Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering

Jared S. Murray (2016)

Sep 20, 2016

Fellegi-Sunter Framework

Recall in Fellegi-Sunter model (1969), we assume each pair of records (a, b) is from either $M = \{(a, b)|a = b\}$ or $U = \{(a, b)|a \neq b\}$.

For each pair, we observe a comparison vector $\gamma(a, b)$ and make decision based on the following quantities:

$$P[(a, b) \in M] = p_M \tag{1}$$

$$P[\gamma_{ab} = g|(a, b) \in M] = \pi_{g|M} \tag{2}$$

$$P[\gamma_{ab} = g|(a, b) \in U] = \pi_{g|U} \tag{3}$$

$$P[\gamma_{ab} = g] = p_M \pi_{g|M} + (1 - p_M)\pi_{g|U} \tag{4}$$

Fellegi-Sunter Framework

Goal: Partition the comparison space into Γ_M, Γ_U, Γ_C:

1. Γ_M: match region
2. Γ_U: non-match region
3. Γ_C: indeterminate region

based on

$$w_g = \frac{P[\gamma_{ab} = g|(a,b) \in M]}{P[\gamma_{ab} = g|(a,b) \in U]} = \frac{\pi_{g|M}}{\pi_{g|U}}$$

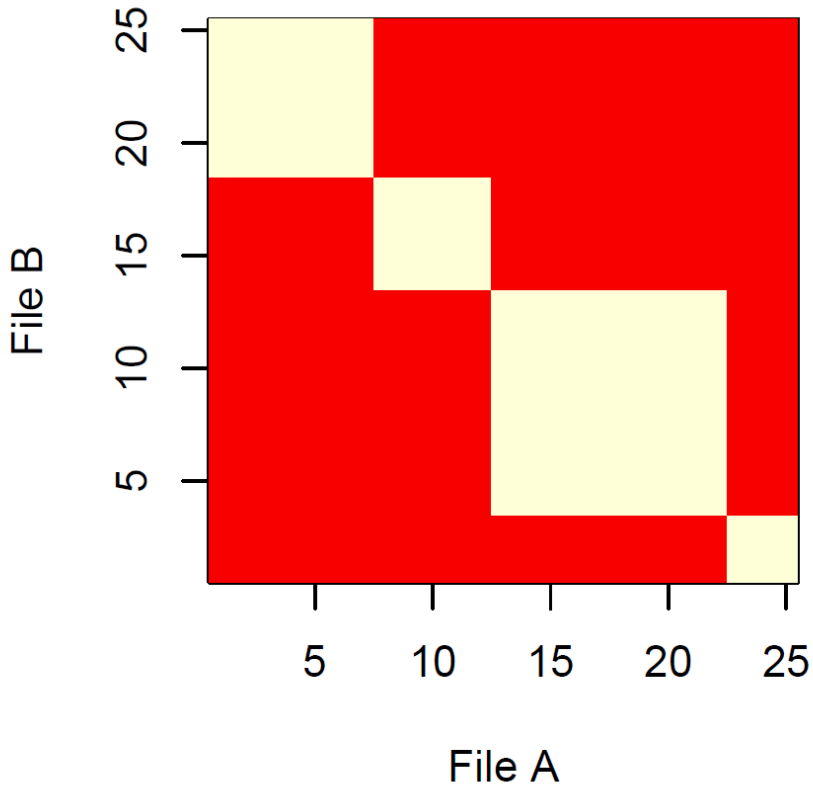
How do these estimates change after blocking, indexing, and filtering?

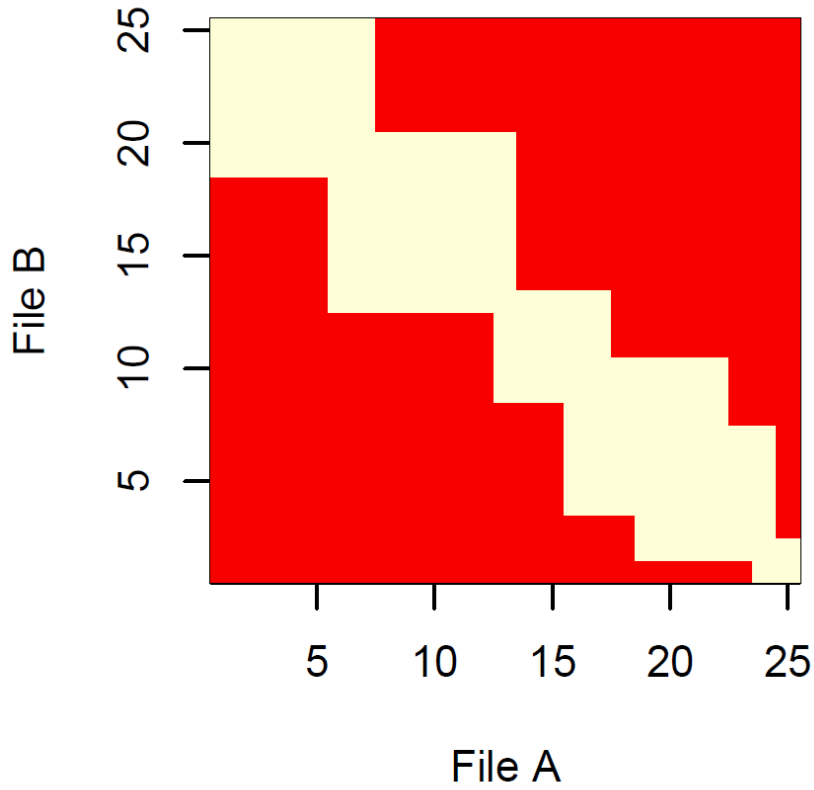
Reducing the Number of Comparison

1. **Indexing**: Any method to reduce the number of candidate pairs.
2. **Blocking**: Any indexing method that partitions records based on the values of a key. (Zip code matches.)
3. **Filtering**: Discard pairs not excluded by initial indexing but which are still unlikely to be matched. Filtering can be interpreted as a special indexing method after a full comparison. (Drop $\gamma = (0, \ldots, 0)$)

Indexing by disjunctions can be considered as a union of multiple results from blocking, which is a very common kind of methods in practice.

Reducing the Number of Comparison





Modeling after Blocking

Let β_{ab} an additional comparison for some blocking criterion.

$$P[(a,b) \in M | \beta_{ab} = 1] = p_{M|\beta} \tag{5}$$

$$P[\gamma_{ab} = g | (a,b) \in M | \beta_{ab} = 1] = \pi_{g|M,\beta} \tag{6}$$

$$P[\gamma_{ab} = g | (a,b) \in U | \beta_{ab} = 1] = \pi_{g|U,\beta} \tag{7}$$

$$P[\gamma_{ab} = g | \beta_{ab} = 1] = p_{M|\beta}\pi_{g|M,\beta} + (1 - p_{M|\beta})\pi_{g|U,\beta} \tag{8}$$

Throw pairs with $\beta_{ab} = 0$.

Modeling after Blocking

Throw pairs with $\beta_{ab} = 0$:

1. $p_{M|\beta} >> p_M$
2. $\pi_{g|M,\beta} \approx \pi_{g|M}$
3. $\pi_{g|U,\beta} < \pi_{g|U}$ when g is "close" to $(0, 0, \ldots, 0)$
4. $\pi_{g|U,\beta} > \pi_{g|U}$ when g is "close" to $(1, 1, \ldots, 1)$

Modeling after Indexing/Filtering

Let ℓ_{ab} an additional comparison for some indexing criterion.

We do not need to include the comparison β_{ab} in γ_{ab} since it is always one in retained pairs. However, this is not true for indexing by disjunction and filtering.

If we index by the disjunction of exact matches on age and zip code:

$$P((a,b) \text{ disagree on age and zip code} | \beta_{ab} = 1, \ell_{ab} = 1) = 0$$

Remark: Indexing by disjunction may change the support of the comparison space. Filtering always changes the support of the comparison space.

Modeling Record Pairs after Indexing by Disjunctions/Filtering

Indexing by disjunction gives similar bias on parameters but also incomplete cells in the contingency table.

Suppose we have two comparisons, then:

$$\begin{array}{c} \\ \gamma(1) \end{array} \begin{array}{c} \gamma(2) \\ \begin{pmatrix} n_{00} & n_{10} \\ n_{01} & n_{11} \end{pmatrix} \end{array} \tag{9}$$

Filtering ignores n_{00}, if the rule is to drop all pairs of records giving full-zero comparison vectors.

Use structural zero formulation and modify the estimates.

Modeling Record Pairs after Indexing by Disjunctions/Filtering

Simply treating structural zeros as sampling zeros biases estimates(e.g. Manrique-Vallier and Reiter (2014)).

We need to constrain the support from Γ to $\Gamma_\ell = \{\gamma_{ab} | P(\gamma_{ab} | \beta_{ab}, \ell_{ab}) > 0\}$ and fit a conditional quasi-independence (CQI) model (Goodman, 1968; Fienberg, 1970; Bishop et al., 1975):

$$\pi_{g|M,\beta,\ell} \propto \prod_{j=1}^{p} \psi^{(j)}_{g(j)|M,\beta,\ell} 1(g \in \Gamma_\ell) \tag{10}$$

$$\pi_{g|U,\beta,\ell} \propto \prod_{j=1}^{p} \psi^{(j)}_{g(j)|U,\beta,\ell} 1(g \in \Gamma_\ell) \tag{11}$$

which can be estimated via EM.

Modeling Record Pairs after Indexing by Disjunctions/Filtering

Recall originally we assume:

$$\pi_{g|M} \propto \prod_{j=1}^{p} \rho_{j|M}^{g(j)} (1 - \rho_{j|M})^{1-g(j)} \tag{12}$$

$$\pi_{g|U} \propto \prod_{j=1}^{p} \rho_{j|U}^{g(j)} (1 - \rho_{j|U})^{1-g(j)} \tag{13}$$

where $\rho_{j|M} = P((a, b)$ match on comparison j$|(a, b) \in M)$

Results: Synthetic Data (RLdata 10000)

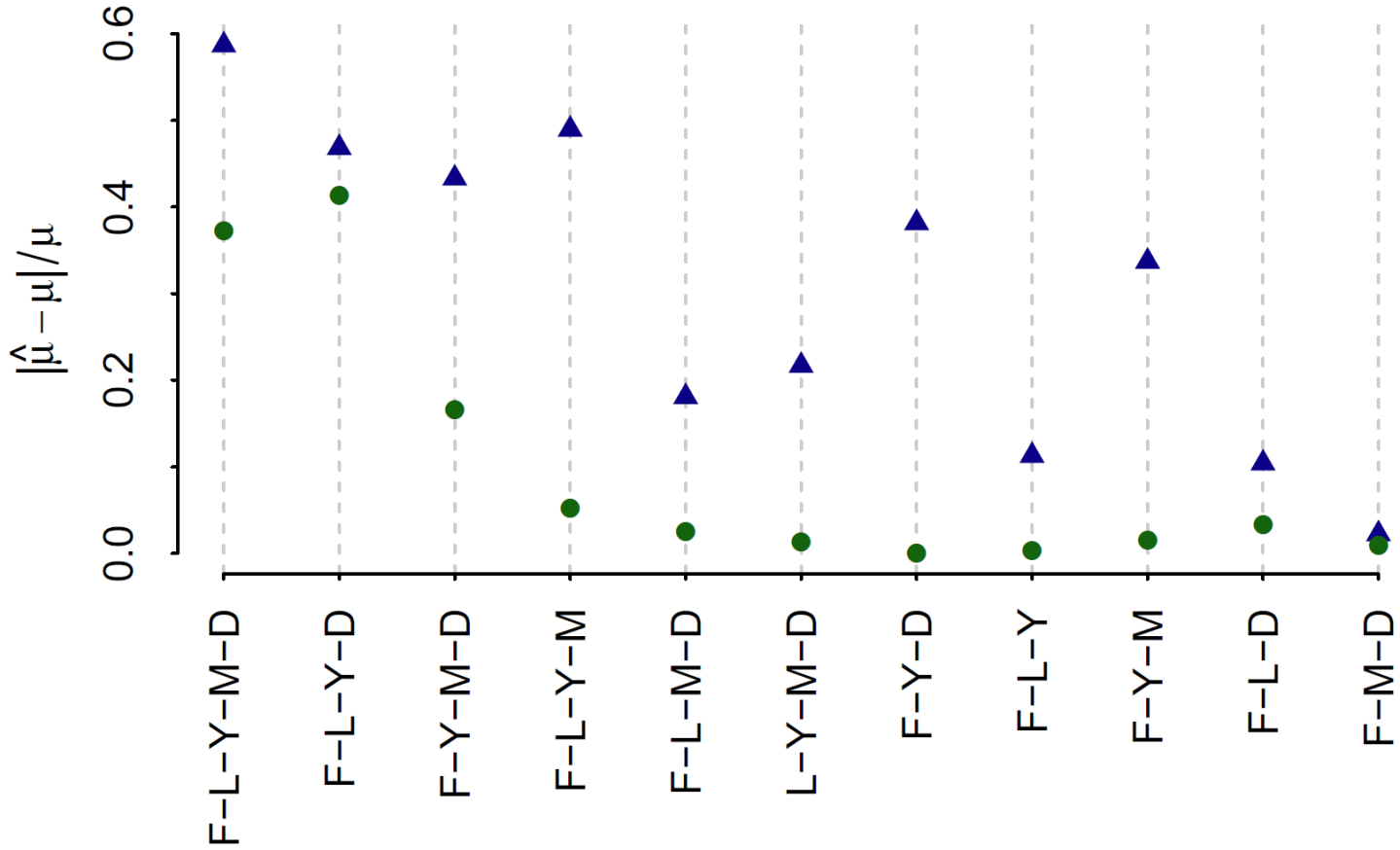


Figure: Absolute relative discrepancy in the estimate of error rates with filtering (Grean) and without filtering (Blue)