WASHINGTON STATE DEPARTMENT OF HEALTH

Hyperblock

Unsupervised Hyperdimensional Fuzzy Blocking for Equitable Public Health Entity Resolution

Sean Coffinger, MA
Health Statistics Manager
Center for Health Statistics, DCHS
Washington State Department of Health

Washington State Department of
HEALTH

DOH 422-286 April 2025
To request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email doh.information@doh.wa.gov.

Table of Contents

Contents

NOTE/ACKNOWLEDGEMENTS: (Option to include information about report and/or names and titles of authors below TOC)

Introduction

Entity Resolution (ER) is a foundational task of data integration and aims to detect entities with different information that correspond to the same object [1]. In public health and associated research, we often aim to merge databases containing various records collected by multiple sources. This particular ER task is referred to here as “record linkage” and has the specific aim to match records from one source with another.

Each record linkage process typically comprises five distinct steps: data preprocessing, blocking (or indexing) and filtering, comparison, classification, and evaluation [2]. This study is entirely focused on optimizing the second procedure, blocking/indexing and filtering. After preprocessing, which standardizes and cleans input data, blocking/indexing aims to reduce the volume and/or search space for candidate comparison. Filtering is then applied to further restrict the number of pairwise comparisons necessary for comparison and classification [3]. These procedures are essential for optimizing linkage processes and building sustainable pipelines when dealing with big data [4]. Whether it be done by partitioning the data into several blocks [5] or sorting similar records based on custom criteria [6], the data must be segmented to allow all true matching records to be compared while minimizing the number of cumulative pairwise comparisons necessary to recall all true matching records.

Several effective non-learning, supervised, and unsupervised blocking methods have been previously proposed [2]. Standard blocking and iterative blocking form the basis for many implemented public health record linkage processes. These strategies assign blocking keys according to predetermined criteria based on the values in various features. The logic can be as simple as a single exact match between sources (i.e., exact date of birth match is the most frequently implemented blocking strategy in our state) or more complex using multiple layers of fuzzy logic or string distances to create multiple blocking keys. These basic methods can be tuned to streamline robust linkages with high recall; however, the necessary volume of inclusion required to achieve high recall remains very large. On the other hand, if the method is too strict, recall will be low for record pairs with higher rates of inexactness, which disproportionally impacts underrepresented subgroups [7].

A plethora of supervised methods have also been proposed: e.g., Token Blocking [8], Attribute Clustering Blocking [9], and Semantic Graph Blocking [10]. These methods are very effective and efficient; however, they require supervision (through training or direct intervention) and computational expertise. Highlighted by COVID-19, we have seen many public health jurisdictions relying on antiquated data systems, underfunded and depleted informatics workforces, and limited capacity to modernize [11]. An optimal solution for the public health sector must be simple, deployable, and generalizable. Unsupervised methods can enable near real-time blocking and can be generalizable, distributable, and sustainable. M. Kerjriwal and D.P. Miranker have developed a few robust unsupervised methods [12, 13], but here we propose a fundamentally different unsupervised approach that emphasizes equitable recall at the cost of some precision while being easily retrofitted into existing fuzzy-join procedures.

Furthermore, like many other industries, public health systems rely on multiple stacked systems with intermediate pipelines. Different jurisdictions, health entities, private operations, and public services all vary in the robustness and efficiency of these often-convoluted systems. Due to the variability observed across public health data systems and data flow, errors caused by human intervention can present themselves at multiple stages prior to data linkage. The most common means of comparing information for both blocking and linkage involves string comparisons. These string comparisons can be as simple as the number of shared letters/numbers or as complex as cosine vector similarity scores. Many linkages from start to finish use comparative values at the text level to account for the variability in a pair of true-linking records with some degree of inexactness. Other methods go beyond text-based methods, like Soundex, which has been used for years to capture similar-sounding identifiers that may not match lexically. This leap to sensory-based similarity metrics has roots across many disciplines but is lacking in entity resolution blocking techniques.

Here, we aim to contribute to the library of blocking and filtering methods for data linkage strategies by introducing a simple unsupervised blocking procedure, which we term *Hyperblock*. Here we demonstrate how Hyperblock can block and filter two data sources effectively when the rate of inexactness between the two databases is high. For this case study, we will link Washington state marriage events with Washington state death certificates. We evaluate this method against the most common established blocking procedures by comparing recall and total block sizes, while disaggregating by race/ethnicity to demonstrate the equity impact of robust public health blocking strategies. Our goal is to introduce Hyperblock as a simple, distributable, implementable solution that prioritizes equitable recall while minimizing volume compared to commonly practiced public health methods.

Methods

All computations were performed in R version 4.1.3 [14]. Running the linkage was a Microsoft Azure Virtual Machine (OS windows 10) with dual 18 core Intel® Xeon® Platinum CPUs and 144GB of physical RAM.

Washington state marriage and death certificates were queried from Washington Health and Life Events System (WHALES) database and initially matched by exact social security numbers available in each relevant table. Once joined, the resulting data frame represented known true-links between the two sources, with the assumption that exact social security number matches provide suitable criteria for ground truth testing. Only three identifier features were selected from each data source to use for downstream blocking: first name, last name, and date of birth (DOB). Additionally, race and ethnicity variables were extracted from the death database for downstream race/ethnicity disaggregation. All values were standardized: capitalized, punctuation removed, non-standard letters converted into standard Latin alphabet, date of births standardized and validated, and double values (e.g. double surnames) were concatenated. The final data set representing known true-links between the two sources totaled 51,401 links. Table 1 displays the links disaggregated by race/ethnicity designations (White/Non-Hispanic, Non-White and/or Hispanic, or Missing Race/Eth) as well as the proportion of inexactness present in true-link. To be considered an inexact true-link, one or more of the three identifier fields must not be exact matches. Individuals with any missing or invalid identifier fields were removed from the analysis and those with one or more missing race and/or ethnicity fields were counted as such.

Table 1 -

Race/Ethnicity	True-Links (n)	% Exact First Name	% Exact Last Name	% Exact DOB
White/Non-Hispanic	40,290	95.7%	70.0%	96.7%
Non-White and/or Hispanic	7,751	93.7%	69.0%	95.4%
Missing Race/Ethnicity	3,360	93.0%	70.8%	92.6%
Total	**51,401**	**95.25%**	**69.91%**	**96.26%**

After preprocessing, each name field from both marriage and death records were segmented into *q*-grams with a length of two, hereby denoted as *bigrams.* Each character field then populated a sparse binary string consisting of all bigram possibilities (26 characters x 26 characters = 676 bigram possibilities). If the name possessed one or more of a certain bigram, that corresponding bigram would be given a value of 1 in the string. If it did not possess a certain bigram, a value of 0 is designated. Date features were treated differently, with a string possessing of all possible day, month, century and year as various features (31 days + 12 months + 2 centuries + 100 years = 145 date features). DOBs were encoded in the same binary way as names depending on corresponding values; however, to account for month and day format switches present in many cultures (Month/Day/Year or Day/Month/Year format switches) and common transcription errors, neighboring days and valid switches were given a value of 0.5.

Two total manifolds were created: one containing text identifiers (First and Last Name) and one containing numeric values (DOB). To build the text identifier manifold, the encoded sparse name strings were subjected to two rounds of uniform manifold approximation and projection (UMAP). UMAP parameters were set at default according to the R package *umap* [15], except for the number of nearest neighbors, which was set at 15. The first UMAP reduced the 676-feature string down to 128 features, then the second UMAP consolidated the string even further down to 4 dimensions. The DOB manifold used a single iteration of UMAP, reducing the number of dimensions directly to 4 from 145. At this point, each identifier possesses a set of 4-dimensional coordinates (X, Y, Z, W). These manifolds were then collected at the record level, and each record possessed a total of 12 coordinates derived from one of the two manifolds.

Next, density-based spatial clustering of applications with noise (DBSCAN) was employed to cluster each 4-dimensional manifold. Each record was assigned a cluster derived from the R package *dbscan* [16] using default parameters, apart from *eps* being set at 0.15 and the minimum points being reduced to 2.

Cluster-inclusion then determined blocking, where records sharing two or more clusters were included as potential links to be compared by the classification algorithm downstream. The process described in the methods is diagramed in Figure 1.

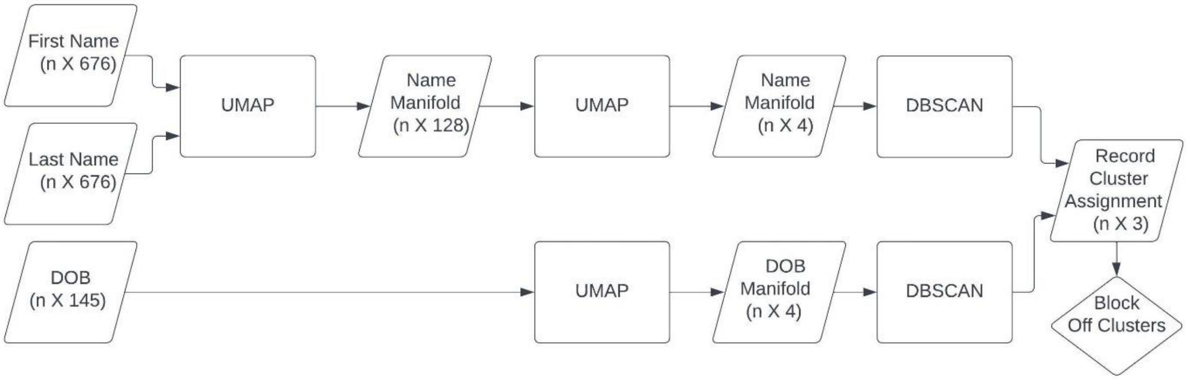

Figure 1 – Manifold dimensionality reduction toward cluster blocking

With a ground-truth established via social security numbers, we evaluate the number of true links captured by this blocking methodology, disaggregated by race and ethnicity. Additionally, we can evaluate the volume required to do so and compare it with various other methods commonly implemented at public health agencies. Here, we compare Hyperblock to the following:

1. No Blocking: The fully inclusive model assumes no blocking step and includes all possible pairwise comparisons
2. Exact-DOB: Exact-DOB requires DOB to match exactly
3. Fuzzy-DOB: fuzzy-DOB methods allow 1 or 2 characters of disagreement measured by Hamming distance
4. Full-fuzzy methods: defined as the fuzzy agreement between two or more identifier variables measured by Levenshtein/Hamming distance less than or equal to 2. We compare inclusion criteria of 1/3 and 2/3, identifying feature matches in Table 3.
    a. For example, to be included in "Full Fuzzy (Levenstein/Hamming ≤ 2, 1+ Feature Matches), one of the identifiers (first name, last name, DOB) must possess a Levenstein/Hamming distance less than or equal to 2.

To illustrate this clearly, Table 2 displays the criteria and feature comparisons of the strategies compared in Table 3.

Table 2 -

Method	*Inclusion Requirements*	*Feature Comparisons*		
		DOB	**First Name**	**Last Name**
No Blocking	All pairwise comparissons included	None	None	None
Exact Match DOB	Exact match of DOB	Exact Match	None	None
Fuzzy DOB *(Hamming ≤ 1)*	Fuzzy match of DOB	Fuzzy Match of Hamming ≤ *1*	None	None
Fuzzy DOB *(Hamming ≤ 2)*	Fuzzy match of DOB	Fuzzy Match of Hamming ≤ *2*	None	None
Full Fuzzy *(Levenstein/Hamming ≤ 2 , 2+ Feature Matches)*	Any two features fuzzy match	Fuzzy Match of Hamming ≤ *2*	Fuzzy Match of Levenstein ≤ *2*	Fuzzy Match of Levenstein ≤ *2*
Full Fuzzy *(Levenstein/Hamming ≤ 2 , 1+ Feature Matches)*	Any one feature fuzzy match	Fuzzy Match of Hamming ≤ *2*	Fuzzy Match of Levenstein ≤ *2*	Fuzzy Match of Levenstein ≤ *2*
Hyperblock *(2+ Feature Matches)*	Any two feature hyperblock match	Cluster Match	Cluster Match	Cluster Match

Here, we exclude advanced methodologies outlined in the introduction due to our research objective, target audience and results presented below.

Results

Hyperblock identified 1810 name clusters and 667 DOB clusters.

Table 3 -

	Block Size (Volume)	*True-Links Included in Blocks (% Recall)*		
Method	**Total Links to be Compared**	**White/Non-Hispanic Links**	**Non-White and/ or Hispanic Links**	**Missing Race/Eth Links**
No Blocking	2,642,062,801	100%	100%	100%
Exact Match DOB	167,998	96.7%	95.4%	92.6%
Fuzzy DOB *(Hamming ≤ 1)*	3,801,669	99.1%	98.6%	96.5%
Fuzzy DOB *(Hamming ≤ 2)*	48,278,657	99.5%	99.2%	97.6%
Full Fuzzy *(Levenstein/Hamming ≤ 2 , 2+ Feature Matches)*	86,983	98.9%	97.9%	96.5%
Full Fuzzy *(Levenstein/Hamming ≤ 2 , 1+ Feature Matches)*	13,442,853	100%	100%	100%
Hyperblock *(2+ Feature Matches)*	18,415,992	100%	100%	99.9%

Hyperblock outperformed many of the rudimentary blocking strategies, including all DOB-based blocking logic strategies. However, performance was comparable to "Full Fuzzy" strategies, and was outperformed by the single feature match strategy. Performance was gauged by the equitable inclusion of links across race/ethnicity disaggregation and the minimization of overall volume.

Discussion

Hyperblock sought to provide a more generalizable, unsupervised blocking approach that aimed to remove any possibility of practitioner bias. We hypothesized that this strategy would more equitably capture links in minoritized subpopulations while maintaining a low volume. The strategy was able to achieve this aim; however, it did not perform as well as more basic strategies in both of these objectives.

Additionally, Hyperblock aimed to be a simple implementation of hyperdimensional blocking for a simple comparison task. We believe we achieved that objective, but the strategy remains far more complicated to implement compared to Full Fuzzy methods, which once again outperformed Hyperblock.

Improvements in encoding strategies and more nuanced tuning of hyperparameters could benefit Hyperblock. Additionally, larger and more diverse datasets may highlight some benefits. However, the results of this investigation are clear: sometimes the simplest methods work. Efforts in improving public health ER tasks should shift towards prioritizing optimization. For example, if better coding and basic computational practices can enable larger volumes of comparisons to be performed independently of hardware requirements, blocking will be less constraining in the linkage pipeline. Our group aims to collaborate with the Data Science and Engineering unit at CHS to pursue optimization of machine learning linkage strategies while ensuring the high-quality products we produce remain.

References

1. V. Christophides, V. Efthymiou, and K. Stefanidis. Entity Resolution in the Web of Data. Morgan & Claypool Publishers, 2015.
2. Christen, P.: Data Matching. Springer (2012).
3. Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys (CSUR)*, *53*(2), 1-42.
4. X. L. Dong and D. Srivastava. Big Data Integration. Morgan & Claypool Publishers, 2015
5. Fellegi, I., Sunter, A.: A theory for record linkage. Journal of the American Statistical Association 64(328) (1969)
6. Hernandez, M.A., Stolfo, S.J.: The merge/purge problem for large databases. In: ACM SIGMOD, San Jose (1995)
7. Coffinger, S., Rothbard, S., Wu, C., Cox, A., Cook, M., & Hutchinson, K. (2023). Prioritizing Equitable Representation, Sustainability, and Accuracy: The Deployment of Machine Learning Linkage Strategies During the COVID-19 Pandemic. Washington Department of Health.
8. G. Papadakis, E. Ioannou, C. Niederée, and P. Fankhauser. Efficient entity resolution for large heterogeneous information spaces. In WSDM, pages 535–544, 2011.
9. G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl. A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE TKDE, 25(12):2665–2682, 2013
10. J. Nin, V. Muntés-Mulero, N. Martínez-Bazan, and J. Larriba-Pey. On the use of semantic blocking techniques for data cleansing and integration. In IDEAS, pages 190–198, 2007.
11. Kadakia KT, Howell MD, DeSalvo KB. Modernizing Public Health Data Systems: Lessons From the Health Information Technology for Economic and Clinical Health (HITECH) Act . *JAMA.* 2021;326(5):385–386. doi:10.1001/jama.2021.12000
12. M. Kejriwal and D. P. Miranker. A two-step blocking scheme learner for scalable link discovery. In OM Workshop, pages 49–60, 2014
13. M. Kejriwal and D. P. Miranker. A DNF blocking scheme learner for heterogeneous datasets. CoRR, abs/1501.01694, 2015.
14. R Core Team (2022). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL https://R-project.org/
15. Tomasz Konopka (2020). umap: Uniform Manifold Approximation and Projection. R package version 0.2.7.0. https://CRAN.R-project.org/package=umap
16. Hahsler M, Piekenbrock M, Doran D (2019). "dbscan: Fast Density-Based Clustering with R." _Journal of Statistical Software_, *91*(1), 1-30. doi: 10.18637/jss.v091.i01 (URL: https://doi.org/10.18637/jss.v091.i01).

DOH 422-286 April 2025