



SUPPLEMENTARY MATERIAL

Supplementary Material: A flexible model for Record Linkage

Kayané Robach ^{1,2,*} Stéphanie van der Pas,^{1,2} Mark van de Wiel^{1,2}
and Michel Hof ^{1,2}

¹Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands and ²Amsterdam Public Health, Methodology, The Netherlands

*Corresponding author. k.c.robach@amsterdamumc.nl

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

This supplementary material contains additional application of *FlexRL*, on the often-used Survey of Household Income and Wealth (SHIW) and RLdata500. As illustrated with the National Long Term Care Survey (NLTCs) data sets in the main article ‘A Flexible model for Record Linkage’, we show the performance of different Record Linkage methods on regional subsets of the data as well as the performance of *FlexRL* on the complete data sets. We study the influence of the linkage probabilities threshold on the final set of links and we explore the robustness of the method under different distortion levels. The experiments can be reproduced using the online vignette, Robach [2024]. The proposed algorithm *FlexRL* is implemented and available in an open source R package.

Key words: Record Linkage, Partially Identifying Variables, Stochastic EM, Latent Variable Model

Overview

This document gather additional data applications for the article ‘A flexible model for Record Linkage’. The open-source software developed is available online, as well as the used data sets and a vignette to reproduce experiments. As the main innovations, our novel method addresses time-varying variables like place of residence and can be used on large data sources without substantial computational resources.

Here we present a real data application on the often-used Survey of Household Income and Wealth (SHIW) in section 2, Steorts et al. [2016], Guha et al. [2022], Pacheco Menezes et al. [2024] and, an application on the synthetic RLData500 in section 3, also broadly used in the record linkage literature, Sariyar and Borg [2010], Steorts [2015], Steorts et al. [2018], Enamorado and Steorts [2020], Omar et al. [2022], Bai et al. [2023], Sosa and Rodríguez [2024]. In each section we first describe the data and present the Partially Identifying Variables (PIVs) used to perform record linkage. The number of unique values of the PIVs assesses their discriminative strength and the level of disagreement and missing values in true links help us forming an opinion on the distortion level of the data. Overall the data applications, we chose to consider all PIVs as stable since we did not have access to informative data to build a model for their instability; this decision is consistent with the level of disagreement in the true links values.

We then compare our method to a simplistic approach, which links the records for which all the PIVs match, regardless of the one-to-one assignment constraint. The amount of *FP* (in comparison with the *TP*) detected by the simplistic approach certify of the level of complexity of the task. We also compare our method to *Exchanger* and *BRL*. However, when the data sets are too large those methods do not run on our computer so we apply the methods on subsets of the data. We present the results with boxplots, showing the performances variability over the data subsets. We also discuss the choice of the linkage probabilities threshold to build a final set of linked records in our method.

The simulation study presented in section 6 of the article showed the robustness of *FlexRL* to the data quality. In a later subsection we characterise the level of distortion of the real data using the sum of medians of disagreements and missing values among the PIVS of true links. Then we artificially distort the data by creating registration errors and we study the performance evolution of our method on different distortion levels.

Application: The Survey of Household Income and Wealth (SHIW)

Description

The SHIW data are made available by the Bank of Italy for research purposes, BancadItalia. The survey has been conducted every two year since 1989 and consist of about 20 000 individuals per census. A unique identifier can be inferred from the family and member identifiers in each sample, so that we can deduce the true linkage structure. We use data from 2016 and 2020 gathering six PIVs: sex, birth year and birth region, marital status, regional code and level of education. The data sets contain approximately 15000 and 16500 records respectively (after filtering the data to obtain the same support), of which 6400 are common to both files. We describe the often used data in table 1. We consider all PIVs stable because we do not have access to the registration time, nor to good explanatory variables to explain the evolution of education level for instance. The numbers of unique values and the proportion of agreement among true links assess the degree of difficulty of the task and portray the true links.

data	unique values	sex	birth year	marital status	regional code	birth region	education
		2	97	4	20	20	6
	type	categorical	categorical	categorical	categorical	categorical	categorical
true links	% agreement	1	.98	.94	1	.94	.77

Table 1. Summary of the full SHIW data of 2016 and 2020. Characteristics of the PIVs and level of agreement among the 6 430 links referring to the same individuals. There are 5% of missing value in birth region in both data sets.

Comparison with the literature on regional subsets

In order to compare *FlexRL* with the methods developed in the literature we divide the data sets into regional subsets as done for the NLTCs in section 6.2. We show the variability in performances over 20 regional subsets of the data: Piedmont, Aosta Valley, Lombardy, Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Liguria, Emilia-Romagna, Tuscany, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria, Sicily, Sardinia.

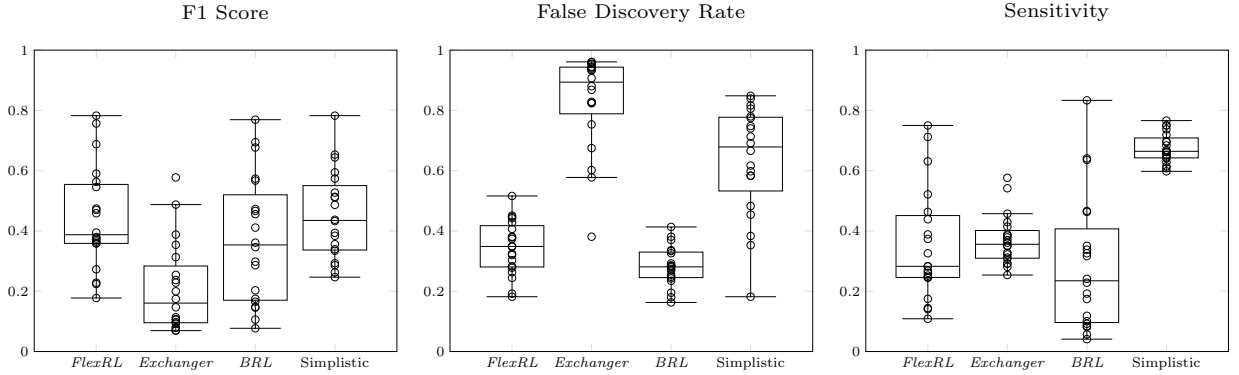


Fig. 1. Boxplots showing the variability of the F1 Score, the False Discovery Rate (FDR) and the Sensitivity of the compared methods on regional subsets of the SHIW data sets: Piedmont, Aosta Valley, Lombardy, Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Liguria, Emilia-Romagna, Tuscany, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria, Sicily, Sardinia.

This study demonstrates the good performance of *FlexRL* in a real setting. *BRL* achieves a similar balance between FDR and sensitivity, however the simplistic approach and *Exchanger* have an excessively high FDR in comparison to the others.

There is a high proportion of agreement in the pairs linked by *BRL*, which explains that the method achieve the lowest level of *FP*. *Exchanger* is less conservative and links more uncertain pairs, to the detriment of its performance, due to a higher number of *FP*. *FlexRL* is positioned between those two methods and capture more *TP* than *BRL* but also more sensible pairs than *Exchanger*.

Performance on the complete data sets

The task is not straightforward on these data sets, as one can judge based on the number of unique values, the distortion level of the data—asserted by disagreement and missing values in true links—and the amount of *FP* detected by the simplistic approach (see table 2).

None of the methods perform very well though the simplistic approach has the highest sensitivity. However, despite more *TP* detected, the simplistic approach leads to a very high amount of *FP*.

The parameters of the *FlexRL* method converge, however some parameters for the probability of mistake are biased, as it is the case for the estimated proportion of links, which is overestimated.

	linked record		FN	F1 Score	FDR	Sensitivity
	TP	FP				
<i>FlexRL</i> (0.5)	2435	6711	3995	.31	.73	.38
Simplistic approach	4318	13807	2112	.35	.76	.67
<i>FlexRL</i> (0.6)	2266	5438	4164	.32	.71	.35
<i>FlexRL</i> (0.7)	2091	4490	4339	.32	.68	.33
<i>FlexRL</i> (0.8)	1941	3749	4489	.32	.66	.30
<i>FlexRL</i> (0.9)	1788	3080	4642	.32	.63	.28

Table 2. Performance of *FlexRL* on the complete SHIW data for several linkage probability threshold (superior to 0.5 to ensure a one-to-one assignment) and of the simplistic approach.

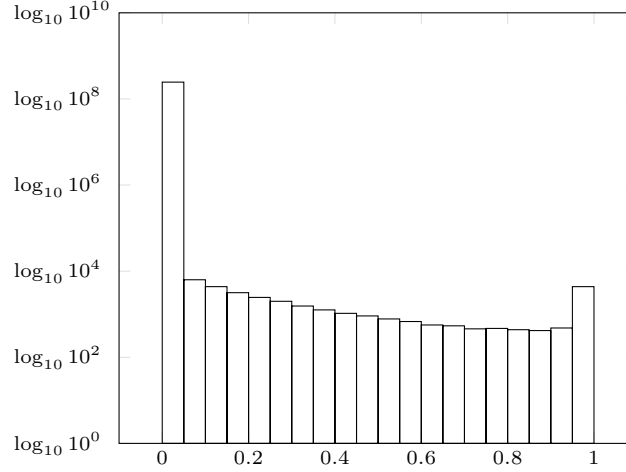


Fig. 2. Linkage probabilities in $\hat{\Delta}$ on a log (base 10) scale for the SHIW data sets. The bimodal distribution is representing the mixture between links and non links (and uncertain cases in between).

Detecting true positive: the influence of the probabilities threshold

Our estimate of the linkage is built by simulating $n_{sim} = 1000$ linkage matrices based on the parameter estimates of the StEM algorithm. From these sampled matrices we can compute the probability for each record pair to be linked.

Then, we need a threshold on the linkage probabilities to measure the performance of our methodology and decide which pairs should be linked. So far we used a sharp truncation of 0.5 on the linkage probabilities in order to compute the partial confusion matrix of our estimate. This strategy allows to maintain a one-to-one assignment constraint in the set of linked records, Tancredi and Liseo [2011, Theorem 4.1], Sadinle [2017, Corollary 1.1] and aligns with the literature methods, Tancredi and Liseo [2011], Steorts et al. [2016], Sadinle [2017].

As explained in the article, section 5.3, we assume the distribution of linkage probabilities to be bimodal, representing a mixture between linked pairs and non linked pairs, and uncertain cases in between, that proves valid according to fig. 2. Therefore, we may adapt the threshold and use a more restrictive value depending on this distribution. We show in table 2 the performance obtained for different threshold. If one would like to choose the threshold such that the estimated FDR would be lower than 10%, one would choose a threshold on the linkage probabilities between 0.6 and 0.7. In that case the FDR is underestimated and the performance obtained is similar to the one with a threshold at 0.5.

Performances and Distortion

We can characterise the level of distortion of the data using the sum of medians of disagreements and missing values among the PIVS of true links. The natural level of distortion of the SHIW is about 4%, which is higher than the NLTCs natural distortion level (0.2%).

To evaluate the robustness of *FlexRL* to the data quality we study the evolution of the performance metrics for screwed versions of the SHIW data sets. We artificially distort the data by changing and removing some values in the PIVs to create registration errors. We compare the performance of our method to the simplistic approach.

Both approaches have their performance decreasing with the increasing distortion, however *FlexRL* shows a more controlled decrease in performance and, the sensitivity level which was higher for the simplistic approach under a low distortion level gets to a similar level as *FlexRL* as the distortion increases.

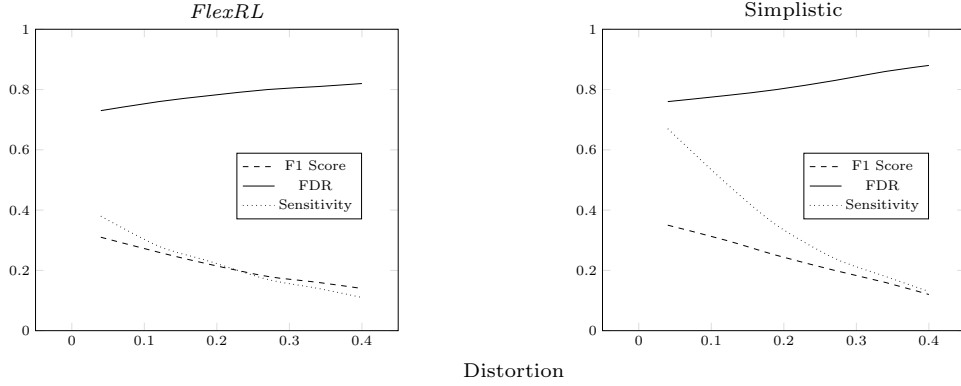


Fig. 3. Evolution of the F1 Score, False Discovery Rate (FDR) and Sensitivity for screwed versions of the complete SHIW data sets. We artificially distort the data and we define the distortion level (x-axis) as the sum of medians of disagreements and missing values among the PIVs of true links.

Application: RLData500

Description

The RLdata500 is a synthetic data set publicly available, created in the context of the ‘RecordLinkage’ R package, Sariyar and Borg [2010], and used to illustrate record linkage since then, Steorts [2015], Steorts et al. [2018], Enamorado and Steorts [2020], Omar et al. [2022], Bai et al. [2023], Sosa and Rodríguez [2024]. A bigger version with 10 000 records is also available and broadly used, Guha et al. [2022], Binette and Steorts [2022], Marchant et al. [2023]. The data consist of 500 records among which 10% are distorted duplicates and, an identifier is available to recover the true linkage structure. We separate the data into two sets \mathcal{A} and \mathcal{B} of 200 and 300 records with the 50 common units to apply our method and *BRL*; *Exchanger* can directly be applied on the data set. We use the five PIVs available to build links: first name, family name, birth day, month and year though this set of PIVs is very unlikely to be available in real life.

Our method handles categorical variables by mapping them to a set of integer values as mentioned in the article section 3.1. Since the names used to link the data are strings, this implies a loss of information, which we mitigate by using a soundex encoding, Russel [1918, 1922]; we describe the data in table 3.

		first name	family name	birth year	birth month	birth day
data	unique values	146	108	86	12	31
	encoded unique values	114	85			
	format	character	character	numeric	numeric	numeric
true links	% agreement	.68	.68	.84	.92	.80
	% agreement in encoded PIVs	.92	.88			

Table 3. Summary of the RLdata500 data. Characteristics of the PIVs in their original format and with soundex encoding for names. Level of agreement among the 50 links referring to the same individuals. No missing value in the dataset.

Performance comparison with the literature

We compare the baseline methods *BRL* and *Exchanger* with *FlexRL* and the simplistic approach applied on two versions of the data, the classical encoding required to launch the method (mapping values to a sequence of integers) and the soundex encoding applied on character type PIVs (first and family names). We provide performance measures of the compared methods in table 4.

	linked record		FN	F1 Score	Precision	Recall
	TP	FP				
<i>FlexRL</i> with soundex	44	0	6	.94	1	.88
<i>FlexRL</i>	38	0	12	.86	1	.76
<i>Exchanger</i>	49	0	1	.99	1	.98
<i>BRL</i>	50	1	0	.99	.98	1
simplistic approach with soundex	20	0	30	.57	1	.40
simplistic approach	0	0	50			

Table 4. Performance of the compared methods on the RLdata500 data with 50 links.

When unraveling the profile of linked or non linked data in terms of agreement level between values of record pairs, we notice that the lower performance of *FlexRL* is not due to the textual reduction of information induced by encoding the data as categorical. Indeed the *FN* are not caused by typographical errors that our method could not detect. The lower performance may be explained by the few number of true links, which makes it harder for our method to estimate the probability of mistakes. The simplistic approach does not link any pair without the soundex encoding because all the links present some registration errors. All methods perform better than the simplistic approach on these data.

References

- E. A. Bai, O. Binette, and J. P. Reiter. Optimal F-score clustering for bipartite record linkage. <https://arxiv.org/abs/2311.13923>, 2023.
- BancadItalia. Bank of Italy, Survey on Household Income and Wealth. <https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/distribuzione-microdati/index.html>. Banca d'Italia, Indagine sui bilanci delle famiglie italiane.
- O. Binette and R. C. Steorts. (almost) all of entity resolution. *Science Advances*, 8(12), 2022.
- T. Enamorado and R. C. Steorts. Probabilistic blocking and distributed Bayesian entity resolution. In *Privacy in Statistical Databases*, pages 224–239. Springer International Publishing, Springer International Publishing, 2020.
- S. Guha, J. P. Reiter, and A. Mercatanti. Bayesian causal inference with bipartite record linkage. *Bayesian Analysis*, 17(4): 1275–1299, 2022.
- N. G. Marchant, B. I. P. Rubinstein, and R. C. Steorts. Bayesian graphical entity resolution using exchangeable random partition priors. *Journal of Survey Statistics and Methodology*, 11(3):569–596, 2023.
- Z. A. Omar, M. A. Abu Bakar, Z. H. Zamzuri, and N. M. Ariff. Duplicate detection using unsupervised random forests: A preliminary analysis. In *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pages 66–71. Institute of Electrical and Electronics Engineers, 2022.
- T. Pacheco Menezes, T. Brendan Murphy, and M. Fop. Hausdorff distance-based record linkage for improved matching of households and individuals in different databases. <https://arxiv.org/pdf/2404.05566>, 2024.
- K. Robach. *FlexRL*. GitHub, 2024. URL <https://github.com/robachowyk/FlexRL>. Github package.
- R. C. Russel. Index. <https://patentimages.storage.googleapis.com/31/35/a1/f697a3ab85ced6/US1261167.pdf>, 1918.
- R. C. Russel. Index. <https://patentimages.storage.googleapis.com/82/e0/32/7b94720218b2d0/US1435663.pdf>, 1922.
- M. Sadinle. Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017.
- M. Sariyar and A. Borg. The recordlinkage package: Detecting errors in data. *The R Journal*, 2(2):61–67, 2010.
- J. Sosa and A. Rodríguez. A Bayesian approach for de-duplication in the presence of relational data. *Journal of Applied Statistics*, 51(2):197–215, 2024.
- R. C. Steorts. Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875, 2015.
- R. C. Steorts, R. Hall, and S. E. Fienberg. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.
- R. C. Steorts, A. Tancredi, and B. Liseo. Generalized Bayesian record linkage and regression with exact error propagation. In *Privacy in Statistical Databases*, pages 297–313. Springer International Publishing, Springer International Publishing, 2018.
- A. Tancredi and B. Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B), 2011.