Using Random Samples in Entity Resolution Applications: An Example Solution to Homework 2

Olivier Binette

February 25, 2020

The goal of this homework is to investigate how "representative" samples can be obtained in the context of entity resolution, for the purpose of evaluating ER performance metrics. The four tasks of the homework use the `RLdata500` dataset to walk us through an exploration of the issue, the proposal of a solution, and its evaluation.

Here I consider the practical scenario where ground truth is only available for selected subsets of the data and is not available for the whole. That is, while unique entity identifiers are available for the `RLdata500` dataset, we will for the most part ignore them. They are only used as part of the exploratory data analysis and to obtain ground truth on samples of records (in practice, ground truth for small samples of records would be obtained through clerical review).

Furthermore, I focus on the problem of estimating the *level of duplication* in the dataset. While this is simpler than the problem of estimating general ER performance metrics, the main issues remain the same. We can view approaches for estimating the level of duplication as providing basic frameworks under which estimation techniques for other quantities could be developped.

Task 1

Start by doing an exploratory analysis of the data set. What do you find?

Solution

Table 1 shows the structure of the `RLdata500` dataset and its first few rows, when sorted by last name.

First name		**Last name**		**Birth date**		
`fname_c1`	`fname_c2`	`lname_c1`	`lname_c2`	`by`	`bm`	`bd`
RAINER	NA	ALBRECHT	NA	1976	11	10
UWE	NA	ALBRECHT	NA	1967	1	8
ANNA	NA	ALBRECHT	NA	1964	10	22
HANNELORE	NA	ALBRECHT	NA	1963	12	4
ANNA	GUDRUN	ALBRECHT	NA	1948	10	30

Table 1: First five rows of the `RLdata500` dataset when sorted by last name.

The first and last names are each separated in two components. Birth year, month, and day are separately recorded.

In Figure 1, we look at the frequency distribution of the first and last names (first components only) and of the birth date fields. Note that there are no missing values among these attributes. As for secondary name components, only 28 records have a second first name, and only 8 records have a second last name.

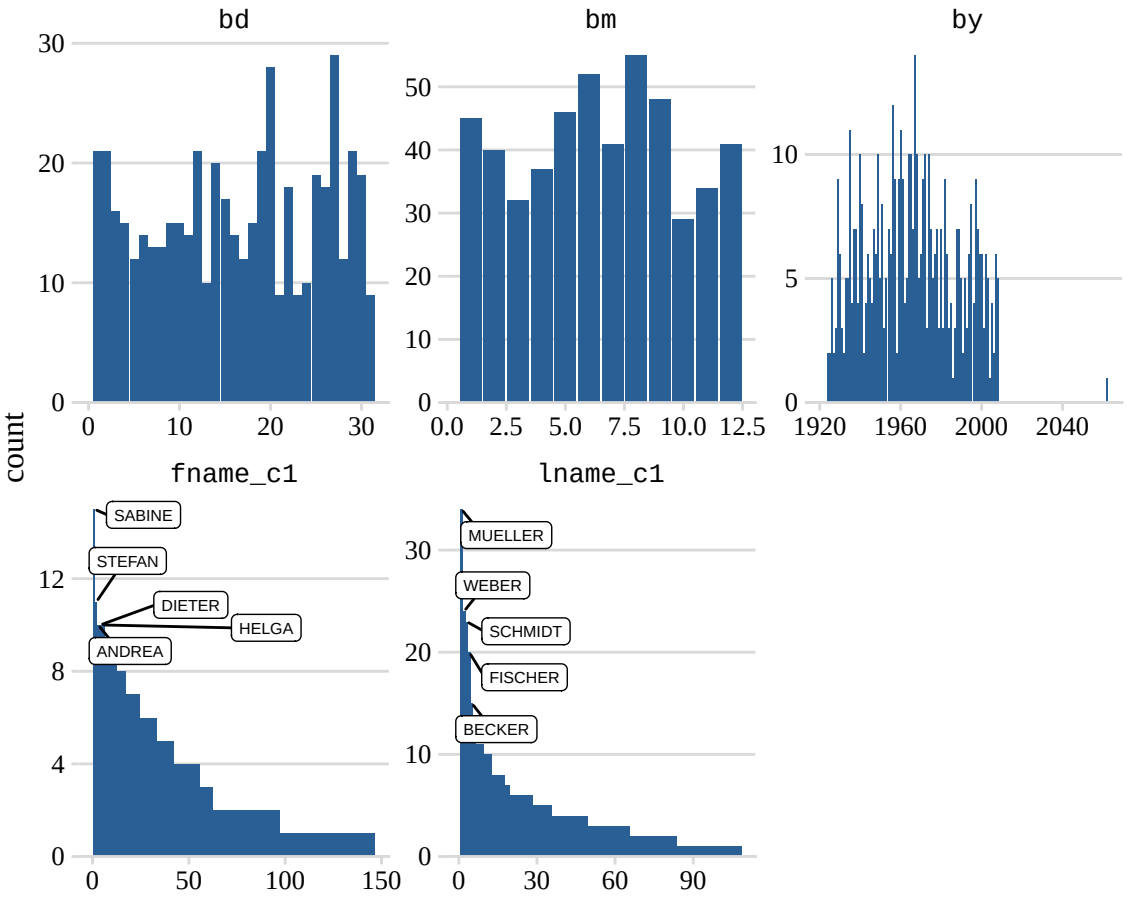

Figure 1: Frequency distribution of main record attributes. Note that first and last names have been reordered by frequency and the x-axis corresponds to unique name index.

The birth day `bd` and birth month `bm` seem roughly uniformly distributed, while birth year `by` is more concentrated around 1960. An erroneous birth year of 2062 is listed on one of the record. We can observe more duplication among last names than among first names. First name may therefore be more discriminative of distinct individuals than last name, assuming comparable error levels.

Finally, we visualize the differences between duplicated records using the `visdat` package. Recall that `RLdata500` contains 50 duplicated records, each with a corresponding original. Figure 2 illustrates the differences between original and duplicated records.

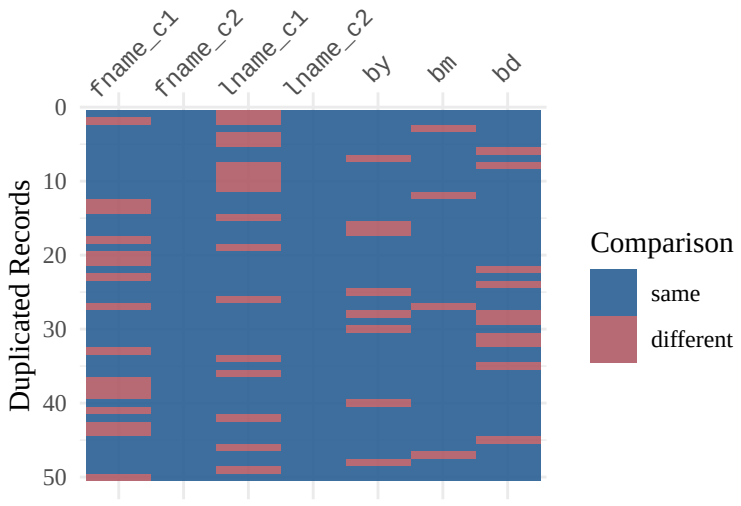

Figure 2: Visualization of the differences between the 50 original records that have been duplicated and slightly modified in the `RLdata500` dataset. Each row represent one of the duplicated record. Each column indicates whether the duplicated record matches its original version in the given field. Observe that each duplicated record differs from its original by exactly one attribute.

Task 2

What happens if you randomly sample 10 records from the original dataset? Do this a few times and describe what happens? Is this representative of the original dataset? Explain and be specific.

Solution

Let's first sample 10 records from the original dataset and take a look at the result in Table 2.

First name		Last name		Birth date			
fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	ID
KATHRIN	NA	KOCH	NA	2005	12	1	259
ROLF	NA	NEUMANN	NA	1967	3	29	68
BRIGITTE	EDITH	PETERS	NA	1956	8	21	389
INGRID	NA	KOERTIG	NA	1960	5	6	430
ANJA	URSULA	WEBER	NA	1995	6	26	415
MONIKA	NA	WEISS	NA	1994	8	19	369
GUENTHER	FRIEDRICH	WOLF	NA	1975	4	23	116
SUSANNE	NA	WEBER	NA	1997	11	25	265
DANIEL	NA	SCHMIDT	NA	1978	3	4	133
SABINE	NA	GRAF	NA	1980	9	5	295

Table 2: Ten random rows from the `RLdata500` dataset with unique identifiers.

In comparison to the full dataset, there is no duplicated record in this sample. Furthermore, there is no duplicate first name, no duplicate last name, no duplicate birth year, and no duplicate birth day. This particular sample therefore provides little to no useful information regarding the level of duplication in the data or regarding the distribution of the attributes.

Now supposed we wished to estimate the precentage of duplicate records, or *level of duplication*, in the whole dataset using such random samples. This problem of estimating the number of duplicate records is also called *unique entity estimation* (Chen, Shrivastava, and Steorts 2018); the goal is to estimate the number of unique entities represented in the dataset.

Would the percentage of duplication in random samples be representative of duplicate in the whole? Figure 3 shows the distribution of the duplication level in 100,000 random samples of size 10 and compares it to the level of duplication in the whole dataset (10%).

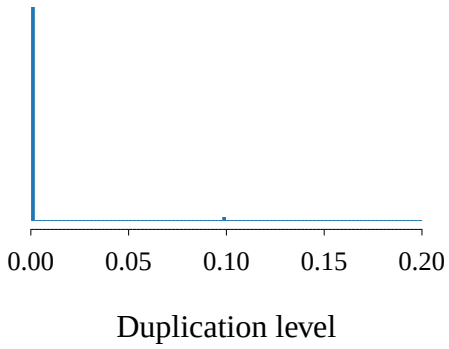

Figure 3: Histogram of duplication levels in 100,000 random samples of size 10 from the `RLdata500` dataset.

The mean level of duplication in the samples is only around 0.002, far from the target 10%.

The naive duplication estimator, taking the observed mean duplication of a random sample, is **highly biased** here. To see why this is the case, consider the coreference matrix C, defined as $C = [c_{i,j}]$ with $c_{i,j} = 1$ if records i and j match, and $c_{i,j} = 0$ otherwise. If we sample k records, this corresponds to sampling $k(k-1)/2$ entries in the lower triangular section of C. The expected number of duplicates in this section is then around $\ell k(k-1)/(n-1)$. While we can adjust for the factor of $k(k-1)/(n-1)$ to obtain an unbiased estimator, the result would be highly inefficient (see Raj (1961), Section 3, for a proof of unbiasedness and a computation of the variance of this estimator).

We would face similar problems if trying to compute precision and recall of a proposed ER method on a subset of the data. An ER method which does not match anything would perform quite well on subsets of the data in terms of both precision and recall. However, its recall would be close to zero on the whole dataset.

There is therefore a need to both:

1. account for the unrepresentativeness of record samples in ER applications (such as by using adjustment factors to obtain unbiased estimators), and
2. propose ways to obtain more representative samples (as to improve the efficiency of estimators).

Tasks 3 and 4 deal with points (1) and (2).

Tasks 3 and 4

Propose something that works better than random sampling and explain why this works better. Propose evaluation metrics, visualizations, etc, to support any of your claims.

Solution

Recall that we focus on the problem of estimating the level of duplication in the whole dataset (this is the unique entity estimation problem discussed in the solution to Task 2).

Here I propose to use a blocking approach: given any set of blocks which partition the record space, a number of them will be sampled with probability proportional to their size. The level of duplication in the dataset is then estimated as the average $\hat{\ell}$ of the level of duplication within each block.

Proposition: If the blocking approach has recall R, then $\mathbb{E}[\hat{\ell}] = R\ell$.

Proof: Let b_i, $i = 1, 2, \ldots, p$ be the sizes of the blocks, and let $N = \sum_i b_i$ be the total number of records. Each block b_i is sampled with probability b_i/N. Now let D be the total number of duplicate records and let d_i be the number of duplicates in block i. Since the blocking approach has recall R, we therefore have $\sum_i d_i = RD$. We can then compute

$$\mathbb{E}[\hat{\ell}] = \sum_{i=1}^{p} \frac{d_i}{b_i} \frac{b_i}{N} = \frac{1}{N} \sum_{i=1}^{p} d_i = \frac{RD}{N} = R\ell.$$

Note that the recall R can be estimated by sampling multiple blocks, and therefore the estimator $\hat{\ell}$ can be recall-adjusted to be approximately unbiased.

To illustrate this approach, consider blocking by the first letter of the last name. This blocking approach has perfect recall $R = 1$. In Figure 4, we illustrate the duplication level within each block, as well as the expectation of $\hat{\ell}$ and the value $R\ell$.

Next consider blocking by birth day `bm`, which has lower recall of 0.8. Figure 5 shows the results in this case.

In both cases our proposition is satisfied.

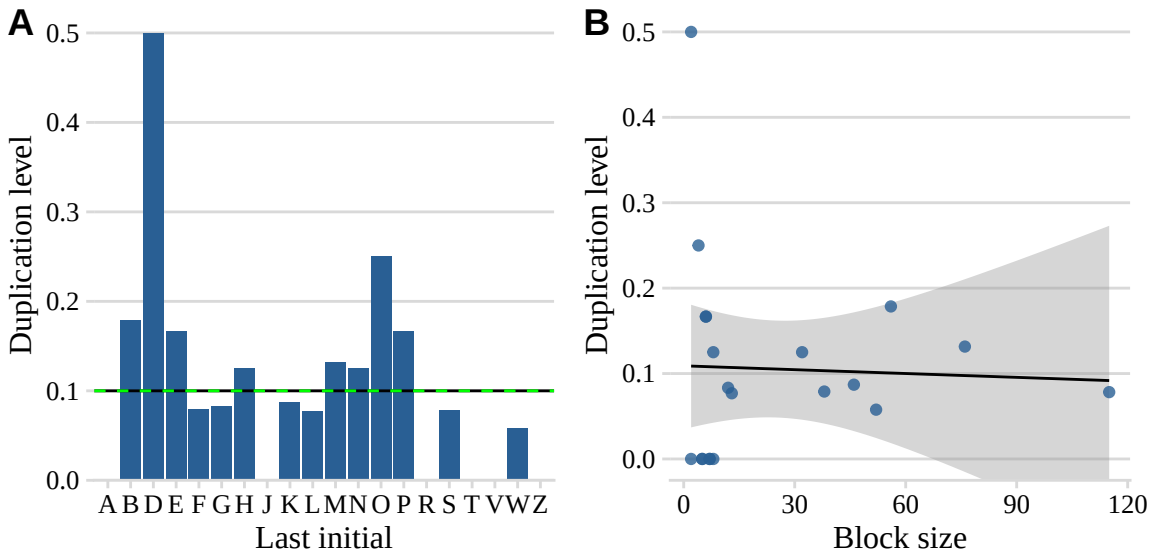

Figure 4: Panel **A**: Duplication level within each block for last name initial blocking. The horizontal black line represents the expected value of $\hat{\ell}$ and the coïnciding dotted green line represents the value $R\ell$. Panel **B**: Scatter plot of block size and duplication level, with a linear regression line and 95% confidence band.

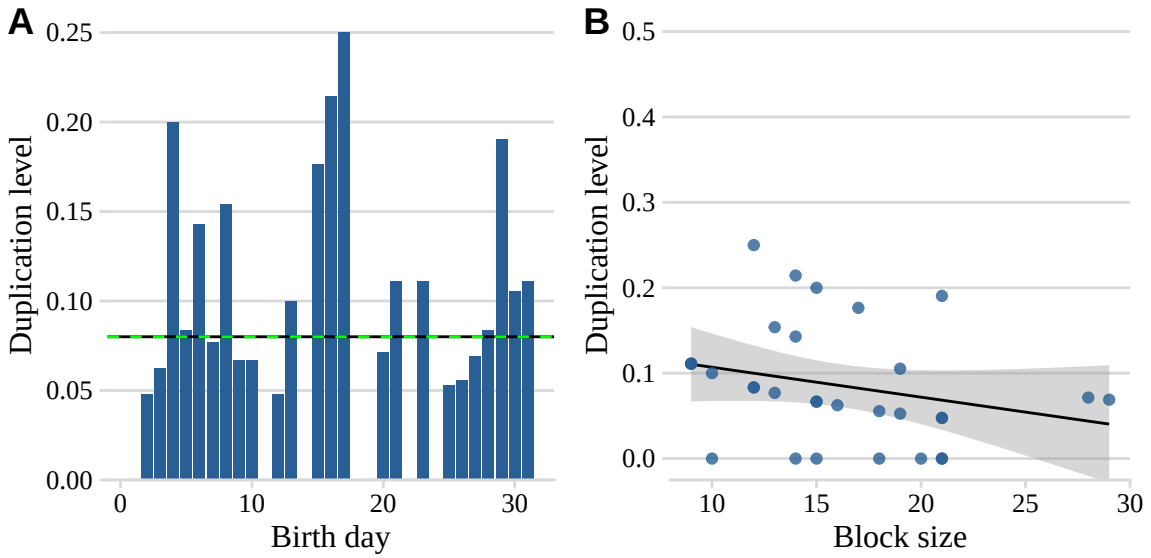

Figure 5: Panel **A**: Duplication level within each block for birth day blocking. The horizontal black line represents the expected value of $\hat{\ell}$ and the coïnciding dotted green line represents the value $R\ell$. Panel **B**: Scatter plot of block size and duplication level, with a linear regression line and 95% confidence band.

Practical implications

Recall how Sadinle (2014) used a single block to evaluate precision and recall of his proposed record linkage approach for the El Salvadorian data set. It is currently unclear if an adaptation of our approach would be preferable to Sadinle's approach. That is, the issue is to determine if sampling a single large block of size N to evaluate level of duplication is preferable to sampling a larger number of blocks of size $n_1, n_2, \ldots, n_k$, with $\sum_i n_i = N$, using our technique and adjusted estimators.

To gain insight into this issue, consider the following experiment, which compares our approach (**method 1**) to the equivalent of Sadinle's approach (**method 2**) for the purpose of estimating the level of duplication. We block by birth day (recall is 0.8) and sample $k = 10$ blocks with probability proportional to their size, without replacement. On average, around 175 records are sampled. Under **method 1**, we compute the duplication level within each block, average those, and adjust the result using a naive recall estimator (a simple bias-adjusted estimator). Under **method 2**, we simply compute the duplication level in the aggregation of all sampled blocks. This is experiment is replicated 5000 times and properties of the estimators are shown in Table 3.

Method	Mean	RMSE
1	0.104	0.029
2	0.083	0.025

Table 3: Comparison of **method 1** and **method 2**, under birth day blocking and sampling $k = 10$ blocks, in terms of mean value and root mean squared error (RMSE). Here the estimand is the duplication level of 0.1.

Method 1 is much less biased than **method 2**, but has a slightly higher root mean squared error. The higher variance of **method 1** is due to the estimation of the recall R and the resulting ratio estimator. By regularizing the recall estimate, we can actually obtain a method which has lower RMSE than both **method 1** and **method 2**. This approach is evaluated under **method 3** in Table 4.

Method	Mean	RMSE
3	0.092	0.022

Table 4: Evaluation of **method 3**, under birth day blocking and sampling $k = 10$ blocks, in terms of mean value and root mean squared error (RMSE). Here the estimand is the duplication level of 0.1.

The issue of estimating recall

The main bottleneck in **method 1** and **method 3** is estimating recall for bias adjustment. Ideally we would be able to estimate recall without looking at all possible links across a set of blocks. This would greatly increase the efficiency of the estimator in terms of the number of possible links that have to be inspected in order to obtain a precise estimate. This is not something we explore further in this homework.

Discussion

In this homework, we considered the problem of estimating the level of duplication ℓ of a dataset. We proposed to use *blocking* and to do probability sampling of blocks rather than sampling records at random. Duplication level within sampled blocks was averaged in order to obtain an estimator $\hat{\ell}$. We observed that $\mathbb{E}[\hat{\ell}] = R\ell$, where R is the recall of the blocking approach. In cases where $R \approx 1$, our approach therefore provides a nearly unbiased estimator of the duplication level.

Using recall estimators, we also obtained recall-adjusted estimators of the duplication level (**method 1** and **method 3**). These estimators were compared to the naive approach, comparable to what was used in Sadinle

(2014), of using the observed duplication level in the aggregation of the sampled blocks (**method 2**). In our experiments, it appeared that **method 1** is nearly unbiased, while **method 3** balances bias and variance for the lowest RMSE.

References

Chen, Beidi, Anshumali Shrivastava, and Rebecca C Steorts. 2018. "Unique Entity Estimation with Application to the Syrian Conflict." *The Annals of Applied Statistics* 12 (2): 1039–67.

Raj, Des. 1961. "On Matching Lists by Samples." *Journal of the American Statistical Association* 56 (293): 151–55.

Sadinle, Mauricio. 2014. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics* 8 (4): 2404–34.