A Comparison of Rule-based and Supervised Machine Learning Approaches for Record Linkage of Italian Historical Data

By Saverio Minardi, Suzanne Greco and Nicola Barban

To cite this article: Minardi, S., Greco, S., & Barban, N. (2025). A Comparison of Rule-based and Supervised Machine Learning Approaches for Record Linkage of Italian Historical Data. *Historical Life Course Studies, 15*, 28–46. https://doi.org/10.51964/hlcs18990

HISTORICAL LIFE COURSE STUDIES

VOLUME 15

2025

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies was established within *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation, the International Institute of Social History, the European Society of Historical Demography, Radboud University Press, Lund University and HiDO Scientific Research Network Historical Demography. Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

HISTORICAL LIFE COURSE STUDIES
VOLUME 15 (2025), 28–46, published 03-06-2025

A Comparison of Rule-based and Supervised Machine Learning Approaches for Record Linkage of Italian Historical Data

Saverio Minardi
University of Bologna

Suzanne Greco
ItalianParishRecords.org, USA

Nicola Barban
University of Bologna

ABSTRACT

Parish and civil records are crucial sources for reconstructing historical socio-demographic processes. However, their analysis presents significant challenges, particularly the need to digitize data and link life events across documents that lack formal identifiers. With the growing availability of digitized records, the development and evaluation of automated linkage techniques have become increasingly important. This study compares rule-based and supervised machine learning approaches for linking birth and death records derived from crowdsourced transcriptions of Italian parish and civil registers. Using a set of hand-linked data as a benchmark, we assess the performance of both approaches in terms of precision and recall, under standard conditions and in scenarios where key disambiguating information is missing. Our findings suggest that the machine learning approach outperforms the rule-based method both under standard conditions and when information is incomplete, making it the preferred option when training data are available. Nonetheless, the rule-based method can still achieve high precision when configured with sufficiently strict matching thresholds. While the focus of this exercise is on linking birth and death records, the procedures can be adapted to a wide range of historical reconstruction projects based on names and dates.

Keywords: Record linkage, Parish records, Historical demography

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs18990

1 INTRODUCTION

Parish and civil records have long been essential sources in historical demography, providing invaluable insights into periods when centralized data collection was not common practice (e.g., Rettaroli & Scalone, 2012; Scalone & Samoggia, 2018). These documents provide detailed information about individuals residing in specific geographically defined communities such as parishes or municipalities.

However, several challenges make the study of parish data and civil registers both time-consuming and complex. First, it requires collecting preserved records from scattered locations and transcribing them into digital form. This transcription process can be difficult, as the sources are not always stored in centralized archives, and original manuscripts are often poorly preserved or difficult to read. Second, information on individuals' life courses is recorded in separate registers — such as baptisms, funerals, and marriages — without formal identification numbers. As a result, record linkage based on the available information is essential for reconstructing individual biographies and family relationships (Del Panta & Rettaroli, 1994).

In the earliest approaches, manuscripts were linked through meticulous manual searches. However, as the volume of available records grew, such approaches became unfeasible and, starting from the early 70s, several automated or semi-automated computer applications were developed across a wide range of projects and experiences (Winchester 1992). More recently, the linkage of U.S. censuses led to the proliferation and comparison of more efficient methods (Abramitzky et al., 2021; Feigenbaum, 2016; Fu et al., 2014; Helgertz, 2022; Ruggles et al., 2018).

Despite advances in record linkage, no single formal technique has been established as a standard, and it is unlikely that any one method will be suitable for all situations, populations, or datasets. As a result, linkage techniques must be adapted to the specific characteristics of the data, creating a continuous need to test and compare different approaches across diverse contexts (Abramitzky et al., 2021; Avoundjian et al., 2020; Wen et al., 2022). Moreover, the growing availability of digitized records — driven by recent advancements in handwritten text recognition and the expansion of crowdsourced platforms that compile information from historical registers (e.g., Kahle et al., 2017; Pujadas-Mora et al., 2022) — has made the development and evaluation of automated, efficient, and broadly applicable record linkage techniques increasingly important.

The present article contributes to the growing literature on historical record linkage by applying and comparing two easily implementable and automated procedures for matching birth and death records from crowdsourced transcriptions of Italian parish and civil records dating from the 16th century onwards. We draw on a novel dataset transcribed by volunteers on the website ItalianParishRecords.org, a U.S.-based organization that digitizes and indexes Italian parish and civil records.

We compare a rule-based string similarity approach — in which records are linked based on fixed thresholds applied to string similarity scores across multiple fields (e.g., Christen, 2012, p. 139) — with a supervised machine learning approach, where a model is trained on human-labelled data to learn how to combine various similarity measures into predicted match probabilities (e.g., Feigenbaum, 2016). The rule-based method requires researchers to set field-specific thresholds; if the similarity between two records exceeds all thresholds, the records are deterministically linked. This approach does not require a set of hand-labelled training data, and and thresholds can be defined based on researcher judgment. In contrast, the machine learning approach learns to combine similarities across fields into a single predicted match probability and automatically determines optimal thresholds for linkage based on labelled training data that approximate true matches.

Both approaches have received particular attention in recent literature on census linkage (e.g., Abramitzky et al., 2021; Wen et al., 2022), however, parish and civil records present several unique characteristics that make the straightforward application of techniques developed for other data sources potentially less effective. First, linkage in these registers relies almost exclusively on string-based information, such as the names of individuals and their relatives. Second, since the registers contain no time-invariant characteristics other than names, only a limited number of blocking criteria can be applied. Third, due to high child mortality and the common practice of reusing the names of deceased children (Herlihy, 1988), the data often include multiple records that are identical in both given names and parental names.

Given these conditions, we evaluate the performance of the two approaches in terms of recall — the number of correct matches identified among all true matches — and precision — the number of

correct matches among those identified. Additionally, the article assesses their performance under varying data quality scenarios by removing information on one or both parents.

Results suggest that both procedures are valid options, but a machine learning approach is preferable when the researcher can create a training dataset as it allows higher levels of both recall and precision. Nevertheless, results highlight that even a rule-based string similarity approach with sufficiently high thresholds can return high precision and be a valid option for applied research.

Despite the specific application to birth and death registers, the main features considered for this linkage — names, last names, location, dates, and parents' names — are the same information in other registers such as marriages or status animarum. These approaches can therefore be easily extendable to the matching of other registers and eventually to family reconstitutions.

The remainder of this paper begins with a review of current approaches to automated record linkage of historical data. It then introduces the data used in this linkage project and describes the creation of the manually matched subsample. The following sections outline the logic and steps underlying the two linkage procedures. Next, the results section compares the performance of the two approaches against the manually matched data, and the concluding section discusses the findings and their implications for researchers seeking to implement similar approaches on comparable historical data sources.

2 CURRENT APPROACHES TO RECORD LINKAGE

Record linkage integrates information from multiple sources to identify all records related to a specific individual. It is essential for historical analysis, as data on a person's life are typically distributed across various documents — such as baptism or funeral registers — without a unique identifier. To reconstruct individual life histories, researchers must use techniques that link records from different sources to the same person.

In theory, identifying individuals across census periods may seem straightforward, as people are expected to retain certain immutable characteristics such as names and birth dates. In practice, however, the process is complicated by proxy reporting, non-standardized spelling (including abbreviations, nicknames, suffixes, and prefixes) and spelling errors, enumeration mistakes, transcription and registration errors, the prevalence of common names, and missing information. These challenges make it difficult to determine with certainty whether two records refer to the same person.

The earliest solution to this problem was a manual reconstruction technique developed by Louis Henry and Michel Fleury in the 1950s for small French parishes. This method relied heavily on the researcher's personal judgment and experience. However, its application to larger populations is problematic: it is time-consuming, prone to inconsistencies in matching decisions, and yields linked datasets that are difficult to reproduce.

As the volume and geographical coverage of available data expanded, manual reconstruction became increasingly impractical. In response, various projects developed fully automated (Winchester, 1992) or semi-automated (Fure, 2000) record linkage techniques.[1] An example of the latter is the work by Breschi et al. (2020), who employed a semi-automatic method to reconstruct individual and family histories in the municipality of Casalguidi, Tuscany. Their approach begins with the automatic linkage of exact matches and is followed by a semi-automated nominative linkage process, where the researcher selects the correct match from a subset of candidates defined by less restrictive criteria.

1 Various large projects have applied semi-automated or fully automated record-linking procedures and family reconstructions since the 1970s. Examples are the PRDH at the University of Montreal (Dillon et al., 2018), the IREP (former SOREP) at the University of Quebec at Chicoutimi (Bouchard et al., 1986), the English Cambridge Group for the History of Population and Social Structure (Wrigley et al., 1997), the Scanian Economic Demographic Database (Dribe & Quaranta, 2020), POPLINK from Sweden (Westberg et al, 2016), BALSAC from Canada (Vézina & Bournival, 2020), the Historical Sample of the Netherlands (Mandemakers, 2002), The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database (Pujadas-Mora et al., 2022).

Automated historical record linkage aims to create a fully routinized procedure for performing this task. However, data errors, limited information on individuals, and the variability of historical records make it challenging to establish a standardized linkage method that can accommodate every possible matching scenario. Automated linkage procedures must balance two main objectives. First, they should capture as many true matches as possible, minimizing Type II errors — cases where records referring to the same individual are not linked. Second, they should avoid false matches, reducing Type I errors — instances where unrelated records are incorrectly linked. However, there is a trade-off between the type I and II errors. Meaning that within a given method achieving a higher linkage rate will tend to come at the expense of greater linking error, whereas low rates of false positive matches will result in lower match rates. This creates a conundrum for researchers because expanding the universe of potential matches increases the likelihood of finding a true match but also increases the risk of false positives.

Recently, several projects aimed at linking different waves of U.S. censuses have generated a substantial body of literature comparing and evaluating various linkage procedures and their properties (Abramitzky et al., 2021; Bailey et al., 2020; Ruggles et al., 2018). The most straightforward methods are rule-based approaches, which rely on researcher-defined criteria to deterministically decide whether two records refer to the same individual. These rules must strike a balance: they need to be flexible enough to account for minor discrepancies caused by reporting or recording errors (such as spelling variations or typographical mistakes), yet strict enough to avoid mistakenly linking different individuals with similar characteristics.

In the case of parish registers, most of the information is found in the names and surnames of the probands and their parents. As a result, name misspellings represent the most critical challenge. Rule-based approaches have addressed this issue in several effective ways, typically by standardizing name spellings or by quantifying the differences between the spellings of two strings.

Ferrie's (1996) approach to link men in the 1850 and 1860 U.S. Census is an early example. Ferrie uses a sample of uncommon names from the 1850 Census, standardises them using NYSIIS[2] codes, and truncates the untransformed names after the fourth letter. He then links his sample to the 1860 Census and eliminates ambiguous candidates according to fixed rules.

Subsequent development of Ferrie's procedure uses Jaro-Winkler scores instead of name standardisations to quantify the dissimilarity of two names (Abramitzky et al., 2019). The Jaro-Winkler string distance calculates the similarity of two strings. The metric is based on the "edit distance" between two strings (the number of modifications required to turn one string into another). Any two records with a string distance above a defined cut-off can subsequently be considered a match.

A complementary approach to rule-based methods is the fully automated probabilistic procedure introduced by Abramitzky et al. (2019), which builds on standard techniques from the statistical literature. Rather than relying on hand-linked training data, this method uses the Expectation-Maximization (EM) algorithm to estimate the probability that two records refer to the same individual, based on features such as Jaro-Winkler name similarity scores and age differences. These probabilities can then be used to construct linked samples according to researcher-specified thresholds that balance false positives and false negatives.

Another widely used procedure is based on supervised machine-learning approaches. For instance, the IPUMS Linked Representative Samples (IPUMS-LRS) used the same variables as Ferrie to estimate the probability that two records refer to the same individual. Using a set of manually verified matches and non-matches as training data, Support Vector Machines (SVMs) were employed to classify potential record links. The model learned from the features of known links — such as name similarity and age difference — and applied this knowledge to evaluate new pairs (Goeken et al., 2011). More recently, Feigenbaum (2016) used a supervised learning model, specifically a probit regression, to link the 1915 Iowa State Census to the 1940 U.S. Census. His model, trained on verified matches, used features like Jaro-Winkler name similarity scores and age differences to predict whether a candidate pair constituted a true match. A match was declared only when its predicted probability exceeded a fixed threshold and was sufficiently greater than any competing alternative. Supervised machine learning approaches have been progressively refined using more features to estimate linkage probability (Helgertz et al., 2022), and recent work by Price et al. (2021) further improves performance by leveraging high-quality training data derived from user-verified genealogical links.

2 The New York State Identification and Intelligence System is a phonetic algorithm that standardizes names based on their pronunciation, allowing matches even when spellings differ slightly.

Building on this line of research, Feigenbaum et al. (2025) provide a detailed assessment of how different approaches to constructing training data influence the performance of supervised record linkage algorithms. Using hand-linked and genealogically sourced training data to link men from the 1900 U.S. Census to later censuses, they examine the trade-offs between data quality, time investment, and algorithmic performance. While they find that richer contextual information can improve the quality of hand-linked training data, they also show that supervised algorithms remain relatively robust even when trained on noisier data.

Gautam et al. (2020) propose a novel approach to record linkage by leveraging Evolution Knowledge Graphs (EKGs) to model temporal changes in individual attributes. Their Weighted Embedding-based Record Linkage (WERL) method optimizes attribute importance for matching records, improving linkage accuracy across historical, medical, and academic datasets. Compared to traditional methods, their approach enhances flexibility and robustness, particularly in handling missing data and evolving attributes.

Recent comparative research does not point towards the superiority of one single linking method, as the best procedure also relies on the characteristics of the data, the research question, and the historical context.

3 DATA

Parish records have long been a vital source of information for studying socio-demographic processes in periods preceding the centralized collection of population registers. They have played a key role in advancing our understanding of numerous demographic phenomena, including infant mortality (Fornasin et al., 2016; Piccione et al., 2014; Scalone et al., 2017; Tymicki, 2009), marriage patterns (Dribe & Lundh, 2010; Ruiu & Breschi, 2015), migration (Breschi et al., 2011; Manfredini, 2003), the demographic dynamics of isolated populations (Rettaroli et al., 2019), and the onset and development of the demographic transition (Breschi et al., 2009; Breschi et al., 2014; Minello et al., 2017; Rettaroli et al., 2017). Most existing studies have focused on small communities, where parish records were carefully transcribed and linked by trained researchers, ensuring high data quality and meticulous family reconstructions.

In this project, we experiment by linking several Italian birth and death parish and civil records transcribed by volunteers on ItalianParishRecords.org[3]. The website collects scans and links to sources of Italian parish and civil records from 1500 onward. A considerable amount of these registers has been transcribed to spreadsheets by volunteers, therefore providing a large amount of analysable data over a long period and across multiple locations. Unlike previous studies, these data result from a collective effort from volunteers passionate about local history or attempting to reconstruct their family history and are not collected with the attempt to document socio-demographic processes. As a result, the data cover a significantly higher number of parishes all over Italy, but each parish's data quality and time coverage are, on average, lower. Nevertheless, the main individual information recorded is the same as any other parish collection making their linkage a valuable case study for fully automated record linkage with large samples. Moreover, as the project progresses and an increasing number of records are transcribed, the dataset becomes an increasingly valuable and unique resource for historical demography.

It is important to acknowledge that, since precise reconstruction was not the primary goal, the inputs were neither systematically double-checked nor consistently collected across time, locations, or registers. As a result, the data present higher levels of imprecision — such as spelling mistakes, missing records, and incomplete registers — that are typical of historical sources, making them an exemplary test case for the limitations of family reconstruction.

Despite these issues, the data are a unique source and possess characteristics that are valuable from a research standpoint. Crowdsourced transcriptions are an emerging option for digitization; as the number of records increases, they are likely to become an increasingly useful resource. Since imperfections in transcription are an inherent feature of such datasets, it is important to consider linking and analytical procedures that account for these limitations. Moreover, given their large size,

3 The data used for this exercise were downloaded from ItalianParishRecords.org February 1, 2022.

these data offer a unique opportunity to experiment with automated linkage algorithms specifically designed for parish and civil registers. As newer and more accurate digitization projects emerge — for example, those using optical character recognition — experimentation on this dataset can serve as a valuable starting point. The procedures developed are likely to become more precise and efficient as the quality of digitization improves.

The available records contained information in unstructured strings on the first and last names of the reference individual; date of the record; parents' names and last names; sex (inferred from the name), and parish name. Death registers further contained the age at death in unstructured strings for approximately 50% of the sample in death registers. The total number of birth records was 744,432 (1479–1910), and deaths were 426,344 (1524–2020). Most records are from southern Italy, especially the region of Calabria, during the 19th century.[4]

All information was contained in unstructured strings; therefore, we performed extensive data cleaning and processing before matching the death and birth registers. First, the year of birth and death was extracted from dates using any appearance of four consecutive digits. Parish names were cleaned of non-alphabetical characters and converted to contemporary municipalities. Individuals' names, last names, and parents' names were cleaned of all non-alphabetical characters and indications of death for parents "Fu" or "Dec" or indication of unknown (e.g., Ignoto, incognito, unknown). Sex was manually assigned to a list of first names.

4 HAND-LINKED SAMPLE FOR TRAINING AND TESTING

A subset of manually verified matches was created for two uses. First, to train the machine learning algorithm and, second, to test the quality of the two procedures. This subset data were built starting from 1,000 random observations from the birth records. These birth records were then connected to their potential matches in the death registers.

Given a subset of birth records X1 and all death records X2, the first step was to extract records in X2 that are plausible matches to records in X1. Variables that define a possible match are called blocking variables. In an extreme case, with no possible blocking variables, the set of possible matches is defined by the entire Cartesian products of X1 and X2 and has X1 * X2 observations.

Since very little information is available, only two blocking criteria were initially applied to creating the training data.

First, for a death record to be a potential match of a birth record, it must be recorded between 0 and 120 years from the birth event. This reflected a minimum age at death of 0 (i.e., people cannot die before birth) and a maximum age of 120 years. Second, the Jaro-Winkler score between the first and last names of the birth and death records had to be above 0.8. Gender is not considered a blocking variable because it is inferred through individual first names, and simple misspells — such as changing an "a" to an "o" — could lead to the assignation of the wrong sex.

We chose not to include birth year — obtained for the reported age at death in death records — as a blocking criterion for several reasons. First, age information in death records is frequently missing (approximately 50% of downloaded records) or described only vaguely (e.g., "child" or "infant"), necessitating an age-independent linkage approach. Second, even when age is reported, significant age heaping can compromise accuracy, particularly when distinguishing individuals with identical names and similar birth years within the same family. Third, other historical sources may lack age information altogether, making an age-independent approach more broadly applicable. However, when reliable age at death or birth year information is available, researchers can easily adapt the approach by incorporating them in the blocking criteria.

At least one potential match was identified for 891 birth records resulting in a final training sample of 41,684 possible combinations.

4 See Table A1 in the Appendix for the distribution of records across regions and Figure A1 of the Appendix for distribution across years.

A human researcher carefully inspected training data, and correct matches were determined. Manual inspection of the training data allowed the establishment of two further blocking criteria. First, birth and death records needed to be from the same municipality. This meant that migrants were not part of the sample, which is a common limitation in the analysis of parish records. Second, records with no comparable information on at least one parent were excluded because it was impossible to disambiguate individuals with the same name and last name without at least one parent.

We explored the feasibility of linkage without information on either parent, finding that the absence of parental names typically resulted in deterministic linkage failure in the training data. Given the frequent reuse of first names within large families in the Italian historical context, the presence of at least one parent's name is a necessary condition for reliably assigning a link. After applying these extra blocking criteria, the training data was reduced to 4,359 dyads for 418 birth records.

A recurring issue was the disambiguation of records with identical or very similar names, last names, and parents' names but different birth or death dates. These exact multiple matches are due to the practices of renaming newborn children with the name already used for previously deceased siblings. In these cases, we introduce a standardised procedure for both the manual validation of the training data and the automated record linkages. When one birth can be matched to multiple deaths, the earliest death record is considered a match. When multiple births are assigned to one death, the latest birth is considered the correct match. These choices avoid the overlapping in the same time-point of two individuals with identical names in the same household. Table 1 helps to clarify this point. It shows multiple possible matches to a single birth record. Four death records are identical in all characteristics except the year of death, meaning that parents reassigned the same name four times after the child's death.

If we were to consider a correct match a death record dead later than the first one, it would result in individuals with the same name living in the same household at the same moment. For instance, if we considered the death in 1875 as the correct match, that would mean that, at least in 1874, two Vita Pisciotta existed in the same place with the same parents, which is a highly unlikely occurrence. The only match that avoids an impossible overlapping is the earliest one. When two records are both possible matches but impossible to disambiguate, as in the case of duplicates, they are both considered not matched. Eventually, 149 correct matches were identified. Descriptive statistics for the training data are reported in Table A2 of the Appendix.

Table 1 *Example of multiple possible matches in the death records to one birth record*

Birth					**Death**				
Name	**Surname**	**Father**	**Mother**	**Year**	**Name**	**Surname**	**Father**	**Mother**	**Year**
Vita	Pisciotta	Francesco Pisciotta	La Rocca Maria	1872	Vita	Pisciotta	Francesco Pisciotta	La Rocca Maria	1874
					Vita	Pisciotta	Francesco Pisciotta	La Rocca Maria	1875
					Vita	Pisciotta	Francesco Pisciotta	Larocca Maria	1877
					Vita	Pisciotta	Francesco Pisciotta	Larocca Maria	1884

5 SUPERVISED MACHINE-LEARNING APPROACH

Our machine learning approach follows the general strategy established by Feigenbaum (2016) and Helgertz et al. (2022). This approach involves training a model on a set of manually verified matches and then using the trained algorithm to predict correct matches across the entire dataset. Both Feigenbaum (2016) and Helgertz et al. (2022) utilized generalized linear models (GLMs), such as logit or probit, demonstrating their effectiveness in census linkage contexts and, in many cases, outperforming other prediction methods. Consistent with this literature, we initially adopted a logistic regression model trained on researcher-verified matches. However, given our data's distinct characteristics — primarily consisting of many string-similarity measures and minimal categorical information — we also implemented a random forest model. This method is advantageous because it can capture complex, non-linear relationships among predictors, potentially significantly improving linkage accuracy in our context.

The models leverage several variables pertaining to the dyad of potential links,[5] and based on those, predict a probability that the dyad is a match.

A dyad of records is then considered a possible match if it satisfies two conditions. First, the match must have the highest predicted probability for a given birth record and exceed a threshold value (b_1), with probabilities estimated using either logistic regression or a random forest model.

Second, for a dyad $X1_i * X2_i$, the ratio of the highest probability among any possible matches to $X1_i$ to the probability of $X1_i * X2_i$ is below a threshold (b_2), meaning that a record needs to have a predicted probability sufficiently higher than other possible matches to be considered the only possible match. If two or more matches for the same records have a predicted probability above b_1 and a ratio to the best predicted probability below b_2, they are all considered possible matches.

Table 2 reports a sketched example of these parameters and thresholds without parents' information for brevity. The maximum probability for $X1_1$ is associated with $X2_1$ (.97), but the second possible match has a sufficiently high predicted probability of .80, resulting in a ratio of 1.21. Considering a threshold (b_1) for the predicted probability equal to .79 (or any other level below .8) and a threshold (b_2) for the ratio to the highest probability equal to 1.3, both $X2_1$ and $X2_2$ are considered possible matches. If only one match results from this procedure (for instance setting b_2 to 1.1) that is considered a correct match. In case of multiple matches, the chronological disambiguation procedure explained above is applied (See Table 1).

Unlike Feigenbaum (2016), who resolves ambiguous matches by rejecting high-probability links that are too similar to the second-best candidate, our method retains all high-probability links above a set threshold. We then resolve ambiguity by applying temporal logic, discarding only those matches that imply impossible overlapping lifespans. This approach is particularly suitable for contexts like ours, where frequent reuse of names within families means multiple high-probability links often genuinely exist, and probability-based exclusion alone would incorrectly discard valid matches.

In order to determine the values of the thresholds b_1 and b_2, we employ cross-validation on the training dataset. During this process, we search through the space of possible values of b_1 and b_2 to find the optimal combination that simultaneously maximises recall and precision.

Table 2 *Example of thresholds for machine-learning approach*

Births (X1)				Deaths (X2)				Parameters	
Id_b	Name	Surname	year	Id_d	Name_d	surname_d	year_d	Probability	Ratio
1	Barbara	Pellegrino	1767	1	Barbara	Pellegrino	1768	.97	(.97/.97) 1.00
				2	Barbara	Pelegrino	1837	.80	(.97/.80) 1.21
				3	Barbara	Petriello	1814	.40	(.97/.40) 2.42

5 Full results for the logistic model on the training data are reported in Table A3 in the Appendix.

Recall is measured as the true positive rate (TPR):

$$TPR = \frac{True\ \ positives}{True\ \ positives\ +\ False\ \ negatives}$$

This measure records the ratio of true positives over the total number of positives. A high TPR indicates that a high share of the real matches is found and matched. If TPR equals one, all correct matches have been found.

Precision is measured through the positive predictive value (PPV):

$$PPV = \frac{True\ \ positives}{True\ \ positives\ +\ False\ \ positives}$$

PPV indicates how many of all the matches assigned by the algorithm are true matches. A PPV equal to one indicates that all matches are correct matches. Ideally, an algorithm would find all possible matches (TPR = 1) and only the correct ones (PPV = 1).

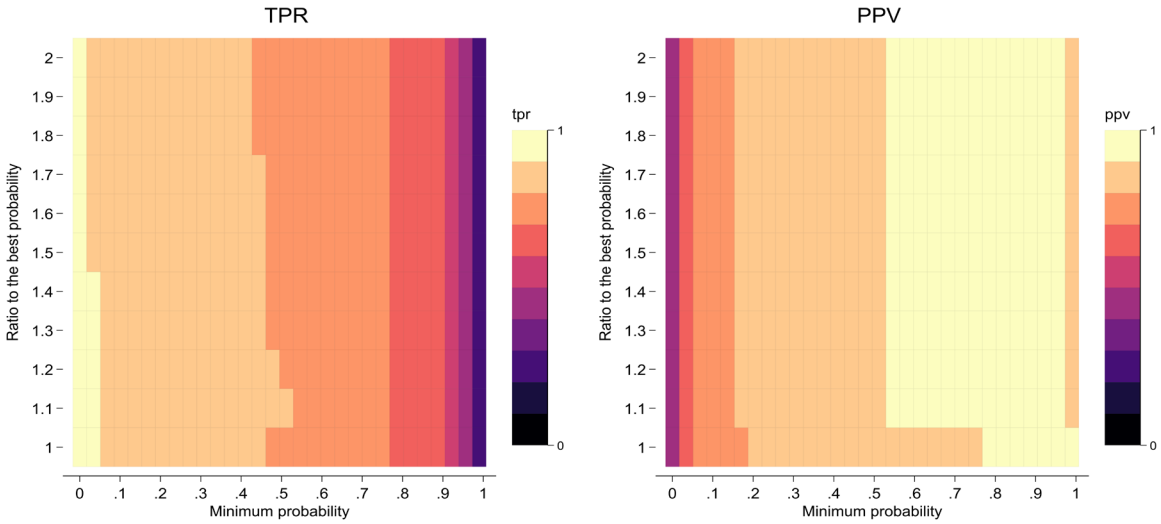
However, the two measures are negatively correlated. The more matches the algorithm finds, the less restrictive it is and hence less precise. Tightening matching criteria reduces false positives but also decreases the number of correct matches identified.

We therefore need to identify the levels of b_1 and b_2 that yield the optimal combination of PPV and TPR. Several approaches can be used to select these optimal parameters, typically depending on the specific objectives and preferences of researchers. Depending on their data and goals, researchers might prioritize higher precision (PPV) at the cost of lower recall (TPR), or vice versa. For instance, Feigenbaum (2016) initially selects thresholds by maximizing the sum of TPR and PPV (giving equal weight to each) and subsequently experiments with alternative weighting schemes, such as maximizing 2 × TPR + PPV. Helgertz et al. (2022), in contrast, select parameters by maximizing the Matthews Correlation Coefficient. Since our goal is to compare methods, we do not present a single best combination but instead graphically illustrate all optimal combinations achievable with each method.

In order to find these values, one-half of the training sample is used to estimate the logistic regression or the random forest and the second half to predict the matches. By comparing predicted matches to the ones assigned by the researchers, it is possible to compute TPR and PPV for each possible combination of b_1 and b_2. This exercise is performed over 100 random splits of the training sample into halves, and the average value across these 100 iterations is considered.

The result of this procedure for the logistic regression is graphically portrayed in Figure 1. Each square represents the average TPR and PPV for each combination of thresholds over 100 random splits.

Figure 1 *TPR and PPV for different thresholds of the minimum possible probability (b_1) and the ratio to the best possible probability (b_2) using the machine-learning approach*



The thresholds that maximise TPR + PPV are a minimum predicted probability (b_1) of .37 and a ratio of the best-predicted probability of the same record divided by the observed predicted probability (b_2) of 1.1, which returns a TPR of .83 and a PPV of .88. Using the random forest to predict probability the thresholds that maximise TPR + PPV are .34 and 1.4 for a TPR of .87 and a PPV of .88.

It must be noted that, in this case, equal weight is assigned to TPR and PPV in the selection of the thresholds. However, researchers can select thresholds that give more importance to precision rather than recall by maximising different functions of TPR and PPV, such as (0.5 * TPR + PPV).

6 RULE-BASED APPROACH

The main advantage of a rule-based approach compared to a supervised approach is that, in principle, it does not require the construction of a training dataset. Researchers do not need to have a training data with ground truth to estimate a probabilistic model and can set the similarity thresholds between attributes of two records based on their judgement and experience. On the other hand, for a rule-based approach we need to set a threshold for each feature we consider, while for the supervised approach all features are summarised in a single probability.

In this case, we leverage the training data in order to estimate the recall and precision of the rule-based approach. This approach further allows us to identify the thresholds that return the best possible outcomes. However, it must be noted that researchers adopting a rule-based approach will not be able to do so without the help of a training dataset.

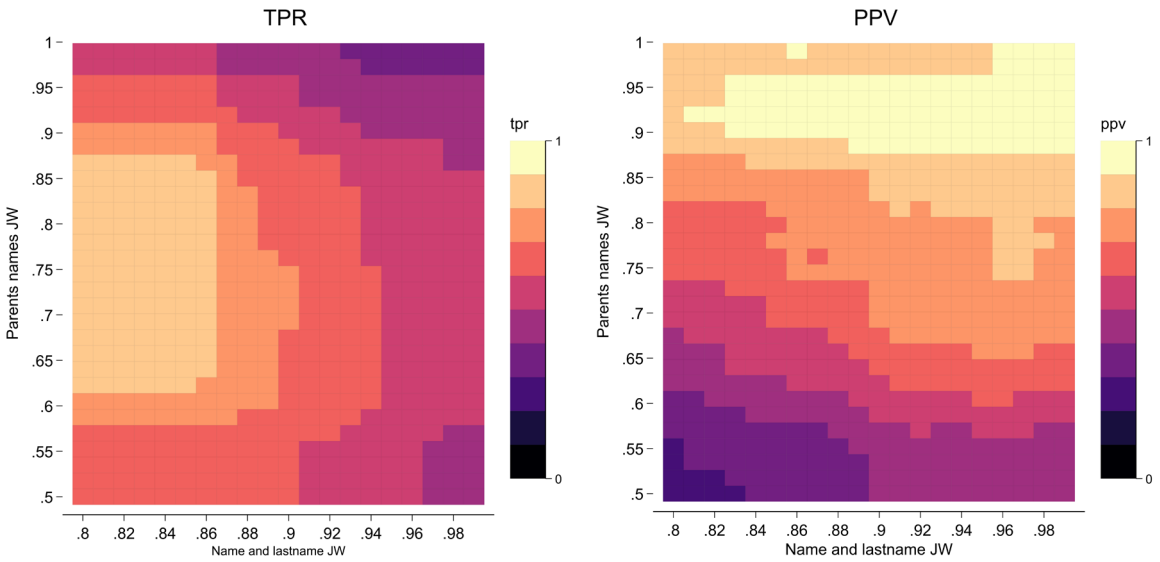
Two records are considered a match if a series of attributes are sufficiently similar (e.g., Dribe et al., 2023). In this case, we consider the first name, last name, father's first and last name, and mother's first and last name.[6] A record is considered a match if the Jaro-Winkler score between the names and last names is above a threshold of s_1, and the Jaro-Winkler for the parents' names above a threshold of s_2 when both records have a non-empty field in the parent name.

In order to investigate the TPR and PPV space where researchers would fall on average using different s_1 and s_2, the PPV and TPR for any possible s_1 and s_2 are estimated on the same 100 half samples on which the machine learning procedure is tested to allow comparability. The resulting space of combination and the estimates of TPR and PPV are reported in Figure 2.

The best possible combination maximising the simple TPR + PPV is a JW of .85 for name and last name and .88 for mothers' and fathers' names and last names, which return a TPR of .78 and a PPV of .900. While these thresholds are selected by comparison to training data they are not far for common practices in record linkage, suggesting that, even in the absence of trained data, sufficiently high similarity thresholds can guarantee precise linkages, while however discarding considerable links.

6 Fathers and mothers first and last names are recorded in a single field. Given that the names, multiple names, and last names in each field are not reported in any consistent order it is not possible to separate them and the full field is compared. To improve the measurement of the Jaro-Winkler score, strings in each field are alphabetically ordered. For instance, "Rosa Michelina Passarelli" is compared to "Passarelli Rosa Michelina" returning a Jaro-Winkler of .76. By alphabetically ordering both as "Michelina Passarelli Rosa" we can achieve a Jaro-Winkler score of 1. In some cases, for instance when extra words are included in one of the two fields, alphabetical sorting may result in a lower score. For instance, when "Antonia Rubino Michele" is compared to "Antonia Rubino" the non-sorted comparison has a score of .92, while the sorted of .90. A similar scenario occurs with fathers for whom often only the first name is reported assuming that last name is identical to that of the child. Alphabetically sorting in these cases is not the best option. For instance, when comparing "Domenico" to "Domenico Costanzo", the score between non-sorted strings is .89, while sorted is .45. To overcome this limit, for each dyad, we use the highest score between the one computed sorting the words in each field and the one without sorting.

Figure 2 *TPR and PPV for different thresholds of the minimum JW for names and last names (s_1) and the minimum JW for parents last names and first names (s_2) using the rule-based string similarity approach*



7 COMPARING MACHINE-LEARNING AND RULE-BASED ALGORITHMS' PERFORMANCE

The question underlying this work is which of the two procedures is most advisable for researchers trying to match large amounts of data digitized from parish and civil records. There is no definite response to this question, as the best performance depends on the researcher's needs and objectives. Here we describe the outcomes of the two procedures regarding precision, recall, and feasibility.

Figure 3 reports the full range of maximum PPV values corresponding to each achievable TPR level for the two procedures. At low levels of recall, both methods yield high and comparable precision. However, as recall increases, only the machine-learning approach — particularly the random forest model — maintains satisfactory precision.

The key result is that a machine learning approach can always replicate the performance of a rule-based approach if the researcher desires so, but not vice-versa. Most importantly, it must be noted that in this case, thresholds for the rule-based approach have been set using training data, an option that defeats the purpose of adopting a rule-based approach in many cases. A researcher would not know exactly what level of precision and recall along the line are approximating.

Machine learning, therefore, allows much greater flexibility, not only in the number of parameters that can be considered, since it does not require to specify a rule for each of them, but also in the trade-off between recall and precision available to the researcher.

On the other hand, a rule-based approach can still return high levels of precision with high enough thresholds at the expense of lower levels of recall.

Figure 3 reports the algorithms' performance in a quite ideal setting; indeed, the combination of the municipality of birth, first and last names, and both parents' names provide, in many cases, an almost unique identifier. However, such conditions are not always in place, and researchers may lack information. Therefore, we test both algorithms' flexibility and performance without selected information: father, mother, and parents.

The algorithms are trained on the same training data where the trainee could assign the correct match with all information available, but the removed information is not used in the algorithm's tuning and prediction.

Figure 3 *Average best possible values of PPV for possible values of TPR using machine learning and rule-based string similarity approaches*

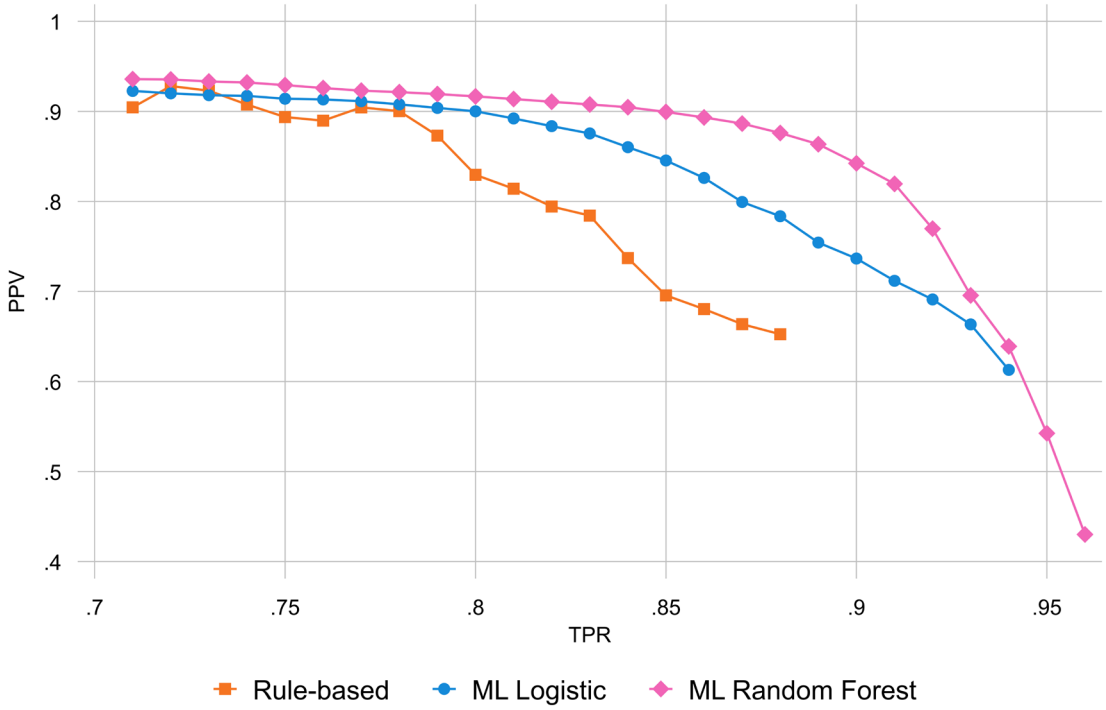


Figure 4 *Best possible values of PPV for possible values of TPR using machine learning and rule-based string similarity approaches and removing information on parents*

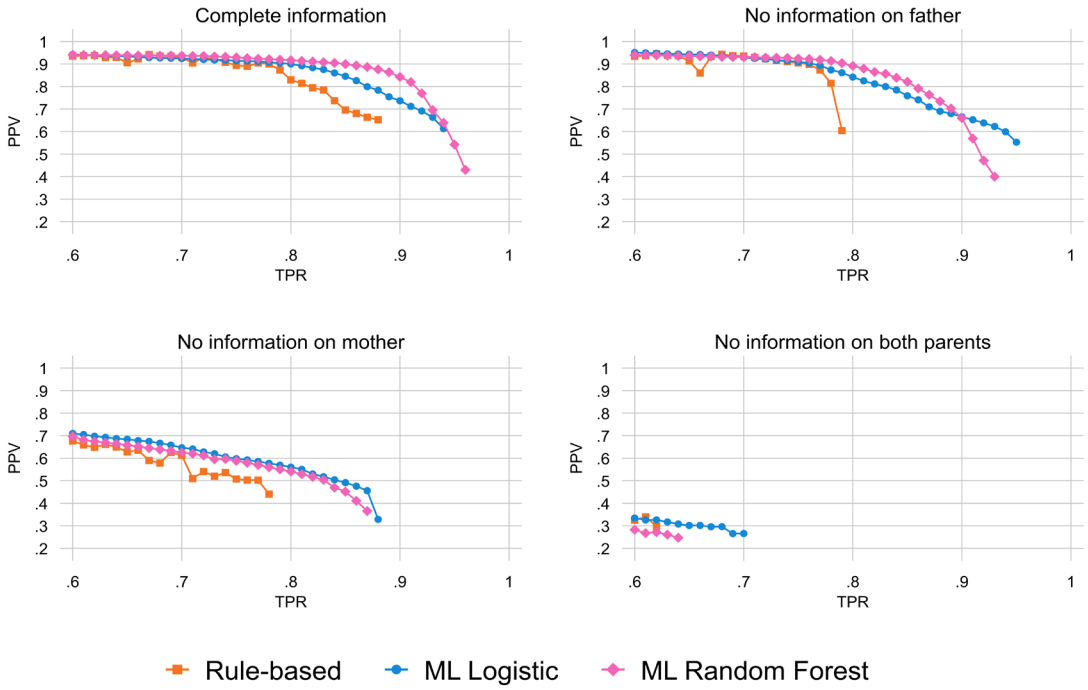


Figure 4 reports the results from this exercise resulting from 100 random sample splits. In both cases, father information does not influence precision levels as long as recall is below .8. This is likely because the last names of the reference individual partially capture father information. However, at higher levels of recall, only the probabilistic algorithm can maintain sufficiently high levels of precision. The rule-based similarity approach can never achieve a TPR of .8, and precision drops quickly at sufficiently high levels of recall.

With all approaches, the absence of information on mothers' names results in very low precision. Given that fathers' information is often already proxied by the last name, information on the mother is necessary for the disambiguation of multiple matches. Here, incorporating marriage records could be central, as they would enable the identification of the appropriate mother when missing and help resolve ambiguities in the linkage process. Finally, as already noted by visual inspection of the training data, records without information on at least one parent are often impossible to disambiguate, returning unacceptable levels of both precision and recall.

8 DISCUSSION

As the amount of available transcribed parish records increases, applied scholars will likely need to link individual records from different registers. The present article has reviewed, applied, and compared two recent approaches to linking historical records of birth and death registers transcribed by online volunteers from ItalianParishRecords.org.

Overall, we find that both automated methods perform well and can be effective options to build linked samples with few false connections in the case of complete information. At lower recall and higher precision levels, the two approaches performed similarly; however, if research intends to achieve higher recall, the machine-learning approach maintained higher levels of precision. Overall, machine-learning approaches allow the researcher more options in tuning the output of their linkage to their research.

Nevertheless, even a rule-based string similarity approach can reach high precision levels with sufficiently high cut-offs. It must be noted, however, that in this case, the best possible parameters for the rule-based approach have been set with the aid of training data. Generally, researchers using a rule-based string similarity approach will end up somewhere on the orange line in Figure 4, but will not have indication of where without tuning their choices against a training dataset.

Indeed, a general conclusion is that, regardless of the chosen approach, researchers should take advantage of a training dataset whenever possible. Training data allow much higher control over the performance of the linkage and allow researchers to acquire deep knowledge of the detailed characteristics of their datasets.

It must be noted that the data linked in this paper were not transcribed with the primary intent of being used for demographic research. As a result, the periods and geographic areas covered were scattered, and transcription quality was quite low. These limitations made linkage even harder, the performance of record linkage on parish records will likely increase with the improvement of transcription procedures and data quality.

We acknowledge that representativeness is a critical issue in historical record linkage studies, particularly given the inherent limitations of parish and civil records. Our primary focus in this article is on precision and recall, aiming to maximize linkage accuracy. Assessing representativeness is challenging for several reasons. First, the available variables for evaluating representativeness beyond municipality and birth cohort are limited. Second, the volunteer-based data collection introduces uneven coverage across locations and time periods, potentially affecting linkage opportunities differentially. Furthermore, linked parish records are inherently non-representative due to factors such as out-migration and the resulting difficulties in linking individuals who left a parish. Consequently, disentangling the representativity issues arising specifically from the linkage procedures from those generally present in historical linkage projects is difficult. Given these constraints, we have not explicitly investigated representativeness. The procedures outlined here focus primarily on maximizing the number of correct links. Assessing whether the resulting sample accurately represents the target population, as well as identifying potential solutions to representativeness issues, is left to researchers applying these methods, depending on their specific data and target populations.

Finally, here we focused on connecting births and deaths since more complete data for more periods were available in those registers. The transcription process is still ongoing, and, at this moment, registers other than births and deaths are too few to extend the linkage process further. Connecting birth and deaths is certainly useful for the study of mortality and longevity; however, the very same procedures

can be applied to the linkage of different registers, such as marriage records or family reconstructions, since the key information used in this exercise (e.g. names and last names, names and last names of relatives, dates of events) are the same in other registers.

ACKNOWLEDGEMENT

This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement 865356).

REFERENCES

Abramitzky, R., Boustan, L., & Eriksson, K. (2019). To the new world and back again: Return migrants in the age of mass migration. *ILR Review, 72*(2), 300–322. https://doi.org/10.1177/0019793917726981

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature, 59*(3), 865–918. https://doi.org/10.1257/jel.20201599

Avoundjian, T., Dombrowski, J. C., Golden, M. R., Hughes, J. P., Guthrie, B. L., Baseman, J., & Sadinle, M. (2020). Comparing methods for record linkage for public health action: Matching algorithm validation study. *JMIR Public Health and Surveillance, 6*(2), e15917. https://doi.org/10.2196/15917

Bailey, M. J., Cole, C., Henderson, M., & Massey, C. (2020). How well do automated linking methods perform? Lessons from US historical data. *Journal of Economic Literature, 58*(4), 997–1044. https://doi.org/10.1257/jel.20191526

Bouchard, G., Roy, R., & Casgrain, B. (1986). De la micro à la macro-reconstitution des familles le systeme SOREP [From micro- to macro-reconstitution of families: The SOREP system]. *Genus, 42*(3/4), 33–54.

Breschi, M., Fornasin, A., & Manfredini, M. (2011). Demographic responses to short-term stress in a 19th century Tuscan population: The case of household out-migration. *Demographic Research, 25*, 491–512. https://doi.org/10.4054/DemRes.2011.25.15

Breschi, M., Fornasin, A., & Manfredini, M. (2020). The richness of Italian historical demography. *Historical Life Course Studies, 9*, 228–240. https://doi.org/10.51964/hlcs9304

Breschi, M., Fornasin, A., Manfredini, M., Pozzi, L., Rettaroli, R., & Scalone, F. (2014). Social and economic determinants of reproductive behavior before the fertility decline. The case of six Italian communities during the nineteenth century. *European Journal of Population, 30*(3), 291–315. https://doi.org/10.1007/s10680-013-9303-8

Breschi, M., Fornasin, A., Pozzi, L., Rettaroli, R., and Scalone, F. (2009). The onset of fertility transition in Italy 1800–1900. In: A. Fornasin & M. Manfredini (Eds.), *Fertility in Italy at the turn of the twentieth century* (pp. 11–29). Forum.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.

Del Panta, L., & Rettaroli, R. (1994). *Introduzione alla demografia storica* [Introduction to historical demography]. Manuali Laterza.

Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The programme de recherche en démographie historique: Past, present and future developments in family reconstitution. *The History of the Family, 23*(1), 20–53. https://doi.org/10.1080/1081602X.2016.1222501

Dribe, M., Eriksson, B., & Helgertz, J. (2023). From Sweden to America: Migrant selection in the transatlantic migration, 1890–1910. *European Review of Economic History, 27*(1), 24–44. https://doi.org/10.1093/ereh/heac007

Dribe, M., & Lundh, C. (2010). Marriage choices and social reproduction: The interrelationship between partner selection and intergenerational socioeconomic mobility in 19th-century Sweden. *Demographic Research, 22*, 347–382. https://doi.org/10.4054/DemRes.2010.22.14

Dribe, M., & Quaranta, L. (2020). The Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies, 9*, 158–172. https://doi.org/10.51964/hlcs9302

Feigenbaum, J. J. (2016). *Automated census record linking: A machine learning approach* (Working paper). https://open.bu.edu/handle/2144/27526.

Feigenbaum, J. J., Helgertz, J., & Price, J. (2025). Examining the role of training data for supervised methods of automated record linkage: Lessons for best practice in economic history. *Explorations in Economic History, 96*, 101656. https://doi.org/10.1016/j.eeh.2025.101656

Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 29*(4), 141–156. https://doi.org/10.1080/01615440.1996.10112735

Fornasin, A., Breschi, M., & Manfredini, M. (2016). Environment, housing, and infant mortality: Udine, 1807–1815. In D. Ramiro Fariñas & M. Oris (Eds.), *New approaches to death in cities during the health transition* (pp. 43–54). Springer. https://doi.org/10.1007/978-3-319-43002-7_3

Fu, Z., Boot, H. M., Christen, P., & Zhou, J. (2014). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing, 8*(2), 204–225. https://doi.org/10.3366/ijhac.2014.0130

Fure, E. (2000). Interactive record linkage: The cumulative construction of life courses. *Demographic Research, 3*, Arcticle 11. https://doi.org/10.4054/DemRes.2000.3.11

Gautam, B., Terrades, O. R., Pujades, J. M., & Valls, M. (2020). *Knowledge graph based methods for record linkage* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2003.03136

Goeken, R., Huynh, L., Lynch, T. A., & Vick, R. (2011). New methods of census record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44*(1), 7–14. https://doi.org/10.1080/01615440.2010.517152

Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 55*(1), 12–29. https://doi.org/10.1080/01615440.2021.1985027

Herlihy, D. (1988). Tuscan Names, 1200–1530. *Renaissance Quarterly, 41*(4), 561–582. https://doi.org/10.2307/2861882

Kahle, P., Colutto, S., Hackl, G., & Muhlberger, G. (2017). Transkribus — A service platform for transcription, recognition and retrieval of historical documents. *14th IAPR International Proceedings of the Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan* (pp. 19–24). IEEE. https://doi.org/10.1109/ICDAR.2017.307

Mandemakers, K. (2002). Building life course datasets from population registers by the Historical Sample of the Netherlands (HSN). *History and Computing, 14*(1–2), 87–107. https://doi.org/10.3366/hac.2002.14.1-2.87

Manfredini, M. (2003). Families in motion: The role and characteristics of household migration in a 19th-century rural Italian parish. *The History of the Family, 8*(2), 317–343. https://doi.org/10.1016/S1081-602X(03)00031-9

Minello, A., Dalla Zuanna, G., & Alfani, G. (2017). First signs of transition: The parallel decline of early baptism and early mortality in the province of Padua (northeast Italy), 1816–1870. *Demographic Research, 36*, Article 27, 759–802. https://doi.org/10.4054/DemRes.2017.36.27

Piccione, L., Dalla Zuanna, G., & Minello, A. (2014). Mortality selection in the first three months of life and survival in the following thirty-three months in rural Veneto (North-East Italy) from 1816 to 1835. *Demographic Research, 31*, Article 39, 1199–1228. https://doi.org/10.4054/DemRes.2014.31.39

Price, J., Buckles, K., van Leeuwen, J., & Riley, I. (2021). Combining family history and machine learning to link historical records: The Census Tree data set. *Explorations in Economic History, 80*, 101391. https://doi.org/10.1016/j.eeh.2021.101391

Pujadas-Mora, J. M., Fornés, A., Ramos Terrades, O., Lladós, J., Chen, J., Valls-Fígols, M., & Cabré, A. (2022). The Barcelona historical marriage database and the Baix Llobregat demographic database. From algorithms for handwriting recognition to individual-level demographic and socioeconomic data. *Historical Life Course Studies, 12*, 99–132. https://doi.org/10.51964/hlcs11971

Rettaroli, R., Samoggia, A., & Scalone, F. (2017). Does socioeconomic status matter? The fertility transition in a northern Italian village (marriage cohorts 1900–1940). *Demographic Research, 37*, Article 15, 455–492. https://doi.org/10.4054/DemRes.2017.37.15

Rettaroli, R., & Scalone, F. (2012). Reproductive behavior during the pre-transitional period: Evidence from rural Bologna. *The Journal of Interdisciplinary History, 42*(4), 615–643. https://doi.org/10.1162/JINH_a_00307

Rettaroli, R., Scalone, F., & Del Panta, L. (2019). The demography of isolated populations. A research note on a German-speaking community in a northern Italian valley between the 18th and 19th century. *Popolazione e storia, 19*(2), 105–123. https://doi.org/10.4424/ps2018-10

Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology, 44*(1), 19–37. https://doi.org/10.1146/annurev-soc-073117-041447

Ruiu, G., & Breschi, M. (2015). For the times they are a changin': The respect for religious precepts through the analysis of the seasonality of marriages. Italy, 1862–2012. *Demographic Research, 33*, Article 7, 179–210. https://doi.org/10.4054/DemRes.2015.33.7

Scalone, F., Agati, P., Angeli, A., & Donno, A. (2017). Exploring unobserved heterogeneity in perinatal and neonatal mortality risks: The case of an Italian sharecropping community, 1900–39. *Population Studies, 71*(1), 23–41. https://doi.org/10.1080/00324728.2016.1254812

Scalone, F., & Samoggia, A. (2018). Neonatal mortality, cold weather, and socioeconomic status in two northern Italian rural parishes, 1820–1900. *Demographic Research, 39*, Article 18, 525–560. https://doi.org/10.4054/DemRes.2018.39.18

Tymicki, K. (2009). The correlates of infant and childhood mortality: A theoretical overview and new evidence from the analysis of longitudinal data of the Bejsce (Poland) parish register reconstitution study of the 18th–20th centuries. *Demographic Research, 20*, Article 23, 559–594. https://doi.org/10.4054/DemRes.2009.20.23

Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies, 9*, 114–129. https://doi.org/10.51964/hlcs9299

Wen, F., In, J., & Breen, R. J. (2022). A comprehensive assessment of census record linking methods: Comparing deterministic, probabilistic, and machine learning approaches. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4241435

Westberg, A., Engberg, E., & Edvinsson, S. (2016). A unique source for innovative longitudinal research: The POPLINK database. *Historical Life Course Studies, 3*, 20–31. https://doi.org/10.51964/hlcs9351

Winchester, I. (1992). What every historian needs to know about record linkage for the microcomputer era. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 25*(4), 149–165. https://doi.org/10.1080/01615440.1992.10112722

Wrigley, E. A., Davies, R. S., Oeppen, J. E., & Schofield, R. S. (1997). *English population history from family reconstitution 1580–1837* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511660344

APPENDIX

Table A1 *Sample distribution by region*

	Births		Deaths	
Region	**N**	**%**	**N**	**%**
Abruzzo	44,373	5.96	18,689	4.38
Basilicata	22,004	2.96	41,278	9.68
Calabria	334,231	44.9	203,697	47.78
Campania	128,328	17.24	80,083	18.78
Emilia Romagna	2,246	0.3	2,186	0.51
Lazio	2,200	0.3	11,327	2.66
Lombardia	4,136	0.56	3,303	0.77
Piemonte	129,151	17.35	4,897	1.15
Puglia	6,035	0.81	5,130	1.2
Sardegna			3,503	0.82
Sicilia	71,728	9.64	52,251	12.26
Total	744,432	100	426,344	100%

Figure A1 *Distribution of observations by year and type of records*

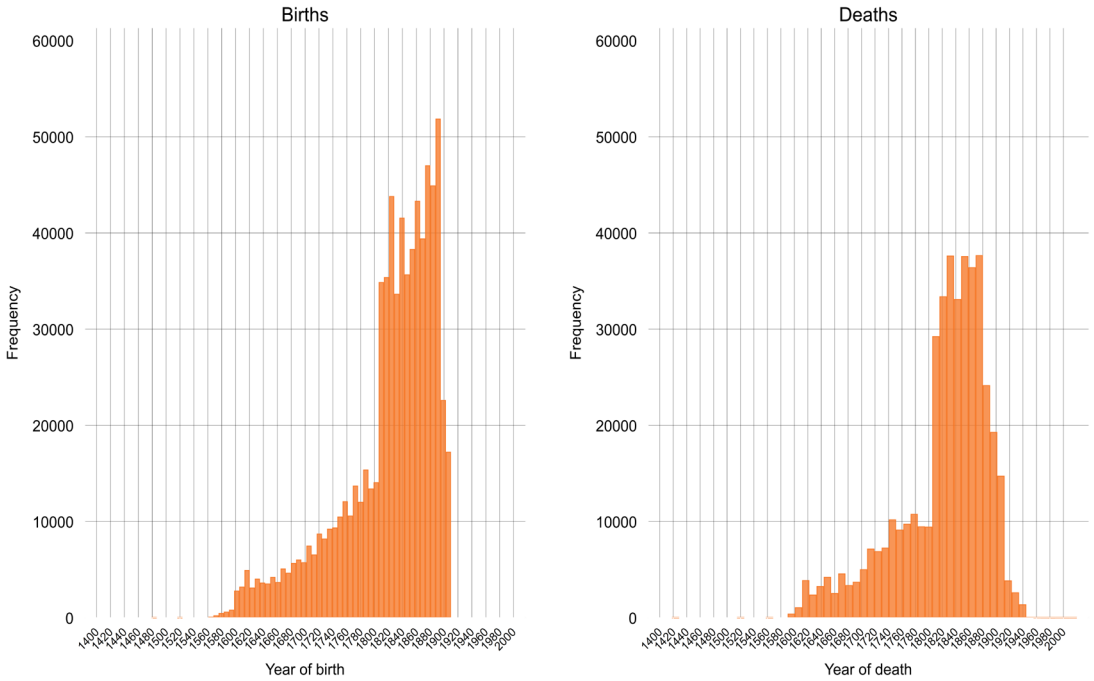


Table A2 Descriptives statistics of training dataset of possible matches

	mean	sd	min	max
Correct and multiple matches	0.041	0.198	0.000	1.000
JW Name	0.913	0.065	0.800	1.000
JW First word name	0.947	0.132	0.000	1.000
JW Last Name	0.925	0.080	0.800	1.000
JW Father	0.638	0.238	0.000	1.000
JW Mother	0.566	0.205	0.000	1.000
Same sex	0.758	0.428	0.000	1.000
Number of hits	36.939	41.569	1.000	188.000
First letter last name equal	0.924	0.265	0.000	1.000
First letter first name equal	0.973	0.162	0.000	1.000
Missing parent	0.113	0.317	0.000	1.000
Observations	4,359			

Table A3 *Logit model of correct matches on the characteristics of the potential matches dyads*

	(1)
Variables	**Model 1**
JW first name	249.8*
	(145.9)
JW first name squared	-127.7
	(78.44)
JW first word of the first name	104.3
	(88.32)
JW first word of the first name squared	-56.92
	(49.88)
JW last name	732.0*
	(412.9)
JW last name squared	-379.1*
	(216.6)
JW father	-13.46***
	(3.155)
JW father squared	16.53***
	(2.810)
JW mother	-21.55***
	(2.152)
JW mother squared	24.58***
	(2.213)
Same sex	2.370***
	(0.826)
Number of hits	-0.0323**
	(0.0154)
Number of hits squared	0.000102
	(0.000113)
First letter last name equal	-1.370
	(2.615)
First letter first name equal	-0.215
	(1.483)
Constant	-524.6**
	(213.3)
Observations	4,359

Notes: Standard errors in parentheses.

** p < 0.1, ** p < 0.05, *** p < 0.01.*