Module 5: Fellegi-Sunter Method

Rebecca C. Steorts

joint with Olivier Binette

Reading

- Binette and Steorts (2020)
- Newcombe et al. (1959)
- Fellegi and Sunter (1969)

Agenda

- Soundex algorithm
- Newcombe algorithm
- Fellegi and Sunter method

Load R Packages

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.width=4, fig.height=3,
                      fig.align="center")
library(RecordLinkage)
library(blink)
library(phonics)
```

Background

- Soundex algorithm
- Likelihood ratio tests (LRT)

Soundex

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English.

- The goal is for similar words to be encoded to the same representation so that they can be matched despite minor differences in spelling.
- The Soundex algorithm was one of the first types of blocking used to our knowledge since it's intuitive and easy to use.

Example of Soundex algorithm

```r
soundex("Rebecca")
```

```
## [1] "R120"
```

```r
soundex("Rebekah")
```

```
## [1] "R120"
```

Example of Soundex algorithm

```r
soundex("Beka")
```

```
## [1] "B200"
```

```r
soundex("Becca")
```

```
## [1] "B200"
```

```r
soundex("Becky")
```

```
## [1] "B200"
```

Likelihood ratio test (LRT)

Please review or learn about LRTs if you are not familiar with these as these are the backbone of the Fellegi and Sunter method (1969).

https://www.sciencedirect.com/topics/computer-science/likelihood-ratio

Newcombe's Automatic Linkage of Vital Records

Newcombe et al. (1959). Published in *Science*:

Automatic Linkage of Vital Records*

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

Newcombe's Automatic Linkage of Vital Records

Newcombe et al. (1959) introduced a **probabilistic record linkage** technique and implemented it on the Datatron 205 computer.

Newcombe's Automatic Linkage of Vital Records

Newcombe et al. (1959) introduced a **probabilistic record linkage** technique and implemented it on the Datatron 205 computer.

The authors did the following:

- Stated record linkage as a statistical problem, proposing the first unsupervised probabilistic record linkage method.
- Illustrated that it could be implemented on a computer.

Newcombe's Automatic Linkage of Vital Records

Goal: Link **34,138 birth records** from 1955 in British Columbia **to 114,471 marriage records** in the preceding ten year period.

	Marriage record	Birth record
Husband's family name	Ayad	Ayot
Wife's family name	Barr	Barr
Husband's initials	J Z	J Z
Wife's initials	M T	B T
Husband's birth province	AB	AB
Wife's birth province	PE	PE

Table 1: Example of identity information from comparing marriage and birth records. This is adapted and translated from Table I of Newcombe (1969). AB and PE represent the Canadian provinces of Alberta and Prince Edward Island.

Newcombe's Automatic Linkage of Vital Records

Main contributions:

1. Sort records by the Soundex algorithm of family names.
2. When the Soundex coding agrees, an informal likelihood ratio test (LRT) determines if the record are matches/non-matches.

Newcombe's Automatic Linkage of Vital Records

The **performance of the method** was as follows:

- 10 record pairs were processed per minutes
- About 98.3% of the true matches were detected, and about 0.7% of the linked records were not actual matches.
- "by far the largest part of the effort" was the preparation of punched card files reproducing marriage records in an adequate format.

Newcombe's Automatic Linkage of Vital Records

The **performance of the method** was as follows:

- 10 record pairs were processed per minutes
- About 98.3% of the true matches were detected, and about 0.7% of the linked records were not actual matches.
- "by far the largest part of the effort" was the preparation of punched card files reproducing marriage records in an adequate format.

Unfortunately, we do not know exactly how the probabilities for the likelihood ratio test were computed in all cases.

Probabilistic Record Linkage

The work of Newcombe et al. (1959) led to one of the most seminal papers in the literature — Fellegi and Sunter (1969).

The Fellegi-Sunter model

Fellegi and Sunter (1969). Published in JASA:

A THEORY FOR RECORD LINKAGE*

IVAN P. FELLEGI AND ALAN B. SUNTER

Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

The Fellegi-Sunter model

Fellegi and Sunter (1969) formalizes Newcombe et al. (1959) in a decision-theoretic framework.

The Fellegi-Sunter model

Fellegi and Sunter (1969) formalizes Newcombe et al. (1959) in a decision-theoretic framework.

Given a pair of records, Fellegi and Sunter (1969) considers three possible actions:

- to *link* the record pairs;
- to *possibly link* the record pairs; or
- to *not link* the record pairs.

An "optimal" decision rule is proposed for this.

The Fellegi-Sunter model

Fellegi and Sunter (1969) formalizes Newcombe et al. (1959) in a decision-theoretic framework.

Given a pair of records, Fellegi and Sunter (1969) considers three possible actions:

- to *link* the record pairs;
- to *possibly link* the record pairs; or
- to *not link* the record pairs.

An "optimal" decision rule is proposed for this.

We will focus on the **model** (rather than the decision-theoretic framework).

The Fellegi-Sunter model

Basic elements:

- Two *databases* A and B
  - Duplication *across* but not within databases (bipartite record linkage).
- *Records* with corresponding *attributes* or *fields*
  - Name, age, address, SSN, etc.

The Fellegi-Sunter model

Our goal:

- Figure out which records refer to the same **entity** (a *person*, *object* or *event*.)

The Fellegi-Sunter model

Our goal:

- Figure out which records refer to the same **entity** (a *person*, *object* or *event*.)

How we'll do that:

- We will consider **record pairs** from databases A and B to obtain multidimensional measures of similarity.
- Based on these **measures of similarity**, we will group records together that refer to the same entity.

The Fellegi-Sunter model

Record no.	Field 1 First name	Field 2 Last name	Field 3 Age
1	Olivier	Binette	25
2	Peter	Hoff	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N_1	Beka	Steorts	NA

Record no.	Field 1 First name	Field 2 Last name	Field 3 Age
1	Oliver	Binette	26
2	Brian	K	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N_2	Frances	Hung	NA

Is Olivier Binette the same person as Oliver Binette?

The Fellegi-Sunter model

Fellegi and Sunter (1969) formalizes Newcombe et al. (1959) in a decision-theoretic framework.

We consider **three possible actions** for a given pair of records:

- to *link* them;
- to call them a *possible link*; or
- to *not link* them.

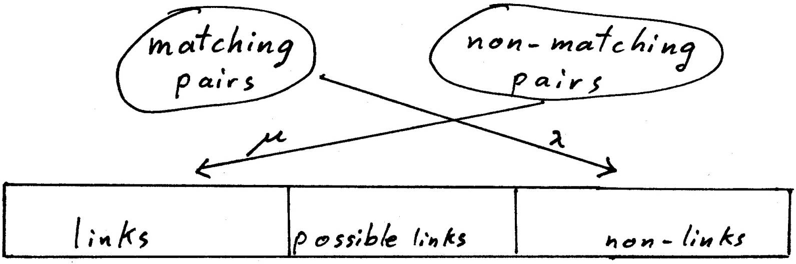
A Theory for Record Linkage

Consider **two error probabilities** (error rates):

$$\mu = \mathbb{P}\left(\text{linking} \mid \text{records do not match}\right),$$

$$\lambda = \mathbb{P}\left(\text{not linking} \mid \text{records do match}\right).$$

A Theory for Record Linkage



Goal of an **optimal decision procedure**:

Minimize the number of *possible links*, while achieving the above error rates at fixed levels μ and λ.

A Fundamental Theorem for Record Linkage

Fellegi and Sunter (1969) showed that the **optimal decision procedure** is obtained by a **likelihood ratio test**.

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

- For the ith pair of records, we compute a corresponding **comparison vector**

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \ldots, \gamma_i^{(k)}).$$

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

- For the ith pair of records, we compute a corresponding **comparison vector**

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \ldots, \gamma_i^{(k)}).$$

- Each γ_i^j compares the jth field of the records.

Example: Let the jth field be "age." Then $\gamma_i^j = 0$ if all ages are the same and $\gamma_i^j = 1$ if ages different.

Comparison Vectors

Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

Comparison Vectors

Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

Levels of agreement/disagreement:

- $\gamma_i^j \in \{0, 1, 2, \ldots, L_j\}$

Comparison Vectors

Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

Levels of agreement/disagreement:

- $\gamma_i^j \in \{0, 1, 2, \ldots, L_j\}$

How they're obtained:

- You choose!
- Use string distance functions to compare names.

Comparison Vectors

How can we visualize the comparison vectors?

$$\gamma_1 = (\gamma_1^{(1)}, \gamma_1^{(2)}, \ldots, \gamma_1^{(k)}) \tag{1}$$
$$\gamma_2 = (\gamma_2^{(1)}, \gamma_2^{(2)}, \ldots, \gamma_2^{(k)}) \tag{2}$$
$$\vdots \tag{3}$$
$$\gamma_{(N_1 \times N_2)} = (\gamma_{(N_1 \times N_2)}^{(1)}, \gamma_{(N_1 \times N_2)}^{(2)}, \ldots, \gamma_{(N_1 \times N_2)}^{(k)}) \tag{4}$$

Let

$$\gamma = (\gamma_1^{(1)}, \gamma_2^{(2)}, \ldots, \gamma_{(N_1 \times N_2)}^{(k)})$$

Likelihood Ratio Test

Define

$$m(\gamma) = \mathbb{P}(\gamma \mid \text{the records are a match}) \tag{5}$$
$$u(\gamma) = \mathbb{P}(\gamma \mid \text{the records are not a match}) \tag{6}$$

Likelihood Ratio Test

Define

$$m(\gamma) = \mathbb{P}(\gamma \mid \text{the records are a match}) \tag{5}$$
$$u(\gamma) = \mathbb{P}(\gamma \mid \text{the records are not a match}) \tag{6}$$

Then the **matching weight** or **log-likelihood ratio** is

$$W(\gamma) = \log(m(\gamma)/u(\gamma)) \tag{7}$$

Likelihood Ratio Test

Two thresholds T_μ and T_λ must be computed as a function of the desired error levels μ and λ.

Specifically, we

- *link* if $W(\gamma) > T_\mu$;
- *possible link* if $T_\mu \geq W(\gamma) > T_\lambda$; and
- do *not link* if otherwise $T_\lambda \geq W(\gamma)$.

We are ignoring the boundary cases (see Appendix 1 of Fellegi-Sunter (1969) for details).

Key Questions

1. How do we compute the probabilities

$$m(\gamma) = \mathbb{P}(\gamma \mid \text{the records are a match}),$$

$$u(\gamma) = \mathbb{P}(\gamma \mid \text{the records are not a match})?$$

2. Do we care about the error rates

$$\mu = \mathbb{P}\left(\text{linking} \middle| \text{records don't match}\right),$$

$$\lambda = \mathbb{P}\left(\text{not linking} \middle| \text{records are a match}\right)?$$

Proposed methods

Fellegi and Sunter (1969) proposed two methods in their paper for calculating $m(\gamma)$ and $u(\gamma)$.

They referred to these as Method 1 and Method 2, and thus, we will stick with the same terminology.

Method 1

Method 1: This is roughly what Newcombe el al (1959) proposed:

- Completely unsupervised
- Uses frequency of occurence of names, ages, addresses as additional information
- Requires prior knowledge of error rates (μ and λ). For some problems, these are known or can be estimated.

Example: At the U.S. Census Bureau, they currently use Method 1 in production for the decennial census and have prior knowledge of the error rates from working on the problem for a very long period of time.

Method 2

The second method applies to the comparison vectors $\gamma = (\gamma_1^{(1)}, \gamma_2^{(2)}, \ldots, \gamma_{(N_1 \times N_2)}^{(k)})$ under the following assumptions:

- $\gamma_i^j \in \{0, 1\}$ is a binary comparison vector
- $\{\gamma_i^j\}_{j=1}^k$ is conditionally independent given the true match or non-match status of the pair of records.

Method 2

Let M be the set of true matches among record pairs, U be the set of true non-matches.

Abusing notation, the idea is to consider the equations:

$$\begin{aligned} P(\gamma) &= P(\gamma \mid M)P(M) + P(\gamma \mid U)P(U) \\ &= \left\{\prod_{i=1}^{k} P(\gamma_i|M)\right\} P(M) + \left\{\prod_{i=1}^{k} P(\gamma_i|U)\right\} P(U) \\ &= \left\{\prod_{i=1}^{k} m(\gamma_i)\right\} P(M) + \left\{\prod_{i=1}^{k} u(\gamma_i)\right\} (1 - P(M)), \end{aligned}$$

which are $2^k - 1$ equations for $2k + 1$ variables; there can be a solution when $k \geq 3$.

To solve for $m(\gamma_i)$ and $u(\gamma_i)$, the EM algorithm is used.[1]

[1] The EM algorithm iteratively solves for the unknown m and u parameters, where it finds a local maximum for the likelihood.

Final Question

Do we care about the error rates

$$\mu = \mathbb{P}\left(\text{linking}\middle|\text{records don't match}\right),$$

$$\lambda = \mathbb{P}\left(\text{not linking}\middle|\text{records are a match}\right)?$$

In practice, we do not know if a given pair of record is a match or not, so put simply, we cannot answer this question directly.

Final Question

We can answer related questions, such as:

- $\mathbb{P}(\text{records don't match} \mid \text{we linked them})$; or
- $\mathbb{P}(\text{records match} \mid \gamma)$

To explore this more in depth, see Binette and Steorts (2020) to see connections to Bayes' rule and see Tepping (1968).

Summary

- Soundex algorithm
- What are the main contributions of Newcombe et al. (1959)?
- Comparison vectors
- What did Fellegi and Sunter (1969) propose (two methods) ?
- How would you summarize the main ideas of this lecture and summarize it to a friend?

Supplement

In the supplement, I provide the math that is the basis this problem, which is important to the paper and extensions moving forward.

Following notation in Kundinger et. al (2024).

Notation

Suppose we have two data sets A and B with n_1 and n_2 records, respectively.

Each records have F fields of information.

We represent the biparite matching by $Z = (Z_1, \ldots, Z_{n_2})$, where

- $Z_j = i$ for $i \in [n_1]$ if record $j \in B$ matches record $i \in A$ and
- $Z_j = n_1 + j$ if record j in B does not match any record in A.

Comparison Data

Let

$$\gamma_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^F)$$

, where γ_{ij}^f is the comparison of record i, j for field f.

Collect all comparison vectors into the matrix

$$\Gamma = \{\gamma_{ij}\}_{i=1,j=1}^{n_1,n_2}.$$