Historical life cycle reconstruction by indexing

Gerrit Bloothooft
Jelte van Boheemen
Marijn Schraagen
Utrecht institute of Linguistics, Utrecht University, The Netherlands
g.bloothooft@uu.nl; jeltevanboheemen@gmail.com; m.p.schraagen@uu.nl

The Netherlands, province of Zeeland
1811 – 20th century vital registration
The Zeeland challenge

application domain

1811
province of Zeeland
1811 – 20th century vital registration

19/09/2016
Zeeland

population size (census)

<table>
<thead>
<tr>
<th>year</th>
<th>people</th>
</tr>
</thead>
<tbody>
<tr>
<td>1830</td>
<td>137,200</td>
</tr>
<tr>
<td>1869</td>
<td>177,569</td>
</tr>
<tr>
<td>1930</td>
<td>247,360</td>
</tr>
</tbody>
</table>

in-migration (census 1869)

<table>
<thead>
<tr>
<th>born in</th>
<th>people</th>
<th>cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>same municipality</td>
<td>118,137</td>
<td>66.0</td>
</tr>
<tr>
<td>Zeeland</td>
<td>46,016</td>
<td>92.4</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>7,810</td>
<td>96.8</td>
</tr>
<tr>
<td>Belgium</td>
<td>5,213</td>
<td>99.8</td>
</tr>
<tr>
<td>other</td>
<td>393</td>
<td>100.0</td>
</tr>
</tbody>
</table>

available data

1,558,205 certificates and 5.6 million individual tokens (key information digitized with the help of volunteers since 1990)

<table>
<thead>
<tr>
<th>type</th>
<th>certificates</th>
<th>people mentioned (million)</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>birth</td>
<td>698,285</td>
<td>2.1</td>
<td>1811-1913</td>
</tr>
<tr>
<td>marriage</td>
<td>192,231</td>
<td>1.2</td>
<td>1811-1938</td>
</tr>
<tr>
<td>divorce</td>
<td>1,690</td>
<td>-</td>
<td>1811-1938</td>
</tr>
<tr>
<td>death</td>
<td>665,399</td>
<td>2.3</td>
<td>1811-1963</td>
</tr>
</tbody>
</table>

LINKS
Zeeland
Cleaned Dataset (Marriages, Births and Deaths), release 2016_01

certificates per year

matching of records

strategies:
- pairwise edit-distance
  - certificate-based or ego-based
  - various (edit-)distance measures
  - problems for large edit-distances
  - quadratic computation problem
- sorted neighborhood
  - ego-based
  - requires standardized data
  - matching within window (size=2) only
  - fast

sorting example

<table>
<thead>
<tr>
<th>ego surname</th>
<th>ego first name</th>
<th>sex</th>
<th>mother surname</th>
<th>mother first name</th>
<th>father surname</th>
<th>father first name</th>
<th>date birth</th>
<th>role</th>
</tr>
</thead>
<tbody>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>dishoek</td>
<td>petronella</td>
<td>doorn abraham</td>
<td>657040</td>
<td>deceased</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>dishoek</td>
<td>petronella</td>
<td>doorn abraham</td>
<td>657103</td>
<td>bride</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>doorn jacoba</td>
<td>doorn johannes</td>
<td>671562</td>
<td>child</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>doorn jacoba</td>
<td>doorn johannes</td>
<td>671566</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>nauta gerrit</td>
<td>doorn lornala</td>
<td>647143</td>
<td>deceased</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>schoner adrieza</td>
<td>doorn hendrik</td>
<td>681241</td>
<td>child</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>vlag</td>
<td>maria</td>
<td>doorn leendert</td>
<td>697569</td>
<td>bride</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doorn wilhelmina</td>
<td>f</td>
<td>vlag</td>
<td>maria</td>
<td>doorn leendert</td>
<td>697661</td>
<td>child</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

optimal field order for sorting key
for birth, marriage, death of ego

1. surname ego [distinguishes more than first name]
2. first name ego
3. sex ego
4. surname mother
5. first name mother
6. surname father [father not always known]
7. first name father
8. birth date ego [imprecise]

place of birth ego is not used [imprecise]
two sorting keys

- Birth: ego, mother, father, birth-date
- Marriage: ego, mother, father, age
- Death: ego, mother, father, age
- Children: ego, partner
- Death partner: ego, partner

Marriage is the bridge between two sortings.

name standardization

**first names**
- 21,157 different initial first names (5.6 million tokens)
- 14,163 names with 762 standards (99.06% of tokens)
- 6,994 names not-standardized (0.94% of tokens)

**surnames** (without particles: *de Vries*)
- 51,380 different surnames
- 39,335 semi-phonetized
- 12,782 limited to first 4 characters

matching of subsequent records

- **requirements**
  - all minus one standardized names match
  - same sex
  - date difference < 400 days

This results in blocks of records which internally subsequently match.

results

- no golden standard
- consistency check (events chronologically correct)
  - 508,862 consistent blocks
  - 2,939 inconsistent blocks (0.56%)
- IISH Amsterdam: edit-distance on certificates
  - 31,773 extra links at IISH (missing parents, missing dates, standardization error)
  - 236,959 links missing (incomplete analysis infant mortality, spelling variation in 1-4 names, cleaning for duplicate records)

non-matching original fields

<table>
<thead>
<tr>
<th>Levenshtein distance</th>
<th>ego surname</th>
<th>mother surname</th>
<th>ego first name</th>
<th>mother first name</th>
<th>father first name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>94.0</td>
<td>91.3</td>
<td>88.3</td>
<td>88.1</td>
<td>89.6</td>
</tr>
<tr>
<td>1</td>
<td>5.0</td>
<td>6.3</td>
<td>7.2</td>
<td>6.8</td>
<td>5.3</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>1.6</td>
<td>2.0</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>&gt;=3</td>
<td>0.2</td>
<td>0.8</td>
<td>2.5</td>
<td>3.3</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Every 10 links have on average at least one name with edit distance >=3.
non-linked ego records

life courses
we have ego: birth, marriage, decease
now add children and integrate more partners
second sorting key (all match)
1. surname ego
2. first name ego
3. sex ego
4. surname partner
5. first name partner
integrate results into previously obtained blocks (with rules)
+ children (birth, marriage, death)
+ partners (death)

life courses (per marriage)
• 285,583 consistent life course sections
  per partner of ego (with on average 11 events)

<table>
<thead>
<tr>
<th>event</th>
<th>fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>birth ego</td>
<td>0.28</td>
</tr>
<tr>
<td>marriage ego</td>
<td>1.00</td>
</tr>
<tr>
<td>death ego</td>
<td>0.63</td>
</tr>
<tr>
<td>birth child</td>
<td>3.44</td>
</tr>
<tr>
<td>marriage child</td>
<td>1.52</td>
</tr>
<tr>
<td>death child</td>
<td>2.39</td>
</tr>
<tr>
<td>death partner</td>
<td>0.62</td>
</tr>
</tbody>
</table>

• 20,061 life courses with multiple marriages

jigsaw of life courses
test life courses that interact
• for additional information
  1888: Catharina Vermeulen, age 45, marries Pieter Hubregtsen, age 29.
  1911: Catharina Vermeulen, age 68, dies.
  1928: Pieter Hubregtsen, age 69, dies.
    no mention of Catharina Vermeulen, because of 2nd marriage
• for consistency
  — marriage sequence (death of earlier partner)
  — child sequence (death of older child with same name)

a life cycle
1842: Catharina Vermeulen is born.
1872: child Maria Herman is born.
1874: child Tenente Herman is born.
1877: child Tenente Herman, age 0, dies.
1879: child Petrus Herman is born.
1881: partner Izaak Herman, age 41, dies.
1883: Catharina Vermeulen, age 38, marries Livinus Scheerens, age 54.
1888: child Abraham Scheerens is born.
1895: child Catholijntje Cornelis, age 39.
1903: child Abraham Scheerens, age 20, marries Adriana Simpelaker, age 19.
1928: child Abraham Scheerens, age 45, marries Catholijntje Cornelis, age 35.
1954: child Abraham Scheerens, age 72, dies.
1991: Catharina Vermeulen, age 68, dies.

issues for discussion
• need for standardization
  — level of standardization, multiple standards
  — sorting under conditions of less rich information
  — application domains
• use of patronymics in sorting
  — missing surnames (<1811)
• step-wise building of life cycles
  — Bertrand Russell: whenever possible, substitute constructions out of known entities for inferences to unknown entities
• golden standard and comparison procedures
  — testing the quality of results
  — Occam’s razor: the least number of individuals that can explain all data
to be continued on Thursday

The Zeeland challenge