The Record Linking Glass Ceiling. Applying Automated Methods to the Census and Women's Marriage Records, 1881–1911

By Emma Diduch

To cite this article: Diduch, E. (2024). The Record Linking Glass Ceiling. Applying Automated Methods to the Census and Women's Marriage Records, 1881–1911. *Historical Life Course Studies, 14*, 126–143. https://doi.org/10.51964/hlcs19189

HISTORICAL LIFE COURSE STUDIES

VOLUME 14

2024

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 14 (2024), 126–143, published 11-11-2024

The Record Linking Glass Ceiling

Applying Automated Methods to the Census and Women's Marriage Records, 1881–1911

Emma Diduch
University of Cambridge

ABSTRACT

This paper presents the results of a project creating a linked dataset of census and civil registration records from the county of Derbyshire, England. The proposed method includes women at every stage, first by comparing the performance of deterministic, probabilistic, and household-based linking methods and then expanding the linking process to capture women who have changed names between censuses due to marriage. In census-to-census linking the best results are obtained through a combination of probabilistic and household-based methods, linking between 40% and 45% of the starting population in each decade 1881–1911. The quality of these links and possible impacts of migration patterns are discussed with reference to the representativeness of the linked sample. Incorporating transcribed indexes of marriages (which are freely available online) allows women to be followed in the census across their marriages. Combined, this process reduces the gap in linking success between women and men and especially improves match rates for women in their twenties by between fifteen and twenty percentage points. These data have important potential for future record linking efforts and for research exploring women's work, marriage, and fertility in a life course perspective.

Keywords: Record linkage, Marriage, Fertility, Census data, Civil registration, England and Wales

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs19189

1 INTRODUCTION

Until recently, women have often been absent in large-scale historical record linking projects. This absence has been not so much an oversight as a methodological challenge. In times and places where women traditionally changed surnames upon marriage, it is usually impossible to track these changes through a single data source like the decennial census. Not only is there an upper limit of links that can be made for women between two censuses, but the unknown outcomes of marriages and remarriages might also undermine confidence in the veracity of these links.

In addition, historians have long criticized census data as a source for information on women's lives when this information was recorded on their behalf by (usually male) heads of household. For example, changing instructions for enumerators and cultural perceptions of work have been blamed for inconsistencies and errors in the recording of women's occupations in the 19th-century British censuses (Higgs & Wilkinson, 2016; Hill, 1993; McGeevor, 2014). As digitized census datasets became available, efforts to algorithmically link individuals over the decades began by focusing exclusively on men (Ferrie, 1996).

Are data issues entirely to blame for these omissions? The research aims driving the creation of linked datasets have also tipped the gender balance in favour of men. Occupational data on individuals and their co-resident families — particularly fathers and sons — are important for studies of geographic and socioeconomic mobility across generations (Abramitzky, Boustan, Jácome, et al., 2021; Long & Ferrie, 2013). The same methods cannot be easily replicated for women, both because of the differences in surname between mothers and married daughters and because women's work patterns are often more fluid and contingent on marital status and family size.

Nevertheless, these transitions — between childhood and adulthood, between work, marriage, and childbearing — are crucial subjects for both women's history and historical demography, and these topics are motivating a new wave of automated record linking. Following women over extended periods of time must involve additional civil registration sources, which have not been digitized and made publicly available to the same extent as national census data and often suffer from additional quality issues. The difficult process of linking women to their marriage records may also introduce new sources of error or bias. To address these challenges, this paper develops an adapted methodology for linking both women and men in England by combining census records with a new source: indexes of marriage.

Complete civil registration records for this period are not available in the UK, although some studies have made use of limited samples in combination with other data sources (Akgün et al., 2020; Arthi et al., 2022; Newman & Smith, 2023; Reid et al., 2002). Indexes of births, deaths, and marriages have been transcribed and made publicly available but differ from full certificates in content and organization. Crucially, the marriage indexes refer to pages in the original registers which usually contain two marriages, so the linking process must include a strategy for allocating spouses to the correct partners. Combining indexes of marriage with direct census-to-census links improves match rates for women in their twenties by up to twenty percentage points

This paper begins by adapting common record linking approaches using census records for the county of Derbyshire, England for 1881 and 1891. A probabilistic approach is chosen over a deterministic method to improve match rates and is used as the basis for further linking within households. This multi-stage process differs from similar studies by including women at every stage of linking, both before and after marriage. To account for the effects of marriage, women in direct census-to-census linking are grouped by marital status. In the final stage, indexes of marriages are used to link single to married women. After comparing and combining these methods, this paper explores the composition of the linked sample and possible biases. Finally, these results are discussed in relation to the potential of this methodology for further research in historical demography.

1.1 LINKING CENSUS DATA FOR ENGLAND AND WALES

The central challenge of many historical record linking projects concerns the trade-off between the number and the accuracy of links, especially when scaling up from small, manually linked samples to automated linking of entire populations. A growing number of studies of different record linking strategies concentrate on these trade-offs to optimize the balance of sensitivity and specificity in linked data. Maximizing the number of links (i.e., avoiding missed links, or false negatives) comes with the risk of including more erroneous links, or false positives. Several recent articles have compared the performance of three record linking approaches: rule-based or deterministic linking, probabilistic linking, and machine learning algorithms.

These comparisons consistently find that probabilistic linking offers a good balance between match rates and accuracy without requiring large, hand-linked training datasets which are both costly to produce and less replicable for other researchers (Abramitzky, Boustan, Eriksson, et al., 2021; Bailey et al., 2020; Massey, 2017; Wen et al., 2022). In addition, other census linking projects have moved towards a household-based approach to incorporate additional linking variables based on the people with whom an individual is living at each census date (Antonie et al., 2020; Fu et al., 2014; Helgertz et al., 2022). These approaches include links for women, but only in their roles as sisters, daughters, mothers, and wives.

Concerns regarding accuracy and representativeness underpin the choice of linking criteria to minimize sources of both error and bias in the linking process. At the most basic level, record linking methods choose time invariant traits to avoid overrepresenting individuals whose circumstances are more stable over time or incorrectly linking individuals whose attributes have changed (Ruggles et al., 2018). In practice, historical records have many discrepancies in individual entries which must be accounted for in the linking process, ranging from misspelled names or imprecise recording of ages to the practice of women changing surnames at marriage (Hwang & Squires, 2024).

The functional distinctions between different automated linking methods are not necessarily in their ability to identify possible links between records from two censuses, but in their ability to discern between multiple competing links based on a limited number of linking variables. A genealogist or manual record linker often considers additional pieces of information such as address or occupation in choosing the right match for John Smith and Jane Doe, but on a systematic basis these decisions contribute to a biased linked sample where people who change jobs or move house (or simply record these details differently) are more likely to be lost from the historical record. When linking an entire population, therefore, researchers must rely primarily on variables that are expected to be constant over time and are consistently recorded in the source data — in historical censuses, these are usually name, sex, age, and birthplace.

For research in England and Wales, the Integrated-Census Microdata (I-CeM) datasets are based on transcriptions from the census enumerators' books which have been enriched with individual- and household-level coded variables, in addition to preserving original records as transcribed (Schürer & Higgs, 2020a). For example, the I-CeM data include age at last birthday as an integer and as the original character string from entries such as 'newborn' or 'six months old'. An individual's precise birth year is thus expected to be within a two-year window, depending on the timing of census night relative to their birthday.

Full names and addresses are not included in the publicly accessible version of I-CeM data, as they are protected by special license (Schürer & Higgs, 2020b). The protected version of the data can be requested for academic research purposes and merged with the anonymized census files. Names are rendered as transcribed and require cleaning and formatting to successfully compare census entries. For the linking process described below, names are reformatted (lowercase, without special characters), and some common nicknames are lengthened (following Abramitzky, Boustan, Eriksson, et al. (2021)). Middle names are unevenly recorded and in some census files first and middle names must be parsed from a single column.

1.2 MOTIVATION AND LOCATION

In addition to the household and occupational variables available in each census dataset, the release of full-count census data for England and Wales includes the 1911 census and its supplementary fertility questionnaire for married women. These data have allowed for detailed examination of the role of social class, occupational structure, and geographic variation in the late 19th-century fertility transition (Garrett & Reid, 2018; Jaadla et al., 2020). However, important questions remain unanswered about the causes of widespread, geographically and occupationally specific fertility decline.

The fertility information collected in 1911 was retrospective, meaning couples' reported occupations and locations did not reflect their circumstances during the most critical periods of family formation. Studies which have constructed marriage and childbearing estimates from earlier censuses have established strong cross-sectional relationships but cannot assess the effects of these variables within the individual life course (Garrett et al., 2001).

For example, within a fertility distribution inversely stratified by social class, textile workers represented an anomaly in the 1911 census with births per couple closer to those of the lower middle class than of the skilled working class (Garrett, 1990; Woods, 2000). Whether these patterns suggest a higher

opportunity cost associated with women's labour force participation or reflect a distinctive industrial culture that is geographically as much as economically determined, a longitudinal perspective is necessary. Were women in textile districts motivated by high wages to reduce their fertility and continue working in the factories, or were married textile workers selected into their occupations by existing low fertility?

The importance of women's work in analyses of the fertility transition motivates the choice of an area with textile manufacturing as a case study for record linkage between census and marriage records. In contrast with the urbanization and manufacturing monoculture of the textile centres of Lancashire, the county of Derbyshire represents both a long-standing history of textile manufacturing and a greater diversity of occupations and environments. In addition to a series of textile mills along the Derwent River, starting with Arkwright's water-powered spinning mill in Cromford in 1771, the county also had substantial employment in mining, both lead and coal, heavy industry, such as in iron and machine making, and pastoral agriculture (Fitton & Wadsworth, 1958). As both a basis for further study of women's work and fertility and a proof of concept for the use of census and civil registration sources to improve data coverage for women, these results represent an important contribution to the literature on historical record linking.

2 CENSUS-TO-CENSUS LINKING

Many linking studies start by selecting census records for men, later including women grouped with their households. Since the methodology used in this paper includes a separate step to link women to records of their marriage, this first stage aims to link women who did not marry between censuses. The full county census files are first grouped by sex and by marital status for women. This prevents single women in one census from being compared to married women in the next census, and vice versa. Separately, married women in the first census are compared with women in the next census who are either married, widowed, or whose spouses are absent on census night. This approach underrepresents widows because of the possibility of remarriage but reduces the risk of false or missed matches for married women who are widowed between censuses.

Each record linking approach begins by generating pairs of records from two census files. Even restricting the population to one county, the potential combinations of records between the 385,947 people in Derbyshire 1881 and the 432,128 in 1891 would be astronomical without limiting the universe of potential pairs. In addition to separating files by gender and by marital status for women, this is done by 'blocking' or grouping the population by birthplace so that pairs are only generated which exactly agree on county of birth. Pairs are assessed based on the linking variables to create some measure of similarity — whether a string comparison for names such as Jaro-Winkler similarity scores or the absolute difference between imputed birth years. Links are then accepted if they exceed a chosen threshold of similarity.

2.1 DETERMINISTIC VS PROBABILISTIC RECORD LINKING

Deterministic linking methods follow a relatively simple set of rules for comparing pairs to a threshold of similarity but often falter when multiple potential matches for one individual fall above the chosen threshold. Increasing the linking threshold (e.g., by requiring exact match on birth year or same spelling of name) does not solve this problem for individuals with common names and risks discarding true, unique matches which fall below the higher threshold. As a basis for comparison, the Derbyshire data are first linked using an adapted version of the ABE deterministic method, with thresholds of 0.9 Jaro-Winkler string similarity scores for names and age differences of one year or less (Abramitzky, Boustan, Eriksson, et al., 2021). These thresholds are also consistent with the most common reporting errors found by Hwang and Squires (2024) when comparing census data with genealogical profiles.

After blocking and comparing pairs of records from each census, there is significant attrition of potential links due to the need to discard duplicates. For example, combining 192,898 men in Derbyshire in 1881 with 215,172 men in 1891 results in 156,729,233 pairs of records with the same birthplace and initials. Out of these, 167,358 pairs have names and ages similar enough to exceed the matching threshold, but most of these matches are not unique. Removing cases where a man in 1881 has been paired with multiple possible matches in 1891 (and vice versa) leaves 56,242 pairs, representing a match rate of 29.16% for the 1881 male cohort.

Probabilistic linking addresses this problem of multiple matching records, which are especially concentrated among people with common names, by estimating the probability that a pair of records is a match based on the characteristics of the total population (Abramitzky et al., 2020; Fellegi & Sunter, 1969). This allows the similarity threshold for a link to be based on the probability that two random individuals could share the same characteristics by chance. For example, there were 255 men named John Smith who were born in Derbyshire and still living in the county in 1881, but only one man named Walter Shakespeare. Probabilistic linking algorithms weight the similarity of linking variables by their prevalence in the population, so an exact match on name and age for John Smith might be given a lower probability of matching than a match for Walter 'Shakspere' with a year's difference from the expected age.

This method mimics the decisions made by human record linkers when presented with a slate of possible matches to contextualize the best candidates, but obtains the same results every time given consistent inputs and thresholds (Bailey et al., 2023). Studies which have compared the results of automated linking methods to a 'ground truth' sample of manually linked or genealogical data find that probabilistic algorithms have lower false positive error rates, with incorrect matches representing between 5% and 15% of men linked compared with up to 30% of deterministic links (Abramitzky, Boustan, Eriksson, et al., 2021; Bailey et al., 2020; Helgertz et al., 2022; Massey, 2017). Unlike manual linking or trained machine learning algorithms this process is fully automated and replicable.

The *reclin2* package for probabilistic record linking in R allows the user to specify the threshold of probability to select a pair of records as a link and to enforce one-to-one linking in the final linked sample (van der Laan, 2022).[1] For use with historical data here, the package is adapted with a custom comparison function to allow for age discrepancies between possible matches. Several tests (not discussed here) revealed that the best version of this method differs by gender: in addition to the variables used in deterministic linking above, match rates for men are better when middle initials are included among the linking variables, and for women when allowing for a wider range in implied year of birth. Applying probabilistic linking between the 1881 and 1891 Derbyshire censuses and selecting unique pairs with a greater than 50% matching probability obtains 123,265 links, or 31.94% of the starting population in 1881. This threshold is less conservative than some purely probabilistic linking methods because these links are further assessed and validated in household linking stages (see below).

Probabilistic linking increases match rates for all groups, but the improvement is slightly smaller for men than for women (Table 1). This is consistent with the expected effects of probabilistic linking on the chances of matches for common names, as the census data reveal a smaller name pool for men. Nearly 70% of the male population of Derbyshire in 1881 shared just 10 first names; in contrast, women's names are less concentrated with only 56% of women covered by the top 10 first names. This illustrates why the inclusion of middle names for men helps to clarify some of these links.

Match rates for married women are the highest in the probabilistic links, while match rates for single women lag far behind. The practice of the head of a household filling out a census schedule on behalf of their dependants means that women's census records are expected to have more variation as fathers, husbands, and employers may have misstated ages or misspelt names. Thus, even though women have a higher prevalence of middle names in their census records, including this variable in the linking process does not improve match rates but, on the contrary, may decrease matching efficiency by 'tightening' the criteria too far and privileging poor quality matches with middle names over higher quality matches without (Goeken et al., 2011).

Some common surnames, including Smith, Taylor, and Walker, are better represented in the probabilistic links. Other surnames which appear to be less common or unique are less likely to be linked if they overlap with another common name, suggesting that they are misspellings or mistranscriptions of the same name (e.g., Sharp and Sharpe, Davies and Davis, or Higgenbotham, Higgenbottom, and Higginbottam). The same is true for first names because not all spelling differences and nicknames can be accounted for and there remain significant overlaps between distinct names like Helen and Ellen or Abraham and Abram. These patterns underscore the role of probabilistic linking in better assessing potential matches based on a comprehensive similarity score across these attributes rather than on discrete thresholds for each variable.

1 In the case of multiple potential links falling above a 50% probability of matching (because these probabilities are estimated independently), the reclin2 package selects the highest probability link.

Table 1 *Census-to-census links by method for Derbyshire, 1881–1891*

Method	Men	Single women	Married women	Total
Deterministic (% linked)	56,242 (29.16%)	25,809 (22.71%)	20,713 (32.64%)	102,764 (26.63%)
Probablistic	64,704 (33.54%)	34,830 (30.65%)	23,731 (37.40%)	123,265 (31.94%)
Probablistic + Household	73,851 (38.92%)	36,817 (32.40%)	25,458 (40.12%)	137,460 (35.62%)
After iterating	85,362 (44.25%)	37,673 (33.15%)	27,347 (43.10%)	151,798 (39.33%)
1881 Census population	192,898	113,631	63,452	385,947

2.2 LINKING BASED ON HOUSEHOLDS

Several recent articles have expanded the selection of linking variables by including household characteristics or even neighbourhood contexts in scoring potential matches (Akgün et al., 2020; Fu et al., 2014; Helgertz et al., 2022; Wisselgren et al., 2014). This is an important movement towards diversifying the composition of linked datasets by including more women in the linking process but does not solve the problem of linking women across marriage. In addition, this approach advantages links for individuals who remained in the same household across censuses, while research that is interested in mobility, marriage, and family formation must include links to new households. On the other hand, the ability to clarify high-confidence links for individuals who remain in the same household has also been shown to decrease overlap with potential matches for individuals who move, thus improving linking success across the board (Antonie et al., 2020).

For this project, probabilistic linking is used as a basis for further household linking. This approach combines the higher linking rates of household-based methods with the higher confidence links of probabilistic methods for those who changed households between censuses. Evaluating household characteristics provides a useful check on the accuracy of probabilistic links (Ó Gráda et al., 2024) and additional household links are expected to be highly accurate (Helgertz et al., 2022). Household contexts can be used to both rule in links that are otherwise uncertain (whether because of discrepancies in name and age or because of competing possible links) as well as rule out links which are logically improbable (including children with different parents and wives with different husbands).

In this second stage of record linking, households which have a member linked from the first probabilistic stage are used to limit the universe of potential matches for other unlinked household members. To prevent erroneous matches — for example between sisters and wives of the same first name — linked household members must have the same mother, father, or spouse present or have the same relationship to the head of household (including grandchild, servant, boarder, etc.) if a spouse or parent is not present in both censuses. In addition, linked spouses are compared to each other to avoid false links for married women — for example, if the link for Herbert Hodgkinson shows that he married for the first time between censuses, his wife should not be matched to a married woman in the previous census. Otherwise, similarity thresholds can be relaxed since a link between records with greater discrepancy in ages or lower name similarity is more likely if they are listed as a member of the same household in two census years.[2]

Although this process is primarily intended to consider matches for unlinked individuals, all individuals who are part of a linked household are included, thus confirming first-stage links which also pass the household linking criteria. When links from the two stages are combined, if a potential household link disagrees with a probabilistic link for the same individual, the household match is accepted as more likely. First-stage links which fail to pass household-linking thresholds (i.e., have been linked to individuals with different parents or spouses) are discarded.

2 Although there are further ways to enrich individual census records with household variables, including characteristics of siblings, children, or extended kin, this straightforward approach has the advantage of exploiting the existing household descriptors in the I-CeM dataset, including unique household IDs and indicators for spouses and parents within households.

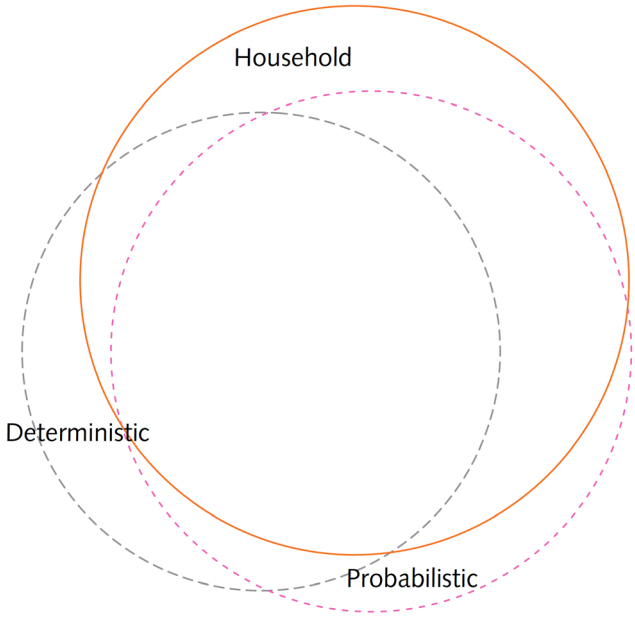
This process also provides a rough estimate of the extent of false positive errors in the probabilistic linking process, although the true extent of error (both false links and missed links) is unknown. In total 10.3% of the probabilistic first stage links are discarded in the household stage — and the proportion discarded varies by subgroup from 8.2% of married women to 13.4% of single women. This error rate is probably a lower-bound estimate but is within the range found by comparison of other probabilistic linking methods to genealogical or hand-linked samples. It is also important to note that these 'ground truth' comparisons are themselves subject to error — human linkers often disagree with each other and where they disagree with algorithmic results there may not be a clear 'winner' (Abramitzky, Boustan, Eriksson, et al., 2021). Individuals for whom links can be confirmed using genealogical records are likely not representative of the total population.

The household linking process improves total match rates significantly, particularly for men and married women. For single women the improvement is more muted as there is an upper limit to the matches that can be made without additional information from marriage records (Table 1). These results support the inclusion of women in individual as well as household-based linking — the expected limit to direct census-to-census links for single women and low error rates in links for married women suggest that the algorithm is performing well for these groups compared with the results of studies which linked men only.

The household linking stage also doubles the number of complete households, in which all members in one census are linked to the next (8,508 distinct households, excluding single-person households). On average, 67.5% of individuals within each household in 1881 are linked to 1891. This is an underestimate of linking success among the linkable population since some household members have died between censuses. Figure 1 compares the additional household links with those from the probabilistic and deterministic methods, showing that most of the new matches are unique to the household stage, not simply confirming links which could have been obtained deterministically. This emphasizes the value added through household linking to identify new links or choose between competing links while allowing for the discrepancies in names and ages which are prevalent in historical data.

To obtain the initial linked census dataset, the probabilistic and household linking stages are repeated with the remaining unlinked populations, under the assumption that the removal of competing matches during the first linking run may help to confirm additional links.[3] Again, improvement in the match rates for this stage of linking is mostly among men and married women (Table 1). In total, 151,798 links are obtained in this phase of census-to-census linking, representing 39.3% of the starting population in 1881 or 48.3% of the 1891 population over 10 years old. This represents a 13 percentage point improvement over the baseline deterministic match rates.

Figure 1 *Venn diagram of links made by deterministic, probabilistic, and household linking methods, 1881–1891*



3 For the second probabilistic run, the probabilities of matches are based on the characteristics of the entire population, as otherwise common names might be even further overrepresented in the unlinked group.

3 REPRESENTATIVENESS OF THE LINKED SAMPLE

Although this methodology was developed using the 1881–1891 census data, application to subsequent censuses for Derbyshire confirms its efficacy. Moving forward in time, linking success for all groups improve for a total match rate of 44% 1891–1901 and 45.4% 1901–1911. These results suggest that the linking results are not an artifact of the characteristics of the 1881 and 1891 censuses and that the probabilistic and household linking steps behave consistently with improved data quality over time, despite differences in enumeration and transcription between censuses.

This paper follows established precedents for testing record linking methods using a sample to avoid the processing time it takes to link entire populations. The fact that this sample is geographically defined rather than randomly selected has implications for patterns of error, although these cannot be systematically evaluated here. In particular, other studies which link from random samples to full count censuses may reach different conclusions about the likelihood of false positive errors, given the larger pool of potential matches in the terminal census and the need to distinguish accurately between them (Bailey et al., 2020; Helgertz et al., 2022). Here, however, the restriction of linking within one county is likely to have a different effect by slightly reducing the potential pool of matches in the terminal census due to outmigration. There might thus be fewer false matches to be made, with migration affecting match rates more than accuracy.

The effects of migration as well as mortality between censuses can be estimated by looking at the terminal census for the 'potentially linkable population' who are native-born and over the age of 10 (Long & Ferrie, 2013; Ruggles et al., 2018). For the whole of England, the native-born population born before 1881 was 16% smaller in the 1891 census. When focusing here on a county rather than a country, the linkable population can be affected by short-distance and return migration as well as by emigration. It is therefore useful to compare the Derbyshire-born population within and outside the county in each census.

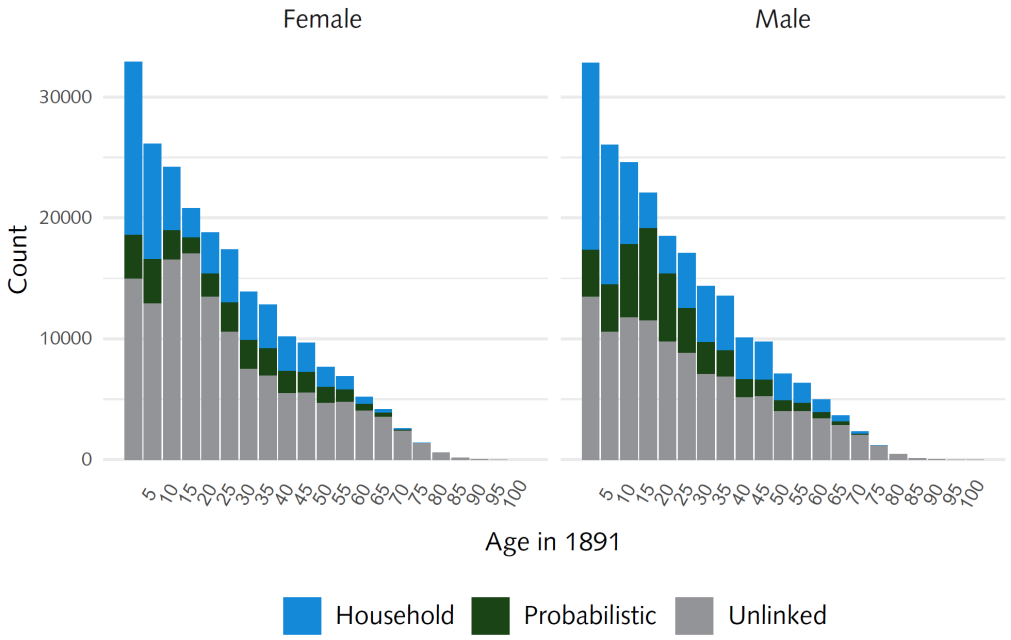
The equivalent loss between 1881 and 1891 for the population born in Derbyshire and living anywhere in England was 7%, suggesting that this group was less likely to emigrate or experienced lower mortality, or both. However, in 1891 a greater proportion of the surviving Derbyshire-born cohort were living outside the county (increasing from 36% in 1881 to 50% in 1891). Comparing the native-born population remaining in the county at each census, the maximum match rate for the linkable population is around 70%, assuming little to no return migration between 1881 and 1891.[4] For single women in the county who are also 'lost' to marriage, the maximum direct match rate is around 40%. The initial census-to-census links thus do not 'overperform' expectations.

Migration effects are reflected in lower match rates for parishes near the county borders where people may have easily moved outside the view of the selected data. Linking within the county prioritizes less-mobile links and might be helped by filtering the universe of potential matches or hurt by not considering likely candidates outside the county. The most common birthplace in each census is Derbyshire itself, with 64.1% of the population in 1881 reporting having been born in the county. This is followed by the neighbouring counties of Nottinghamshire, Staffordshire, and Yorkshire, and match rates are up to twenty percentage points lower for individuals born outside the county (Table A, Appendix).

Other characteristics which are associated with mobility are also underrepresented in the linked sample. Unskilled workers have the lowest match rates, possibly related to migration in search of work and to lower levels of literacy contributing to inconsistent census entries. Both women and men who reported textile occupations in the census are slightly more likely to be linked than those in other broad occupational categories such as in agriculture or domestic service, possibly related to lower migration levels associated with their employment. These patterns are expected given the better linking performance for native born, middle- or upper-class populations found by other studies, and researchers using linked data must be careful to acknowledge the potential impact on their analyses (Helgertz et al., 2022; Massey, 2017).

4 This proportion is calculated from the difference between the populations born in Derbyshire before 1881 and still present at each census (247,294 vs 181,166), adjusted downward to account for under-enumeration and non-unique cases (about 6%) (for more details see Long and Ferrie (2013), and Wen et al. (2022)).

Figure 2 *Census links by age, sex, and method, 1891–1901*



In addition to migration, linking success is also affected by mortality and marriage. Figure 2 shows lower match rates in the youngest and oldest age categories (a higher proportion of each bar in grey), illustrating the effects of childhood and old age mortality. Most links are made or confirmed in the household linking stage, but there is a higher proportion made by only individual probabilistic linking for men in their teens and twenties when the transition to a new household is most likely.

Widows and widowers are much less likely to be linked as a function of both age and, for women, remarriage. Marriage also explains the large deficit in links in Figure 2 for women as they entered their twenties. Just 22% of women aged 15 to 25 at the 1891 census are linked directly to 1901, compared with 47% of men. For married men and women match rates are nearly identical, reinforced by spouses linked together within households (Table B, Appendix). Direct census-to-census linking is successful in linking women up to or after marriage, but the transition between states requires additional information.

4 LINKING COUPLES TO MARRIAGES

Complete and detailed marriage records for England and Wales are not currently available on a large scale. Parish marriage records are more accessible but only include marriages in the Church of England and are decreasingly useful after the introduction of Civil Registration in 1837 and in areas of high non-conformity, including Derbyshire. This makes the process of linking women to their marriages much more complicated but not impossible.

Published indexes to 19th-century marriages have been scanned and transcribed by volunteers so that references for each marriage in the administrative county of Derbyshire can be selected for the period covered by the census data (Free UK Genealogy CIO, 2024).[5] However, the content of the indexes is limited: date of marriage is given as quarter of the year, ages are not recorded, individuals are listed alphabetically instead of with spouses, and women are listed only under their own (maiden) names. Each marriage entry is indexed to a page in the original registry book, where there are usually two couples (four individuals) per page. Thus, linking brides to grooms requires knowledge of the correct pairing of couples, which can be determined by reference to a later census entry based on the first and last names of the husband and the first name of his wife.

5 The volunteer-transcribed entries include flags for possible errors in index entries, including misspelled districts or register volumes outside the expected range. Before linking these are corrected, and errors in the transcription of pages of the register are replaced with missing values. The marriage indexes are also arranged by administrative county and district boundaries, which are not identical to the registration districts included in the census data.

The index transcriptions do not include a gender variable, which must be inferred from first name. It is possible to link to the index on name similarity alone, however not distinguishing between brides and grooms significantly increases the pool of potential matches and affects the choice of linking thresholds. For example, a name similarity threshold high enough to exclude false matches between Charles and Charlotte, with a Jaro-Winkler string similarity of 0.84, would exclude likely matches with differences in spelling or nicknames (e.g., Fanny and Fannie, J-W similarity 0.82). To address this problem, names in the marriage index are assigned a gender using the R package 'gender', which uses historical data on name frequency (Blevins & Mullen, 2015).[6]

Successful linking to the marriage index depends on defining the 'linkable population' of couples who married in the relevant decade, some of whom can be observed in the census-to-census links based on a change in the marital status of the husband. Compared with the approximately 28,000 marriages registered in Derbyshire between 1881 and 1891 (General Register Office, 1897), just over 11,000 men are linked from 'single' to 'married' in these censuses. Another 998 linked single men are classified as 'married, spouse absent' in 1891 and 296 appear to have been married and widowed in the same decade. Without information about their absent spouses, it is difficult to confidently link these last two groups to the correct marriage records.

The sample of potential grooms can be expanded to include unlinked married men in the second census aged 35 and under, representing the cohort likely to have married within the previous 10 years.[7] The wider net captures another 10,000 men, making up much of the deficit between linked and registered marriages. Two factors require caution when linking this expanded sample: there is no way to be sure that these men married in the same decade (except for a smaller number of younger married men) or that they married within the county. This is also a difficult life stage for record linking due to changing authorship of the original records; often a bride has shifted from a census entry filled out by her father to writing her own name in the marriage register to appearing in a post-marriage census entry filled out by her husband.

Linking to the indexes of marriage occurs in two phases. First, single women who are unlinked after the census-to-census stages are taken as potential brides and merged with possible entries in the index by name and district of marriage, under the assumption that women are most likely to marry near their homes. Index entries are then grouped with potential partners on the same page of the marriage register, and each of the possible pairings is compared to a newlywed couple in the next census (whether identified by census linking or by age cohort).

These groupings of single woman, bride, groom, and couple are compared based on husband's first name and surname, bride and wife's first names, and age in each census (Figure 3).[8] Potential links to marriages which imply ages at marriage under 15 years old are discarded. Experiments in manual linking suggested that discrepancies in brides' and wives' birthplaces caused low match rates so this variable is excluded in initial marriage index links in favour of focusing on matches by district of marriage. For example, if a wife had migrated as a child before meeting her future husband, he might assume she had been born in the county. After removing the first round of marriage links, the same process is then repeated for unlinked newlyweds without the restriction on district of marriage. Instead, single women and wives are additionally compared on county of birth.

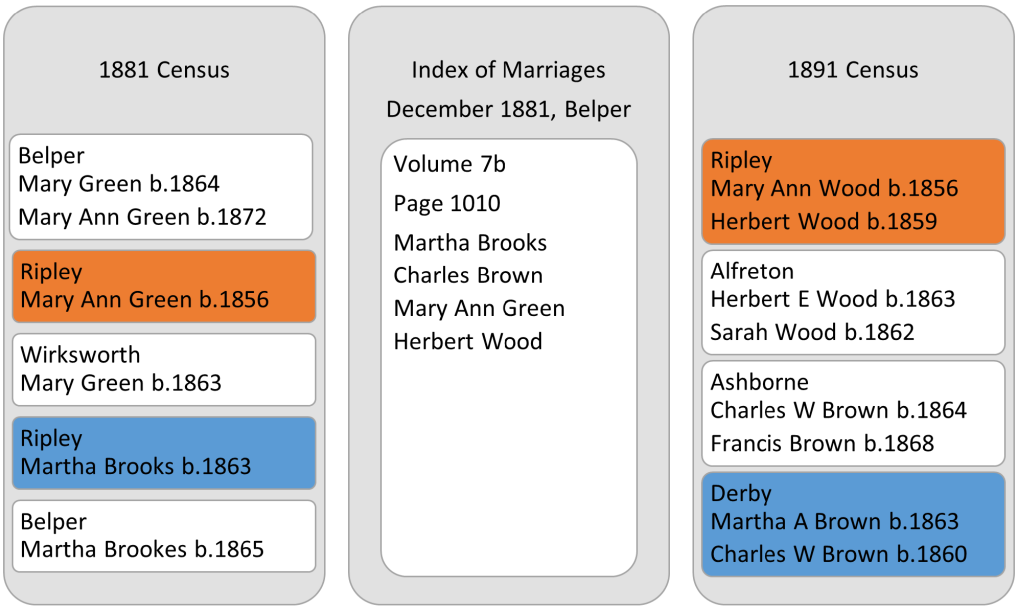
Of the couples who married between 1881 and 1891, 7,529 brides are linked across their marriages in the first stage and another 1,986 are linked in the second. This includes a few links to marriages registered in neighbouring districts which are outside the registration county borders but within the administrative borders (e.g., Mansfield, Rotherham, Worksop). These links imply short distance migration either before or after marriage (see discussion of migration and age at marriage below). The final successful links for women who married between 1881 and 1891 represent around 33% of the marriages registered in Derbyshire in the same decade.

6 This method successfully assigns a gender to 97% of names in the marriage index; the uncategorized entries (i.e., uncommon names, unusual spellings, or names which could belong to either gender) are included in both bride and groom pools.

7 This expanded cohort includes men who married anywhere between the ages of 15 and 35. This age range was chosen to centre on an average age at marriage around 25 years old.

8 The specific parameters are that first names have Jaro-Winkler similarities of at least 0.8, surnames have J-W similarities of at least 0.9, and single woman/wife birthyears be within three years. Competing potential links are selected based on name and age similarity and middle name agreement. If there is no unique best link based on these criteria, then all potential links for the individual are discarded.

Figure 3 *Example links to Index of Marriages (names pseudonymized)*



In total, 43,313 couples are identified by census-to-census linking who married in the period between 1881 and 1911. With the expanded sample by age cohort, 81,039 potential brides and grooms are included in the index linking process for these three decades and 36,334 wives are successfully linked to their pre-marriage census entries, or 44.8% of the total sample. This is probably an underestimate of the true linking efficiency for couples covered by the index of marriages given that an unknown number of the potential grooms were married outside the county.

4.1 QUALITY OF MARRIAGE INDEX LINKS

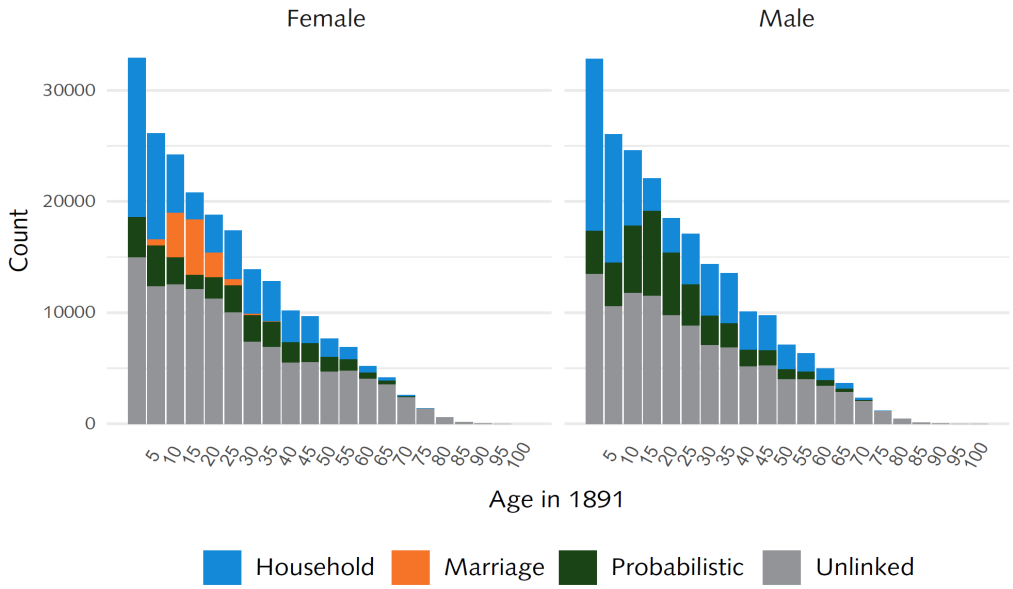
Linking women via indexes of marriage significantly reduces the gender linking gap by improving total match rates for single women to over 40% (Table B, Appendix). This improvement is mostly concentrated among women in their twenties, increasing match rates by 15 to 20 percentage points (Figure 4). The age profile of the total linked sample after linking via the marriage index is very similar for men and women, although match rates for women remain slightly lower overall.

Marriage links are influenced by migration in two ways: first, a skewed sex ratio among the Derbyshire population may have caused local grooms to seek brides elsewhere, as 22% of grooms linked in the census were born outside the county compared with 39% of their wives. In addition, the expanded selection of potential grooms includes men who moved into the county between censuses, with up to 70% reporting birthplaces outside Derbyshire. Only those who married after migration can be found in the index and their wives found in the previous census.

By selecting men based on changed marital status or age cohort, the linked sample is designed to capture first marriages for single, never-married women. The average age gap between spouses is identical to the average for all couples in the census (husbands 1.5 years older than wives). In addition, the dates of linked marriages closely follow trends in the dates of all marriage index entries for the period and distance from a census does not seem to influence the chances of being linked.

The average estimated ages at marriage for linked couples marrying in Derbyshire between 1881 and 1891 are just under 23 for women and 24 for men (Table C, Appendix). These averages are lower than singulate mean ages at marriage (SMAMs) calculated based on the proportions single by age in the census. These population estimates for women and men in Derbyshire in 1891 are 25.5 and 26.7, respectively (Reid et al., 2018). Given the migration patterns discussed earlier, it is not surprising that these sources disagree. SMAMs are vulnerable to movements of single and married people which affect the age structure of the census population. In the linked sample, non-migrants are also likely to marry earlier.

Figure 4 *Census links by age and method, including links via marriage index, 1891–1901*



The accuracy of marriage index links can be evaluated for 26,433 women who are linked to both the index of marriages and to the 1911 census when they were asked to report the duration of their current marriage. This sample only includes women who did not move outside the county and were still married at the time of the census. Age in the linked sample is calculated from age at the nearest census and year of marriage, while marriage dates in the 1911 census are based on retrospectively reported marriage durations. These ages are both approximate (due to the imprecision of marriage dates and birth dates) and subject to misreporting or mistranscription.

Average ages at marriage for each these three decades as implied by ages and marital durations reported in the 1911 census are not as high as the SMAMs calculated from population distributions, but the index links still represent a slight underestimate (Table C, Appendix). Some of the women recorded in 1911 were in second marriages, which are not included in the linking process for single, never-married women. Comparing these estimates at the individual level, the vast majority of reported marriage durations agree with the estimated ages at marriage obtained by record linking: 96% of the estimates from both sources agree within three years and 98% within five years (Figure A, Appendix). The combination of links to the marriage index and subsequent census linking up to 1911 thus seems to be successful in reconstructing individual life courses, but the resulting linked sample does not reflect the full range of marriage experiences in the population.

4.2 FULL LINKED DATASET

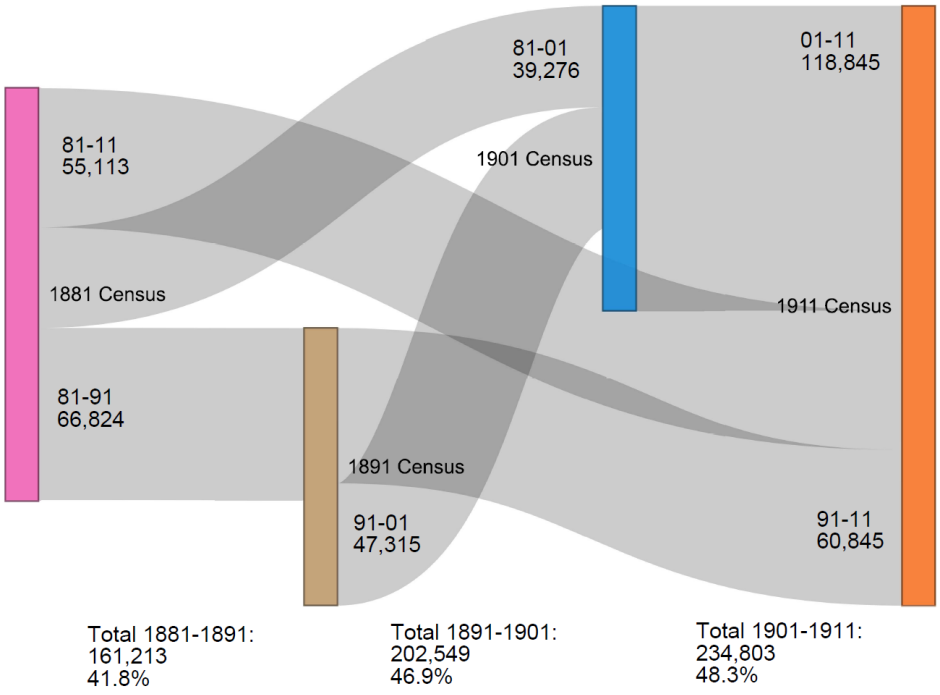
Having completed separate linking processes on all census years 1881 to 1911, it is possible to combine linked panels and follow individuals over more than one decade. Incorporating links via marriage indexes improves match rates over multiple decades by up to five percentage points. In total there are 94,389 individuals who are observed in at least the first three linked censuses, or 35.9% of the population aged 20 and over in 1901. Match rates improve over the final three decades, and there are 115,958 individuals with links from 1891 to 1901 and 1911, or 37.4% of the linkable population in 1911.

Combining these groups of links requires that individuals are present in each consecutive census, although linking between non-consecutive censuses might obtain more matches for individuals who left and re-entered the county. However, a greater time span also carries the risk of more false links, particularly since linking to the marriage indexes over this time span is more difficult. For the entire period, 55,113 people are identified in every census from 1881 to 1911, or 25.5% of the 1911 population over 30 (Figure 5). This group is skewed towards links for men, representing 57% of the whole-period links, but includes 11,880 women who married in one of these three decades of observation.

Without the inclusion of indexes of marriage, following so many women over such an extended period would be nearly impossible. The combination of all census and index links reveals that most of the women linked across marriages are traced over several decades — just 26% are only observed in the

censuses immediately before and after marriage. Out of 71,267 women who are linked across more than one decade, 26,601 of these life courses include the date and location of marriage. This pattern is driven by the strength of combining individual and household linking, since women are likely to be linked in conjunction with their parents before marriage and with their husbands after marriage.

Figure 5 *Decadal census links by duration of observation from first to last linked census entry, 1881–1911*



5 CONCLUSIONS

This paper has aimed to build on the existing record linking literature by adapting several methods to a new population and data source, including both census data and indexed civil registration records. Unlike many other automated linking projects, this methodology includes women at every stage. A multi-step process of individual, household, and marital linking allows for several checks on the accuracy of linked records and improves total coverage of the linked dataset by reducing the gender gap in links. These decadal links also combine to form significant cohorts which can be observed over more than one decade.

The use of open access, transcribed marriage indexes on this scale, overcoming the challenges posed by the limited information available for each couple, is a key proof of concept for future efforts to link larger populations. Rather than forming part of a 'one-size-fits-all' record linking paradigm, the results also encourage awareness of the underlying population dynamics (enumeration practices, life-cycle migration patterns, naming conventions, etc.) which influence both total match rates and possible sources of bias. The focus here has been on the performance of linking methods for women throughout the life course rather than on correcting for migration or assessing patterns of error.

Ongoing research using these data explores the relationships between women's work, marriage, and fertility in a life course perspective which has previously been difficult to achieve with available sources in the UK. Cross-sectional studies of occupational and class differences in fertility decline using census data almost always focus on classification by husband's occupation rather than wife's, and these gender-skewed models tend to erase women's agency (Mackinnon, 1995). Rather than focusing on current occupation, future models can include pre-marital work and occupational mobility for both women and men. Rather than measuring lifetime migration, researchers will be able to explore short-term movement in relation to marriage, work, and family structure. Rather than modelling only total fertility, linked data can be used to reconstruct childbearing trajectories and explore the relative impacts of age at marriage, birth spacing, and parity-specific stopping.

Effective characterization of the demographic transition in 19th-century Derbyshire will depend on the quality and coverage of links between sources and over time created through the processes detailed above. Despite data restrictions and methodological obstacles, this project is an important movement towards viewing women in historical demography as individuals with both a past and a future.

REFERENCES

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

Abramitzky, R., Boustan, L., Jácome, E., & Pérez, S. (2021). Intergenerational mobility of immigrants in the United States over two centuries. *American Economic Review*, *111*(2), 580–608. https://doi.org/10.1257/aer.20191586

Abramitzky, R., Mill, R., & Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *53*(2), 94–111. https://doi.org/10.1080/01615440.2018.1543034

Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben, C., & Williamson, L. (2020). Linking Scottish vital event records using family groups. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *53*(2), 130–146. https://doi.org/10.1080/01615440.2019.1571466

Antonie, L., Inwood, K., Minns, C., & Summerfield, F. (2020). Selection bias encountered in the systematic linking of historical census records. *Social Science History*, *44*(3), 555–570. https://doi.org/10.1017/ssh.2020.15

Arthi, V., Beach, B., & Hanlon, W. W. (2022). Recessions, mortality, and migration bias: Evidence from the Lancashire Cotton Famine. *American Economic Journal: Applied Economics*, *14*(2), 228–255. https://doi.org/10.1257/app.20190131

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

Bailey, M., Lin, P. Z., Mohammed, A. R. S., Mohnen, P., Murray, J., Zhang, M., & Prettyman, A. (2023). The creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database project. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *56*(3), 138–159. https://doi.org/10.1080/01615440.2023.2239699

Blevins, C., & Mullen, L. (2015). Jane, John… Leslie? A historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, *9*(3). https://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

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

Fitton, R. S., & Wadsworth, A. P. (1958). *The Strutts and the Arkwrights, 1758–1830: A study of the early factory system*. Manchester University Press.

Free UK Genealogy CIO. (2024). *FreeBMD*. https://www.freebmd.org.uk/

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

Garrett, E. (1990). The trials of labour: Motherhood versus employment in a nineteenth-century textile centre. *Continuity and Change*, *5*(1), 121–154. https://doi.org/10.1017/S0268416000003908

Garrett, E., & Reid, A. (2018). Composing a national picture from local scenes: New and future insights into the fertility transition. *Local Population Studies*, *100*(1), 60–76. http://www.localpopulationstudies.org.uk/wp-content/uploads/LPS-100-2018-GARRETT-and-REID-pp-60-76.pdf

Garrett, E., Reid, A., Schürer, K., & Szreter, S. (2001). *Changing family size in England and Wales: Place, class, and demography, 1891–1911*. Cambridge University Press.

General Register Office. (1897). *Supplement to the fifty-fifth annual report of the registrar-general of births, deaths, and marriages in England*. H.M. Stationery Office.

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

Higgs, E., & Wilkinson, A. (2016). Women, occupations and work in the Victorian censuses revisited. *History Workshop Journal*, *81*(1), 17–38. https://doi.org/10.1093/hwj/dbw001

Hill, B. (1993). Women, work and the census: A problem for historians of women. *History Workshop*, *35*(1), 78–94. https://doi.org/10.1093/HWJ/35.1.78

Hwang, S. I. M., & Squires, M. (2024). Linked samples and measurement error in historical US census data. *Explorations in Economic History*, *93*, Article 101579. https://doi.org/10.1016/j.eeh.2024.101579

Jaadla, H., Reid, A., Garrett, E., Schürer, K., & Day, J. (2020). Revisiting the fertility transition in England and Wales: The role of social class and migration. *Demography*, *57*(4), 1543–1569. https://doi.org/10.1007/s13524-020-00895-3

Long, J., & Ferrie, J. (2013). Intergenerational occupational mobility in Great Britain and the United States since 1850. *American Economic Review*, *103*(4), 1109–1137. https://doi.org/10.1257/aer.103.4.1109

Mackinnon, A. (1995). Were women present at the demographic transition? Questions from a feminist historian to historical demographers. *Gender & History*, *7*(2), 222–240. https://doi.org/10.1111/j.1468-0424.1995.tb00022.x

Massey, C. G. (2017). Playing with matches: An assessment of accuracy in linked historical data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *50*(3), 129–143. https://doi.org/10.1080/01615440.2017.1288598

McGeevor, S. (2014). How well did the nineteenth century census record women's 'regular' employment in England and Wales? A case study of Hertfordshire in 1851. *The History of the Family*, *19*(4), 489–512. https://doi.org/10.1080/1081602X.2014.968181

Newman, L., & Smith, H. (2023). *Linking Post Office superannuants to certification and registration of death: Sources and methodology* (Addressing Health Working Paper 3). https://doi.org/10.13140/RG.2.2.22319.51364

Ó Gráda, C., Anbinder, T., Connor, D., & Wegge, S. A. (2024). The problem of false positives in automated census linking: Nineteenth-century New York's Irish immigrants as a case study. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *56*(4), 240–259. https://doi.org/10.1080/01615440.2024.2312293

Reid, A., Davies, R., & Garrett, E. (2002). Nineteenth-century Scottish demography from linked censuses and civil registers: A 'sets of related individuals' approach. *History & Computing*, *14*(1–2), 61–86. https://doi.org/10.3366/hac.2002.14.1-2.61

Reid, A. M., Arulanantham, S. J., Day, J. D., Garrett, E. M., Jaadla, H., & Lucas-Smith, M. (2018). *Populations past: Atlas of Victorian and Edwardian population*. https://www.populationspast.org/

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

Schürer, K., & Higgs, E. (2020a). *Integrated Census Microdata (I-CeM), 1851-1911* (SN: 7481) [Dataset]. UK Data Service. https://doi.org/10.5255/UKDA-SN-7481-2

Schürer, K., & Higgs, E. (2020b). *Integrated Census Microdata (I-CeM) names and addresses, 1851–1911: Special licence access* (SN: 7856; Version 2nd Edition) [Dataset]. UK Data Service. https://doi.org/10.5255/UKDA-SN-7856-2

van der Laan, D. J. (2022). reclin2: A toolkit for record linkage and deduplication. *The R Journal*, *14*(2), 325–333. https://doi.org/10.32614/RJ-2022-038

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

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *47*(3), 138–151. https://doi.org/10.1080/01615440.2014.913967

Woods, R. (2000). *The demography of Victorian England and Wales*. Cambridge University Press.

APPENDIX REPRESENTATIVENESS OF LINKED DATA

Table A *Linked data for each decade 1881–1911 compared to census population distributions, match rates by county of birth*

County of birth (top 10)	1881 census	1881–1891	Match rate	1891 census	1891-1901	Match rate	1901 census	1901–1911	Match rate
Derbyshire	247,294	120,225	48.62%	264,608	142,995	54.04%	299,319	163,817	54.73%
Nottinghamshire	21,484	7,029	32.72%	21,185	8,116	38.31%	23,871	9,769	40.92%
Staffordshire	17,591	5,562	31.62%	18,900	7,206	38.13%	20,300	8,493	41.84%
Yorkshire	15,143	4,455	29.42%	17,528	6,359	36.28%	22,221	8,654	38.95%
Leicestershire	13,150	5,211	39.63%	13,955	6,054	43.38%	14,051	6,596	46.94%
Cheshire	9,619	3,210	33.37%	9,756	3,719	38.12%	10,063	4,148	41.22%
Lancashire	9,003	2,291	25.45%	10,263	3,184	31.02%	11,874	3,914	32.96%
Unknown; not coded	8,033	1,281	15.95%	6,810	1,263	18.55%	7,188	1,111	15.46%
Worcestershire	4,528	1,307	28.86%	4,005	1,267	31.64%	5,751	2,431	42.27%
Lincolnshire	3,892	1,120	28.78%	4,873	1,606	32.96%	5,648	2,149	38.05%

Table B *Linked data, including links via index of marriages, for each decade 1881–1911 compared to census population distributions, match rates by sex and marital status*

Marital status	1881 census	1881–1891	Match rate	1891 census	1891–1901	Match rate	1901 census	1901–1911	Match rate
Female	193,043	75,851	39.3%	215,545	95,594	44.3%	244,362	111,472	45.6%
Divorced							9	1	11.1%
Married	63,452	27,248	42.9%	68,738	33,665	49.0%	80,251	41,704	52.0%
Married, spouse absent	3,655	244	6.7%	5,559	382	6.9%	6,323	449	7.1%
Single (with index)	113,631	37,672 (47,187)	33.2 (41.5%)	127,197	47,431 (59,934)	37.3% (47.1%)	142,295	52,968 (67,284)	37.2% (47.3%)
Unknown	181	8	4.4%	439	9	2.1%	300	6	2.0%
Widowed	12,124	1,164	9.6%	13,612	1,604	11.8%	15,184	2,028	13.4%
Male	192,898	85,362	44.3%	215,172	106,940	49.7%	240,717	123,331	51.2%
Divorced							9	7	77.8%
Married	63,562	27,399	43.1%	68,819	34,212	49.7%	80,377	42,329	52.7%
Married, spouse absent	3,286	835	25.4%	5,273	1,638	31.1%	5,581	1,885	33.8%
Single	118,899	55,551	46.7%	133,425	69,064	51.8%	146,285	76,831	52.5%
Unknown	277	52	18.8%	423	103	24.3%	292	90	30.8%
Widowed	6,874	1,525	22.2%	7,232	1,923	26.6%	8,173	2,189	26.8%
Total	385,947	161,213	41.8%	432,126	202,549	46.9%	486,316	234,803	48.3%

Table C *Average age at marriage calculated from links to index of marriages compared to average age at marriage implied by reported age and duration of marriage in the 1911 census, by decade of marriage and district of marriage or enumeration*

Average age at marriage	1881				1891				1901			
	Index entry	Census entry	N index	N census	Index entry	Census entry	N index	N census	Index entry	Census entry	N index	N census
Ashborne	22.9	24.7	358	540	23.4	26.0	442	794	24.4	26.7	425	1,057
Bakewell	23.4	24.8	669	925	23.9	25.6	737	1,387	24.0	26.6	826	2,067
Belper	22.0	23.5	1,485	1,710	22.3	24.5	1,855	2,808	22.8	25.4	2,012	4,256
Chapel-en-le-frith	23.3	24.0	370	839	23.7	25.2	490	1,331	23.6	26.3	513	1,794
Chesterfield	21.4	22.7	2,215	3,368	21.6	23.8	3,263	5,816	22.1	24.6	4,004	9,185
Derby	22.0	23.6	2,345	3,233	22.5	24.3	3,018	5,218	23.1	25.5	3,374	7,451
Hayfield	23.0	24.4	702	535	23.8	25.4	696	847	24.6	26.7	286	1,288
Shardlow	22.3	23.3	835	1,777	22.7	24.7	1,198	2,805	23.1	25.5	1,536	4,154
Total	22.5	23.9	8,979	12,927	23.0	24.9	11,699	21,006	23.5	25.9	12,976	31,252
Total including other districts	22.8	23.9	9,515	13,723	23.3	25.0	12,503	22,096	23.9	26.0	14,316	32,918

Figure A *Comparison of ages at marriage calculated from age and duration of marriage reported in the 1911 census and from year of linked index entry and age at nearest census. Reference lines indicating minimum linking threshold and age agreement +/- five years. Data grouped by decade of marriage as calculated from duration reported in the 1911 census.*