>
<tr>
<td>3</td>
<td>Father's Given Name</td>
<td>2.83686</td>
<td>-2.54161</td>
</tr>
<tr>
<td>4</td>
<td>Father's Surname</td>
<td>3.89474</td>
<td>-2.44506</td>
</tr>
<tr>
<td>5</td>
<td>Mother's Given Name</td>
<td>2.09498</td>
<td>-1.6466</td>
</tr>
<tr>
<td>6</td>
<td>Mother's Surname</td>
<td>3.04619</td>
<td>-8.1628</td>
</tr>
<tr>
<td>7</td>
<td>Spouse's Given Name</td>
<td>3.30857</td>
<td>-2.5861</td>
</tr>
<tr>
<td>8</td>
<td>Spouse's Surname</td>
<td>4.39975</td>
<td>-3.06505</td>
</tr>
<tr>
<td>9</td>
<td>Birth Town</td>
<td>0.00176</td>
<td>-8.1628</td>
</tr>
<tr>
<td>10</td>
<td>Birth County</td>
<td>0.55256</td>
<td>-1.57191</td>
</tr>
<tr>
<td>11</td>
<td>Birth State</td>
<td>0.00604</td>
<td>-8.1628</td>
</tr>
<tr>
<td>12</td>
<td>Birthday</td>
<td>3.43841</td>
<td>-2.16826</td>
</tr>
<tr>
<td>13</td>
<td>Birth Month</td>
<td>1.98113</td>
<td>-0.91975</td>
</tr>
<tr>
<td>14</td>
<td>Birth Year</td>
<td>4.60908</td>
<td>-1.09195</td>
</tr>
<tr>
<td>15</td>
<td>Death Town</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>Death County</td>
<td>0.59431</td>
<td>-8.1628</td>
</tr>
<tr>
<td>17</td>
<td>Death State</td>
<td>0</td>
<td>-8.1628</td>
</tr>
<tr>
<td>18</td>
<td>Death Day</td>
<td>3.47962</td>
<td>-1.70889</td>
</tr>
<tr>
<td>19</td>
<td>Death Month</td>
<td>2.28891</td>
<td>-2.04636</td>
</tr>
<tr>
<td>20</td>
<td>Death Year</td>
<td>4.41364</td>
<td>-2.12932</td>
</tr>
</tbody>
</table>
Application to Genealogical Research

Matches: 1.65% misclassified, 17.52% unclassified

Non-Matches: 1.87% misclassified, 7.71% unclassified

Figure 3: Relative Frequency Histogram with Thresholds when Blocked by Surname and Sex
Application to Genealogical Research

Matches: 4.96% misclassified
Non-Matches: 2.39% misclassified

Figure 4: Relative Frequency Histogram with Thresholds when Blocked by Surname Only
The Future For Our Research

• Extend Visual Basic Program

• Expand Weighting Possibilities

• Obtain More Data

• Build Library of Weights