

UNIVERSITY OF CAMBRIDGE
Faculty of History

Cambridge Working Papers in Economic and Social History

A First Full-Count Linking of English and Welsh Censuses, 1851-1921

Guillaume Proffit, Alexis Litvine, Emma Diduch, Emily Chung, and Robin Linacre

October 27, 2026

The Cambridge Working Papers in Economic and Social History provides an avenue for members of the university to share their research. It showcases unpublished research of Cambridge economic and social historians, including academic staff, students, invited speakers and visitors.

If you would like to include your work in the series, please send it to the editors.

Learn more about Cambridge's Research in Economic and Social History here.

A First Full-Count Linking of English and Welsh Censuses, 1851–1921

Guillaume Proffit*[1], Alexis Litvine†[1], Emma Diduch‡[1], Emily Chung§[1], and Robin Linacre¶[2]

[1]CAMPOP, University of Cambridge
[2]UK Ministry of Justice

*Email: gpap3@cam.ac.uk
†Email: adl38@cam.ac.uk
‡Email: egd28@cam.ac.uk
§Email: evc28@cam.ac.uk
¶Email: robin.linacre@justice.gov.uk

Abstract

Historical census linking provides researchers with evidence for finer-grained studies of population and economy, turning cross-sectional observations into longitudinal data by tracking individuals and their changing characteristics across censuses. In this paper, we adapt an unsupervised Bayesian census linking method to British censuses between 1851 and 1921. We split the process into several stages, which allows us to reduce the amount of blocking required and minimise false negatives. We also combine immutable and mutable characteristics to disambiguate cases with a high level of similarity. Finally, by incorporating the entire set of marriage civil registration indexes, we are able to substantially reduce male skew in the linked population. We provide researchers with the first full-count linking for England and Wales 1851–1921 together with several verification exercises and a description of biases in the linked data.

Key Words: historical censuses; record linkage; longitudinal data; England and Wales; population history.

INTRODUCTION

Since 2014, the Integrated-Census Microdata (I-CeM) project has published datasets derived from manual transcriptions of British censuses at the individual level, beginning with the 1851 census and now extending to 1921 (Schürer, Wakelam, and FINDMYPAST LIMITED 2024; Schürer, Higgs, and FINDMYPAST LIMITED 2024). Researchers can now easily access information on occupation, family structure, place of enumeration, and demographic data for all available censuses.[1] I-CeM also provides standardised coding of occupations, birthplaces, household structures, and consistent geographic identifiers (Schürer 2024). However, this resource provides only snapshots of the English and Welsh (and Scottish, for certain years) populations at decennial intervals; even though the vast majority of people enumerated in one census are present in another, there is no consistent identifier to track individuals over time. In this paper, we introduce a new method to provide the first full-count linking of English and Welsh censuses using I-CeM data, opening new possibilities for longitudinal studies.

Census linking is not a new endeavour. Researchers have long aspired to create longitudinal data from censuses (Malin 1935; Thernstrom 1964), but since the late 1990s, with the increasing availability of both computing power and the release of digitised census data, the rate of methodological innovations in census linking has intensified (Ferrie 1996; Abramitzky, Boustan, and Eriksson 2012; Ruggles, Fitch, and Roberts 2018). Our paper inherits from much of this work.

Intuitively, census linking can appear as a relatively simple, albeit time-consuming, process. Omitting cases where individuals were not yet born, had died, or migrated abroad between census dates, one can reasonably assume that any given individual possessing distinctive characteristics should be identifiable in the next or previous census using these unique features.

Fundamentally, the issue is that identifying the same person in multiple sources, even on a small scale, is far from being self-evident. In the absence of persistent unique identifiers for individuals, census linking must entirely rely on the reported characteristics of individuals at each census point to locate their future or past selves, amongst the 20 to 30 million other individuals found in each census. Two major conceptual issues arise from this: data **variance** and **commonality**.

Variance. Working with historical censuses, it becomes rapidly clear that the observed characteristics of a single person can vary significantly between censuses. Characteristics reported in a census fall into two groups: *immutable* characteristics (names, birthplace and implied year of birth), which are not expected to change over time,[2] and *mutable* characteristics (occupation, co-residents, place of residence) which can be expected to change in the decade between two censuses.

The first source of variance is changes amongst mutable characteristics: a change in occupation, marriage and formation of a new household, migration, etc. The second source of variance is errors in the recording of characteristics; this can affect both mutable and immutable characteristics (although for mutable characteristics it is harder to distinguish an error from genuine change). These errors can be introduced at various stages. First, at enumeration, information can be misreported through ignorance or simplification: ages can be rounded to the nearest value ending in 5 or 0 (a phenomenon referred to as age heaping), and a distant rural birthplace can

1. With the exception of 1871, which has not yet been released.
2. One exception to this rule is women's surnames which conventionally are changed at marriage in many countries.

be simplified as the nearest large town (Day 2022). These phenomena can be compounded by a misunderstanding of the enumeration instructions (in a period where literacy was not universal) or by the desire to hide some family secrets (marital status or the nature of household relationships). Errors can also be introduced by the enumerator who, until 1911, copied the information written in the schedules by the head of household into the Census Enumerators' Books (CEBs). Furthermore, I-CeM itself relies on the recent manual transcription of these enumerators' books (and household schedules for 1911 and 1921), which introduces another source of potential error from either outsourced transcribers (generally unfamiliar with the source or nineteenth-century palaeography) or automatic handwritten text recognition (HTR) algorithms, not to mention cases of poorly legible squiggles, damaged documents, or challenging tabular layouts.

Commonality. The inverse phenomenon can be equally problematic: different individuals can share many characteristics, making them almost indistinguishable. For example, there were 26,749 men named William Smith in the 1911 census. Worse yet, even when looking to other identifying characteristics, there were still 99 boys named William Jones aged 16 born in Glamorganshire, and 29 William Jones aged 20 born in the parish of Walton-on-the-Hill (Cheshire). The degree of commonality in nominal and other characteristics is therefore a significant hurdle to achieving high-accuracy links. Identifying the correct William Jones, if at all possible, would require widening the set of linking criteria used, or careful analysis of ancillary records to complement the census. This issue is by no means an insignificant phenomenon. In 1911, 24.68% of men's full names were not unique in their county of birth and 4.34% of names were not unique in their specific place (in most cases, parish) of birth.

Automating census linking therefore involves defining a protocol to distinguish and quantify what constitutes acceptable intercensal variation from what characterises genuine difference in a pair of compared individuals. This is formally expressed as the statistical measure of confidence in one or more links suggested for each individual. Regarding commonality, in many cases, several similarly probable matches may be found. A census linking pipeline must thus decide when two plausible links for a given person are considered so close as to be confounded, and hence that person to be unlinkable.

In an effort to minimise such cases where we cannot distinguish between possible matches, our work adopts a wider array of linking criteria than has been used in the literature so far. The first generation of automated census linking methods relied only on immutable characteristics that remain stable throughout life (Ferrie 2005; Abramitzky, Boustan, and Eriksson 2012) and, consequently, suffered more from data commonality issues. A recent strand in the literature now takes into account a wider range of characteristics to provide their linking approach with more separation criteria (Antonie et al. 2020; Helgertz et al. 2022; Buckles et al. 2025; Abramitzky et al. 2025). We believe that the risks of inducing bias and error in the linking by using additional information, including mutable characteristics, are counteracted by the greater recall they enable, and thus adopt this approach.

Our linking workflow is therefore made of two broad phases: first, we introduce a foundational method to link British censuses based on extensive pre-processing and enrichment of the census data. We combine mutable and immutable individual characteristics to establish, first, high-confidence links for both men and women in adjacent censuses. With these initial easy-to-establish links established for a proportion of men and women, we next link men and women across the (now much reduced) population using only immutable characteristics to complement the initial links created. Overall, this method allows us to link 52.81% of the individuals in adjacent censuses with a high degree of confidence. We then build on this foundational method

by incorporating full-count indexes to the Marriage Register, adapting the method developed by Diduch (2024). With these indexes, we link married women across censuses (and across surname changes) by linking them and their husbands to adjacent register entries, allowing us to establish more links for this population. In doing so, we believe this to be the first use of full-count registration data for census linking. We achieve a final linking rate of 55.62%, with an error rate of 6.34% on a binary review of the census-to-census links.

At scale, census linking involves an effort to maximise both precision (measured as the proportion of true links in all links) and recall (the proportion of all linkable individuals linked across censuses), while quantifying as precisely as possible any bias introduced by the linking method. The latter is fundamental, since all subsequent analyses will be impacted by bias in either the accuracy or the representativeness of the linked population. The tradeoffs between those three metrics have been a primary differentiator between linking methods in the last twenty years (Ruggles, Fitch, and Roberts 2018). False links (low precision) have the strongest negative effects on downstream analysis due to both potential attenuation bias because of excess noise in the data (Mattheis 2024; Gross and Mueller-Smith 2021; Bailey et al. 2020) and the fact that false links are less directly observable than missed links. A lower recall rate is unlikely to prevent analysis, as sample sizes can remain very high when working with full-count censuses, as long as sample bias is controlled. In designing our methodology, we thus prioritised precision above recall, while adopting an empirical and descriptive approach to measuring bias in our results (Solon 1992; Ruggles, Fitch, and Roberts 2018).

The rest of this article is organised as follows. **Section 1** reviews the state of the art in census linking and describes our choice of linking variables. **Section 2** details our linking strategy and describes the models created. **Section 3** details the linking of civil registration index data. Finally, **Section 4** describes our verification strategy, and **Section 5** presents the results of the model and discusses its limitations.

1 LITERATURE

The I-CeM project has released full-count transcriptions of British censuses covering the period 1851–1921, with the exception of 1871.[3] However, heretofore only a handful of published works have attempted to link these censuses, and these have generally made use of subsamples for local case studies (Long and Ferrie 2013; Diduch 2024; Newton and Bennett 2020; Reid, Davies, and Garrett 2002; Pérez 2019). This contrasts with the situation in the US, Canada, and some Scandinavian countries where major improvements in full-count census linking have been made over the last fifteen years. We therefore adopt methods described in this literature as a guide to linking British historical censuses. This section briefly summarises recent advances in the field and offers key takeaways for our work.

Until the mid-2010s, census linking relied on deterministic methods, covering increasingly large datasets (Ferrie 1996, 2005; Abramitzky, Boustan, and Eriksson 2012, 2014; Collins and Wanamaker 2015; Beach et al. 2016; Abramitzky, Boustan, and Eriksson 2017). A deterministic approach links individuals using fixed rules based on relatively simple similarity thresholds for pairs of census records. However, in the last decade, deterministic approaches have been superseded by both probabilistic and Machine Learning-based (ML) methods. It is generally agreed that these have a greater capacity to distinguish patterns of differences between pairs

3. Although a full release of 1871 I-CeM is in the making.

Figure 1: **Precision and recall of different census-linking methods**

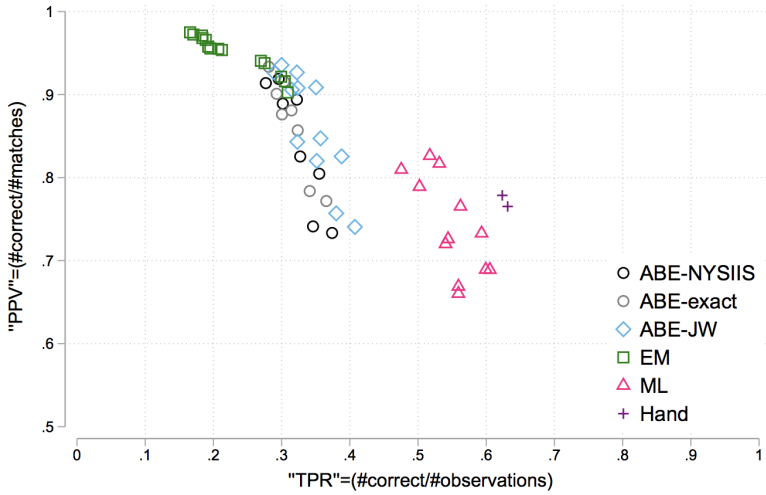


Figure adapted from Abramitzky et al. (2021)

and achieve higher linking and lower error rates than deterministic approaches (Abramitzky et al. 2021; Ó Gráda et al. 2023). These methods work by assigning coefficients to degrees of similarity and/or difference between all pairs considered. The main difference is that ML-based methods operate by learning to reproduce the choices made by humans, using a training dataset, while probabilistic methods use statistical approaches to derive coefficients from the distribution of dissimilarities in the observed data.[4] These two methods constitute the current performance frontier in census linking.

The performance of linking methods can be measured by combining two metrics: precision (the share of "true" pairs among all identified pairs) and recall (the share of true pairs identified out of all possible matches). A method that obtains many correct matches (true positives) but also many false matches (false positives) will have high recall but low precision. Conversely, a conservative approach, which rejects all dubious matches, will have high precision but low recall. Abramitzky et al. (2021) find that probabilistic approaches tend to have higher precision and ML methods higher recall scores; we reproduce here their graph (with precision on the vertical axis and recall on the horizontal axis).[5] Deterministic, ML and probabilistic methods all struggle with linking women, in part because of recording biases against women (as male heads of household most often completed the census form), and largely because of surname changes occurring at marriage. Bailey et al. (2023) and Diduch (2024) have shown that the use of registration data can reduce this bias, but this has not yet been applied on a full count linking with census data.

In this article, we decided against adopting an ML approach because of issues regarding

4. ML methods can also be described as probabilistic, but for the sake of clarity we use the latter term to refer to methods that do not require training data.

5. Because of the various methods in which these two measures are established, Figure 1 does not strictly compare like for like, and both axes should be read as proxies for their respective dimensions.

training data. Unfortunately, a large crowdsourced genealogical dataset such as the Census Tree (Price et al. 2021) is not currently available in the UK. Training a model on poorly documented genealogical data subsamples also increases the risk of incorporating errors and biases. Thus, working with crowdsourced genealogical data requires careful treatment (Chong et al. 2022; Stelter and Alburez-Gutierrez 2022; Calderón-Bernal, Alburez-Gutierrez, and Zagheni 2023; Colasurdo and Omenti 2024; Hollingsworth 1976). Feigenbaum, Helgertz, and Price (2025) have shown that it is possible to train capable linking ML models on relatively small (hence inexpensive to create) datasets, but this remains a sizeable effort, outside of the scope of this paper. Here, we instead adopt an unsupervised probabilistic approach coming from the field of Entity Resolution, the Fellegi-Sunter model with Expectation Maximisation (FSEM). This method has the substantial advantage of being highly interpretable. This section describes the model and our main design choices for the linking pipeline.

Fellegi-Sunter with Expectation Maximisation models

This model was designed in the 1960s for probabilistic record linking in contexts featuring fuzzy data (Fellegi and Sunter 1969), and has already been used for historical census linking (Abramitzky, Mill, and Pérez 2020). Abramitzky et al. (2021) have shown that FSEM models tend towards higher precision and a slightly lower recall, when compared to other deterministic and ML methods (as shown by the location of the green square 'EM' results in the upper left part of **Figure 1**). The key benefit of this approach is that it is 'unsupervised', meaning that the algorithm derives the relative importance of each linking criterion (referred to as the weights) directly from the censuses, without consuming training data. An added benefit of FSEM models is that they are composed of relatively few weights (compared to ML methods), allowing for sophistication in the comparison, whilst retaining a very high degree of interpretability for the researcher. For this article, we adopted Splink, a Python package developed at the UK Ministry of Justice that implements FSEM models to link administrative sources (Linacre et al. 2022). Splink has previously been used in other research contexts (Kates et al. 2025), although never for historical census linking.

The core idea behind FSEM models is that certain characteristics contribute more significantly to the overall probability of matching than others, following a Bayesian approach.[6] Intuitively, an exact match on year of birth provides stronger evidence towards a link than a match on gender alone. In FSEM, the relative importance of a given linking criterion is captured by its *partial match weight*, which indicates the importance of each variable to the final confidence score. For each linking criterion (surname, age, place of birth), the researcher defines arbitrary levels of comparison, representing degrees of similarity between two values, ranging from an exact match to a complete mismatch; each level of comparison has a partial match weight attached, representing the confidence it provides in an overall match. For example, when comparing birthplaces, the comparison levels could be:

- a match on the parish of birth
- a match on the county of birth
- a match on the country of birth

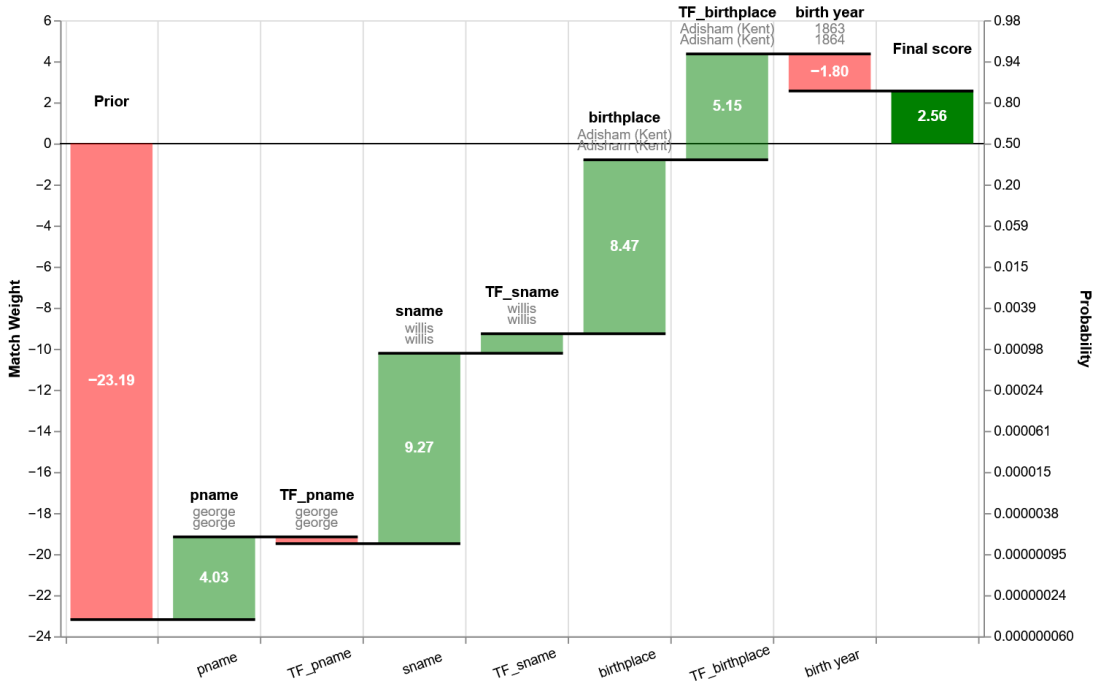
6. For a mathematical approach to FSEM, see Abramitzky, Mill, and Pérez (2020), Enamorado, Fifield, and Imai (2019), and Robin Linacre (2021).

- a complete mismatch

Each level would have monotonically decreasing partial match weights from a strong positive value (for a parish match) to a strongly negative value (for a complete mismatch).

At scale, the model compares pairs of records (or individuals) and calculates the probability that they represent the same person. A pair is initiated with a probability λ, equal to the probability that a random pair is a true match (an extremely small number). The model then compares each variable across the record pair, such as first name, age, and place of birth. Within each comparison, the pair is assigned to a level of comparison depending on which criteria it meets. Each level of comparison is associated with an m and a u probability: m being the probability that a truly matching pair falls into that comparison level, and u being the probability that a truly non-matching pair falls into that same level. The ratio between these two probabilities gives the partial match weight contributed by that comparison level. [7] When these weights are combined across all variables and with the prior match probability λ, they give an overall measure of the similarity of the two records, also called the Bayes Factor, which indicates the odds of a record pair being a true match.[8]

Figure 2: **Match weight waterfall chart for pair scoring using the FSEM model**



7. More specifically, the ratio of m over$u$ gives the Bayes factor, whose log2 provides the partial match weight.

8. Rigorously, these are not strictly probabilities as aspects of our model violate the FSEM assumption of non-collinearity among comparisons. We use the term as a shorthand. However, it should be said that the non-collinearity assumption is systematically broken in practical applications of FSEM, without compromising the feasibility and quality of the linking.

Figure 2 allows us to visualise the scoring of a pair with the FSEM model. From left to right, we can see the value of all partial weights included in the model and their contribution to the final confidence score, indicated on the top right bar. On the extreme left, the value λ (-23.19) indicates the probability (strictly speaking, the prior log-odds weight) that these two pairs, ex-nihilo, happen to be true matches. Each of the following bars shows the scores derived from the partial match weights (in this example, first name, surname, birth place, and age). Some comparison levels (names, occupation, and birth place) also add corrective partial match weights (Term Frequency adjustments, henceforth TF) to account for the frequency of specific values: a negative TF adjustment means that a specific value is more common than average, and hence the added confidence that agreement on this value contributes to the matching probability is moderated downwards. Equally, a match on a rare value (such as a pair agreeing on a very uncommon birthplace, like Adisham in Kent in this example) leads to a commensurate upwards adjustment. Aggregated, they give a final match weight (in this case 2.56) that indicates the likelihood of this pair being a true pair. Here, the final weight is positive, indicating that the two records in adjacent censuses are likely (86%) to represent the same entity (person). Depending on the context, varying thresholds can be applied to accept or reject the pair.

A further crucial element of the FSEM model is its unsupervised nature. Our choice of this method was guided by the model's ability to derive its weights (its m and u values) directly from the data, without requiring training/calibration data. In truth, we could observe m and u directly over a labelled pair dataset, but since the lack of training data is precisely the starting point of this paper, we use two alternative methods to estimate m and u values.

We first estimate u values (the probability that a truly non matching pair would fall in a given level of comparison) by creating a large selection (1bn) of random pairs. At such a sample size, the distribution of random pairs amongst levels of comparison is a reliable proxy for the distribution of non-matching pairs and we can thus directly measure u for each level of comparison.

Estimating m values is more challenging and requires us to use the Expectation-Maximisation (EM) procedure. Working over a given set of pairs containing both matches and non matches, the EM algorithm initiates default m values, and uses the u values that were estimated by the random pair method. The EM algorithm then iteratively refines the m values until the changes are small enough that the values are considered to have converged. Providing the set of pairs on which EM is run is an important design choice: the researcher defines the blocking rules that require a match on certain characteristics (such as first name and place of birth), which restricts the sample provided to the model and gives it a certain density of truly matching pairs inside the set. However, this also means that EM can only estimate the m values for the non-blocked values (for example surname and age). The EM process is thus run many times with various sets of data to achieve coverage of every comparison, and ideally provide different estimates, allowing the final m values to be drawn from averages of several EM sessions. Our approach has been to run a training session for each combination of blocking criteria that leaves 1, 2 and 3 comparisons unblocked and hence trainable.[9]

The final parameter in the model is λ, the initial probability that a random pair of records is a true match. In our case, $\lambda = 1/MatchesExpected$.[10] We estimate the expected number of

9. We do exclude some rare rules with unhealthy convergence.

10. This follows the intuition that, in a near one-to-one linking of two similarly sized tables, the probability that two randomly selected records form a true match is approximately the reciprocal of the number of expected matches.

matches as 90% of individuals aged 10 or older in the subsequent census. Although this is a somewhat arbitrary approach (Long and Ferrie 2013), λ is mostly a scaling factor in the model, and we show that it has little impact on our results. Increasing the expected matching proportion from 40 % to 100 %, our overall linking rates vary by less than 2%. Furthermore, evaluating the values of λ with our final results, we conclude that our approach is rather conservative since we assumed that all individuals aged 10 and above could be linked backward, which is not possible due to blocking or under-enumeration. If we instead set λ to $\lambda = 1/LinksMade$, the value of λ would be much higher (about double, hence making linking relatively easier).

To our knowledge, there is no literature that discusses the best approach for EM training (Linacre 2026). However, the high degree of stability among the m values and the rapid convergence of the algorithm appear to indicate that this approach leads to a suitable estimation of both the m and u parameters; this is further evidenced by the quality of the links produced (as shown by our verification work).

Choice of variables and blocking strategy

A key debate in the census linking literature pertains to the choice of characteristics for linking. During the last two decades, most linking methods have only used immutable characteristics such as names,[11] age, and birthplace (Ruggles, Fitch, and Roberts 2018). Indeed, including mutable characteristics poses two risks. First, it could lead to erroneous linking. A man named William Smith working as an agricultural labourer in 1851 could for example be linked to the wrong William Smith in 1861 if the former had changed occupation during this decade, while another William Smith was recorded as agricultural labourer in 1861. Secondly, mutable characteristics can increase the bias in the linked sample. Using occupation as a linking criterion, for example, will tend to create more links for men than women as male occupations are reported more frequently. Similarly, using household information will bias the data in favour of individuals who remained in the same household (or who kept the same occupation). Recent publications have challenged this, showing that the use of mutable characteristics can be beneficial, because the higher linking rate achieved more than offsets the potential biases introduced by non-universal criteria (Fu et al. 2014; Antonie et al. 2020; Rijpma, Cilliers, and Fourie 2020; Helgertz et al. 2022; Abramitzky et al. 2025). This is because the careful use of mutable characteristics, such as household composition and occupation, can distinguish between potential links that would otherwise be confounded. Our work follows this approach.

A second key component of the census linking literature is the degree of "blocking" required. Comparing all pairs of records for datasets containing more than 30 million individuals would be equivalent to computing a matrix of 9×10^{14} elements, which is computationally implausible. Some restrictions on the number of potential candidates (the subsets to be compared) are therefore necessary. Limiting the comparison to pairs already sharing some commonality – for example, only comparing men to men, or individuals who share a first name and age – is a simple way of reducing the complexity of the computation. However, any blocking necessarily introduces some false negatives, as the blocking characteristics themselves are subject to variance. In this paper, we minimise the degree of blocking required by adopting a multi-stage approach, which allows us to remove progressively high-certainty pairs and, therefore, relax blocking rules for the remaining candidates. The complete workflow is described in Section 2.1.

11. except for women across marriage

2 OUR CORE IMPLEMENTATION

A probabilistic linking model requires at least three design choices. The first pertains to the blocking strategy ('which subset of pairs will be considered for scoring?'), and the other two to the pair scoring itself: the comparisons ('which variables are included?') and levels of comparisons ('how are they compared?'). In this section, we outline our choices with regard to these three aspects.

2.1 Linking strategy

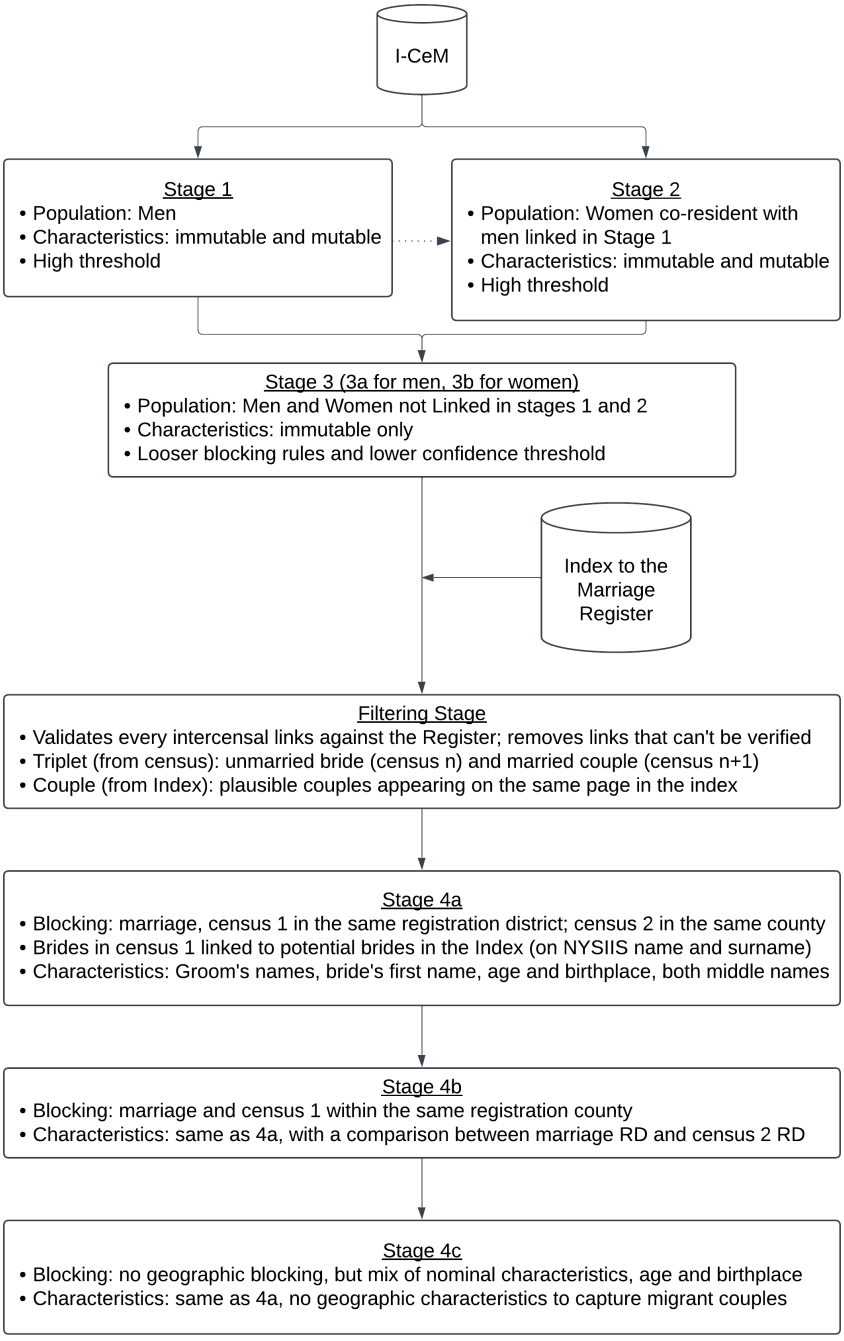
Our linking is separated across several stages in order to minimise the amount of blocking required and to use established links to compare co-residents (Helgertz et al. 2022). **Figure 3** shows the overall structure of our linking approach and the different stages and the order in which they take place.

In Stage 1, we begin by establishing high-confidence links among men (Helgertz et al. 2022; Diduch 2024), using both immutable characteristics (first name, surname, age, and birthplace) and mutable characteristics (occupation, head of household's first name, co-residents' first names, and residential coordinates) with a high threshold (probability > 0.9), to identify highly probable links and reduce downstream computational complexity. With this, we link and remove on average 43.16% of the total male population from the linking pool.

In Stage 2, we find links between women using both mutable and immutable characteristics (identical to Stage 1 characteristics, except for occupations and place of residence), comparing women who are co-resident with men linked in Stage 1. This is because co-residence is a reliable criterion to track individuals, especially for women who suffer from otherwise less well-reported characteristics, making a replication of the Stage 1 approach more difficult. This also allows us to establish high-confidence links for women, with the same aim of reducing computational complexity, linking and removing on average a further 23.13% of the female population.

In Stage 3, we attempt to link the remaining men and women separately, using only immutable characteristics (first name, surname, age, and birthplace). The reduced number of individuals left to link allows us to apply looser blocking rules (described in Section 2.2). In this case, since we only use immutable characteristics, we use a lower confidence threshold (0.8).

Figure 3: **Linking stages**



2.2 Blocking

In each of these stages, we adopt an open-ended, internally cumulative blocking strategy to maximise the number of pairs that are considered for linking and to reduce false negatives. By internally cumulative, we mean that in a given stage, a pair of individuals only has to meet one of the blocking rules to be considered for linking; the pairs yielded from each of the blocking rules are aggregated and scored together. Each stage has a large set of blocking rules which require records to match on a changing set of characteristics, allowing for more variance in the data. The cumulative design allows us to catch pairs for which a variable used in one of the blocking rules is incorrect. For example, a mismatching first name (for example, due to a misspelling) does not prevent a pair from being considered as it may fall under a different blocking rule which may only require a match on first name initial, or not require a match on first name at all. This is crucial. The Census Tree dataset shows for example that all blocking variables have a degree of variance among truly matching pairs, including the most stable ones, such as sex (Price et al. 2021). The specific rules we apply are detailed in the Appendix.

We adopt two more global blocking rules. First, we process men and women separately, which essentially functions as a blocking rule. Individuals with sex marked as unknown (less than 0.05% of the full I-CeM population) are duplicated to be considered both as men and women throughout the linking. When male and female links are aggregated, individuals with unknown sex can therefore end up being linked twice. In such cases, we retain the pair with the highest probability of linking. Secondly, we prevent individuals reported as married, widowed, or divorced (individuals ever married) in a given census from being compared to unmarried individuals (individuals never married) in the later census. This relies on the correct reporting and transcription of the underlying marital status and sex. Empirically, this approach seems robust.[12] These two rules offer a strong discriminating power, allowing for looser blocking elsewhere.

As **Table 1** shows, our blocking strategy is significantly less restrictive than other approaches in the literature when applied to British censuses, thus minimising the risk of false negatives. In this respect, our work is closer to the new generation of methods initiated by Helgertz et al. (2022) and Bailey et al. (2023). Taking as an example the 1881–1891 census pair, we make 4.977 billion comparisons in our first round. Using the household pairs created without any further blocking criteria, we compare another 219 million pairs of individuals. In Stage 3, we make 1.927 billion comparisons of men and 2.005 billion comparisons of women. We are confident that our cumulative blocking strategy allows us to make more plausible comparisons than the methods which apply a single indiscriminate blocking rule.

We also optimise the computing aspects of our census linking method in order to maximise the number of pairs that can be scored (our linking is executed in DuckDB, an in-memory SQL database, allowing for fast and memory-efficient execution). We optimised our linking for execution in an HPC environment with 1 TB of memory and 120 cores for processing.

2.3 Comparisons

Most of the literature restricts its approach to using the immutable characteristics reported in the census: first and last name, age, and place of birth. There are few examples of studies that

12. Price et al. (2021) shows that sex is the second most stable criterion in the Census Tree dataset, with a stability rate of 99.53%.

Table 1: **Comparison of blocking rules in the recent census linking literature applied to British censuses**

Method	Blocking rule	Pairs compared
This paper	Cumulative (see below)	9.128 bn
ABE (2012)	First and last name (NYSIIS) and age	558,392,762
AMP (2020)	Place of birth (US state, Norwegian municipalities), surname and last-name initials, and age < 5	1,622,104,262
Feigenbaum (2016)	State of birth; age difference < 3 yrs; Jaro–Winkler similarity of first and last name > 0.8	366,155,889
Buckles (2023)	First and last name (NYSIIS), age difference < 4, birthplace, sex, and race	1,449,218

As ABE, AMP and Feigenbaum only link men, we extrapolate a blocking rule by sex, reflecting their (implicit) gender criteria.

use mutable characteristics in automated linking (Fu et al. 2014; Antonie et al. 2020; Rijpma, Cilliers, and Fourie 2020; Helgertz et al. 2022; Abramitzky et al. 2025). Here, we use the wealth of information that can be drawn from mutable characteristics to narrow our matches and disambiguate (where possible) cases of over-commonality, not to disqualify plausible links where a mutable characteristic happened to change. FSEM's unsupervised estimation of weights confirm our intuition: the negative weights it derives for mismatches in mutable characteristics are indeed very low, whilst positive weights for matches on these mutable characteristics are healthy, allowing a reinforcement in pair confidence in Stages 1 and 2. Furthermore, our use of these mutable characteristics is associated with a higher confidence threshold required to accept a pair.

In the following, we describe our comparisons, including the enrichment brought to the data and the specific levels of comparison. The levels of comparison are given in decreasing implied levels of similarity, as they are applied in FSEM. That is, the third level of comparison can only be met if the first two levels of comparison (which imply greater similarity) have not been met. We start by discussing immutable characteristics and then move on to mutable ones.

Surnames

We compare surnames using three fuzzy matching and simili-phonetics algorithms to account for enumeration and transcription errors. We first seek an exact match, then a Levenshtein distance of 1, then a Jaro-Winkler coefficient > 0.9, and finally a phonetic approximation using a Metaphone 3 match.[13] The final and broadest level of comparison, common to men and women, checks if the initials of surnames match.

For women, we introduce an ante-penultimate level of comparison, which allows candidate pairs to have completely different surnames as long as both individuals are possible candidates for the marital transition.[14] With this operation, we do not aim to link all women who were

13. Metaphone 3 is a phonetic algorithm providing one or two phonetic keys for a given input string (Philips 2009)

14. In the first census: unmarried women aged between 16 and 26; in the subsequent census, married women aged between 24 and 34. This arbitrary window is only used for intercensal marriage linking; our later marriage linking using the Marriage Index has much looser age conditions. We are aware that more refined data on

likely to have married over the intercensal period, but we want to link a specific subset of this population, based on the observation that their other criteria imply a confident match. This is only a workaround to the issue of the change of the surname of the woman at the time of marriage: we tackle this problem directly later, using the full-count index to the Marriage Register.

First names

We then compare first names, which are highly problematic. The use of nicknames and swapped middle names, compounded with issues in enumeration and transcription, and the high frequency of certain names complicate their use. After some typical cleaning, we restrict first names to the first word of the name string, treating any subsequent words as middle names (these were not enumerated separately). We adapt the sex-specific nickname transposition lists from Abramitzky et al. (2021) and the Census Tree project to fit British data, introducing potential nicknames for 18.34% of all census individuals (Abramitzky, Mill, and Pérez 2020; Price et al. 2021).[15] Following this preprocessing, we first look for an exact match on the first name and then on the candidate nickname. We then use the same structure of fuzzy and similiphonetic algorithms as for surnames, and seek a Levenshtein distance of 1, then a Jaro-Winkler score > 0.9, then a Metaphone 3 match, and finally a match on initials. Our final level of comparison follows Helgertz et al. (2022) and checks for the potential use of a middle name by catching cases where the initial of the middle name in one record matches the initial of the first name in the other (in either direction).

Ages

Ages are reported in the census as completed years at the time of enumeration. In some cases, ages reported as character strings or with months (such as for infants under one year old) have been converted to numeric values, truncated and adjusted by 10 years (or 20 for the 1861–1881 pair) so that they can be compared across censuses.[16] We first look for an exact match. Our second level looks for a difference in implied birth year of 1 or 2 years, to adjust for rounding and the difference between census enumeration and birthdays (Diduch 2024). Our third level seeks to capture cases of age heaping: that is cases where ages don't match but one of the ages ends in 0 or 5, and the ages are at most one decade away.[17]

Birthplaces

For birthplaces, we rely on the work conducted by the I-CeM team to standardise and categorise the places of birth strings into standardised places of birth (in most cases, a parish) and their counties (Schürer, Penkova, and Shi 2015). However, as shown in Day (2022), migrants tended to report larger localities as their place of birth as they moved further away. Therefore, we assign to each individual an alternative birthplace code, which is the likely 'next big town'

the distribution of age at first marriage exist, but selecting too large of a population as "candidate for marital transition" would dilute the estimated weight for this comparison level.

15. For example, these include "john" to "jno", or "alexander" to "alex".

16. We truncate ages instead of rounding them since all censuses took place in the first half of the year.

17. We accept that this catches more than just rounded ages. Nevertheless, it remains the best way to identify potential cases of heaping.

a migrant could have declared. We define those as the nearest place in the first decile of frequency within its county or nationally. Thus, we first compare the primary birthplace and the alternative birthplace. If neither level matches, we check whether the records share a county of birth and either individual is in a transient status.[18] This status indicates that the head of household, filling in the enumeration form, may have reported the individual's place of birth with less precision. Finally, we check whether individuals share a county of birth. The precision of birthplace information in British censuses is extremely helpful in establishing matches, for example compared to American censuses where birthplace is reported at only state-level.

Occupations

To these we add four mutable characteristics in Stages 1 and 2. These are an individual's occupation, the name of the head of household (HoH), the names of household members, and residential coordinates. For **occupations**, we rely on the HISCO occupational typology (Leeuwen, Maas, and Miles 2002). The nested structure of HISCO is helpful for comparing occupations as it provides a rough indication of the 'distance' between occupations. That is to say, an individual coded as 'Other leather goods maker' would be found in the same second-level HISCO code as 'Boot and shoe makers and repairers.' The assumption is that changes in occupations are more likely to occur in adjacent occupational categories. The only issue is that although I-CeM provides HISCO encoding, it seems to have been produced with a simple lookup table between I-CeM's own occupational typology, 'occode', and HISCO, thus largely losing the specificity of reported occupational titles. We use OccCANINE (Dahl, Johansen, and Vedel 2024) to re-encode all occupations of male individuals.[19] OccCANINE assigned codes differ from I-CeM's HISCO codes in 54% of cases. However, only 22% of all individuals with an occupation do change top-level groups. Furthermore, 705 unique HISCO codes are used (compared to only 329 in the I-CeM encoding), allowing us to achieve a more thorough use of the 'occupational space', and thus a greater precision in occupational coding. One of the main changes in the OccCANINE coding is the recoding of all agricultural labourers into general labourers. It is beneficial to group all kinds of labourers in the same HISCO code since we know some agricultural labourers were enumerated as general labourers but cannot distinguish which cases were affected. Furthermore, agricultural labourers are generally found in rural areas, where we are otherwise able to establish more links than in urban centres (see **Figure 12a** in Appendix showing the spatial distribution of the two occupational codes in 1901). The high frequency of the labourer HISCO code is counterbalanced by the use of TF-adjustments, as discussed below. When occupations are reported for both records in a pair, we seek a match at each of the three levels of the HISCO typology: the Unit group, the Minor group and the Major group.

HoH's first name

The approach for the comparison of the HoH's first name is similar to that of the comparison for first names. We use the same cleaning procedure; we first seek an exact match, then a Levenshtein distance of 1, and a Jaro-Winkler ratio > 0.9. The final level catches cases where

18. Individuals are in a transient status if they are non-familial members of the household they reside in, or if they are in the marital transition status introduced above.

19. We use the recommended threshold of 0.22; for any OccCANINE prediction with a probability $< 50\%$, we retain the I-CeM assigned HISCO code

HoH's first names differ, but either census individual is flagged as having a transient status in the household.[20]

First names of the co-residents of the household

The first names of the co-residents of the household are also compared. For each individual, we generate the list of all first names of members of the same household (excluding their own).[21] We thus seek an exact match of the list of co-residents' first names, then 4, 3 and 2 names in common.

Residential coordinates

Finally, the residential coordinates of the records are compared using the method described in Proffit (2024). This method links census addresses to modern gazetteers (and some additional historical gazetteers), to link them to their street (33.08%) and building coordinates where possible (43.26%). For the 24% of people whom we were unable to geocode, we instead used the centroid of their parish of residence. To define our thresholds, we rely on the data from Pooley (1979) for a sample of Liverpool streets. He finds that across two pairs of decades (1851–1861 and 1871-1881), 17.9% of households remained at the same address, while 27.8% moved to another place of residence in Liverpool. Reid et al. have similar findings for Belfast in 1901–1911 (Reid, Garrett, and Szreter (2016)). Although the generalisability of these data to all areas of England and Wales can be questioned, we deem it sufficient to define our basic threshold. Individuals for whom the geocoded distance between census records is less than 1 km are captured, as a heuristic for having remained in the same town or parish.

Term frequency adjustments

We use term-frequency adjustments to account for differences in the frequency of particular values. Agreement on a common birthplace such as London provides less evidence of a true match than agreement on a rare birthplace such as Lavenham. Splink therefore adds a term-frequency adjustment to the partial match weight:

$$\Delta W_{\mathrm{TF}} = w \log_2 \left(\frac{u_{\mathrm{exact}}}{f^*} \right), \qquad f^* = \max\{f_L, f_R, f_{\min}\}. \tag{1}$$

Here, u_{exact} is the non-match probability for exact agreement on the field used for the adjustment; f_L and f_R are the relative frequencies of the values in the two records; $f_{\min}$ is any configured minimum frequency; and w controls the strength of the adjustment. Values more frequent than the baseline receive a negative adjustment, while less frequent values receive a positive adjustment. Where the two records contain different values, the larger frequency gives the more conservative adjustment.

We implement term frequency adjustments for exact matches on surnames, first names, HoH names, initials for all names, nicknames, HISCO groupings, birthplaces, and birth counties.

20. As above, this includes anyone identified above as a candidate for marital transition, and anyone with a non familial relationship to the HoH, such as boarders, visitors, or servants.

21. Except for individuals enumerated in institutions or in households of more than 8 people.

In cases where we compare strings using fuzzy matching, we implement 'half-powered' term frequency adjustments (that is, with a 50% effect) of first names, surnames, and HoH names, for both the Levenshtein and Jaro-Winkler levels. This happens when comparing slightly different strings, in which case we cannot arbitrarily decide which one of the two is the 'correct' string to use for the term frequency adjustment. In these cases, we select the most frequent of the two strings, and apply its (more strongly negative, hence conservative) term-frequency adjustment, but with a 50% effect, to account for the possible error in string selection.[22]

2.4 Disambiguation

Since the FSEM model can produce several potential links for a given individual, we added a disambiguation process. If any individual is involved in several links and the two most likely candidates are, at most, 0.1 percentage point apart, we consider that these links are confounded and dismiss any links for that individual. However, if there are duplicate links beyond that threshold, we discard the duplicates and retain the most probable link, as we consider them to have been disambiguated. We believe that our approach is strict in its requirement of exclusivity: in 1901–1911, 77.42 % of pre-deduplication links have a probability above 95 %, meaning that any link with duplicates is likely to be wholly discarded. This approach is similar to Abramitzky et al. (2021).

3 LINKING WITH CIVIL REGISTRATION

One of the most significant weaknesses of census linking methods to date has been their relative inability to link women between the ages of 15 and 30. This is due to three factors.

First, marriage in the periods covered by our censuses led to a near-universal change in the bride's surname to that of the groom. Given the centrality of names to the linking process, this prevents a link being made for a woman who married between censuses, or increases the likelihood of a false link being made to a woman who remained single.

Second, marriage also usually led to a change of household for the bride and to varying degrees of migration, reducing the set of mutable characteristics available to establish links in Stage 2.

Third, marriage reinforces the misreporting problem, as it involves linking women using records filled out by their HoH (generally their father) to records written by their new husband, and with this changing authorship comes an increased likelihood of changes in the reporting of their birthplace and other personal details.

While early census linking projects excluded women entirely or included them only in household-based linking stages, we include women in Stages 2 and 3 of our method, linking them as household members and as individuals. This can present a problem for our blocking strategy, because we are missing crucial information about whether a potential pair of records is truly feasible, i.e. whether a specific single woman will get married in the next intercensal gap and hence be expected to change surname (or conversely, whether a wife has married during the preceding intercensal gap). This is damaging for our FSEM approach, as it means there are a substantial number of "true" matches which fall into the "all else" level of comparison for surnames, which otherwise mostly sees true non-matches. This artificially raises the m value

22. We specifically elect to use a 50% effect as it is the recommended approach in the Splink documentation.

of this level, with associated effects on other levels in this comparison. A solution would be to establish a dedicated level of comparison to allow for these cases of mismatching surnames among women who have married between censuses; however, as discussed above, it is difficult to identify these candidates.

In Stages 2 and 3, we nonetheless mitigate this comparison problem by introducing a level of comparison covering mismatches in surnames for individuals who are in what we call 'transition status' (i.e. of age and lifecycle status that makes them candidates for marriage in the next or previous ten years). This approach does improve the overall convergence of the model. The trade off is that the core FSEM approach produces two types of error for women across marriage: the incorrect linking of a single woman to a married woman with the same surname, where this agreement has been given too great a weight, and the incorrect linking of individuals in that transition status, where the dedicated level of comparison can be wrongly lenient on a surname mismatch. We mitigate the first problem of linking single to married women by using the Index to the Marriage Register.

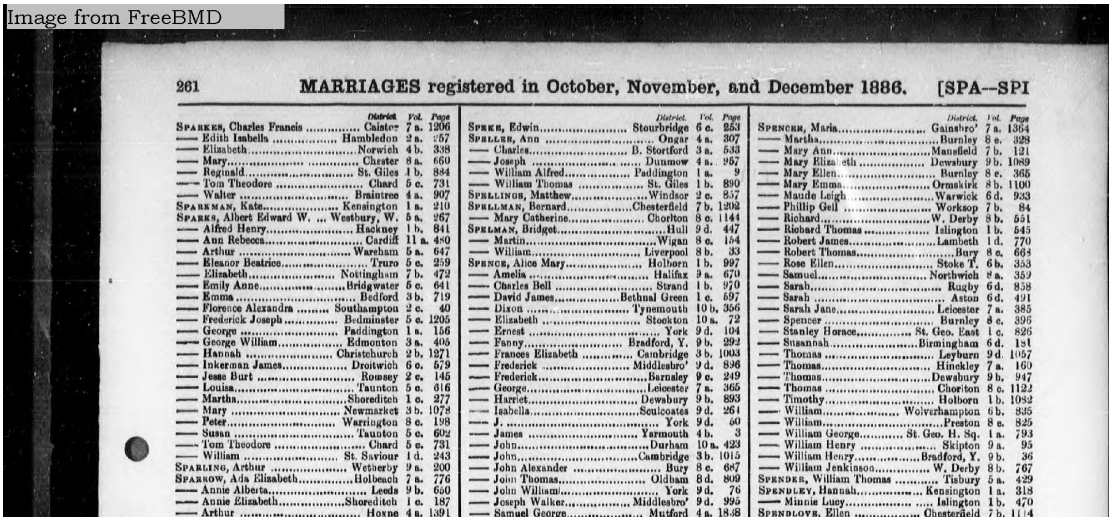
3.1 Marriage Linking Method

Recent works by Bailey et al. (2023) and Diduch (2024) have proposed possible solutions to tackle the problem of linking women across marriage, using registration data. Diduch specifically suggests a method for the British case that uses the Index to the Marriage Register to establish additional links for women across marriage. The main challenge of the method lies in the structure of the source: the Index to the Marriage Register (henceforth Marriage Index) does not list *marriage events* but rather *marrying individuals*, making the identification of couples difficult. As shown in **Figure 4**, the Marriage Index lists individual brides and grooms, separately and alphabetically under their names at the time of marriage; spouses, therefore, generally do not appear near each other in the Index. The purpose of the civil registration indexes is to allow users to find the relevant volumes and pages of the original registers where the full marriage entries are recorded. The full details of these registers have not been comprehensively transcribed or made available to researchers, but possible spouses can still be identified if they are indexed to the same page and volume within a given year, quarter, and district of marriage.

Crucially, in most cases each page of the original marriage register contains two marriages – thus four individuals and four possible combinations into couples in the Index – but this varies slightly depending on the recording practice adopted at local registry offices across the country. The correct pairing of spouses within a given page of the marriage register cannot be determined without reference to the census where husbands and wives are coresident. Further complicating the task, there are errors in the Marriage Index and in its transcription by volunteers for FreeBMD (Births, Marriages, and Deaths), leading to duplicated entries or illegible or erroneous page, volume, district, or year information. Despite these issues, most individuals can still be traced back to 'standard' pages within a volume (72.37% to pages containing four individuals, and a further 10.71% to pages containing only two individuals and thus one marriage). Another challenge presented by the Marriage Index is in the difference in name recording: index entries carry more official, verbose versions of names, often with (several) middle names or middle initials, whereas a census entry might have a single first name, or perhaps initials, nicknames and abbreviations. Despite its limitations, the Marriage Index nonetheless constitutes the most

comprehensive source that can be used to improve linking of women across marriage.[23] The rest of this section describes our approach to this source and its use in linking.

Figure 4: **Example of the Index to the Marriage Register**



Note: This image shows part of a page of the Index to the marriage register. Entries are individuals, followed by the Registration District, Volume and Page in which their full entry in the marriage register can be found. The Year and Quarter can be found in the header. Once sorted by Page, Volume, and Registration District, this allows us to identify the names present on any given page of the actual marriage register, which is the data we are using for the linking to the census.

Conceptually, the approach is to find true combinations, where three links hold: a single woman in Census A matches an index bride, the same index bride matches a wife in Census B when taking the surname of a groom indexed to the same page of the register, and the Census B wife's husband matches the groom located on the same page of the register. A key challenge is determining the pool of 'marrying individuals' in a given decade, as trying to link every spinster or every couple to the index would be not only computationally prohibitive but would also risk making false links for those who did not marry or were already married. The pool of potential brides is narrowed by removing those single women who remained single and were linked directly to the next census in Stages 2 and 3. We select potential newlywed couples if the wife has not been linked from the previous census and her husband is either unlinked or is linked from a single man in the previous census, providing indirect evidence of marriage between the censuses. This restriction prevents us from making any links via the index for married men who may have been widowed and remarried in the same decade, and may be impacted by upstream linking errors, but it is logistically necessary and consistent with our focus on linking first marriages. Consistent with the census-to-census linking, each of the three links must be uncontested: if a census husband is linked twice, if several links disagree on the right combination of partners in a register page, or if the census wife finds too many candidates in the Marriage Index, the link is rejected.

Another challenge to achieving these links is the relative scarcity of information available:

23. Sourcing a single entry from the GRO costs £12.5 with no bulk options available, and the corresponding religious registers are scattered across various local record offices and genealogical companies' holdings.

index entries only provide two forenames and surnames (where the correct grouping of bride and groom can be determined), and coarse information on marriage location and timing which can only be used as linking criteria in a limited way. This forces us to design a separate pipeline with strict blocking criteria.

The first step is to assign sex to the entries in the Marriage Index, based on the given names.[24] Whilst not strictly necessary, assigning sex greatly reduces the number of possible couples to be processed. It is impractical to use the number of individuals on a page as the only signal to assign sex because of significant misreporting or mis-transcription of page/volume numbers.

We add three further implicit blocking criteria. First, for any couple in Census B with a co-resident child, we hypothesise that their marriage occurred no later than the year following their child's implied birth year (this buffer accounts for errors in the enumeration of the child's age). Second, we only consider first marriages for women which imply an age at marriage between 15 and 35, inclusive (since the population to be linked from Census A has been restricted to single women, we are only linking first marriages). Finally, we ignore any register pages which have been indexed with more than 10 individuals, which are often the result of transcription errors or duplicate entries.

For the reasons explained above, our index linking (Stage 4 in **Figure 3**) begins with a filtering stage, where we use the Index to assess all links across marriage made for women during the direct census-to-census linking (Stages 1, 2 and 3). For any link, we create a triplet consisting of the linked Census A single woman and Census B wife, alongside the co-resident husband. We compare those triplets with all possible combinations of index couples, using both spatial and nominal blocking rules. Comparisons cover the husband's first and last name, the bride's first and last name, both middle names (if any), and whether there is any degree of spatial overlap between the census locations of the triplet and the place of marriage.[25] We apply a high confidence threshold (0.9), which alongside other constraints (such as requiring that the husband be co-resident with the wife), leads us to drop the majority (65.20%) of census-to-census links for women across marriage. This ensures that we only keep links with strong support from index entries, with the desired side effect of reducing slightly the computational complexity of the main index linking by establishing some of the correct combinations of couples indexed to the same register pages. Unconfirmed eligible core links are deleted, and their individuals returned to the main marriage-linking candidate pool. Confirmed links are retained in the final dataset, and their register and census entries are excluded from subsequent marriage linking.

24. Although we could use a simple dictionary, this approach would have struggled to discriminate names that suffer from transcription errors. Instead, we used a character-level convolutional neural network (CharCNN). This model analyses character n-gram patterns within names using multiple convolutional filters of varying sizes (3, 4, and 5 characters) to capture both short and long linguistic features that correlate with sex. The model processes names at the character level through embedding layers and applies global max pooling to extract the most salient sex-predictive features, ultimately classifying names through a fully connected output layer. We use census name+sex pairs to train this model: this is somewhat imperfect, as census names tend to be recorded slightly differently (nicknames, initials, etc) than register names (often more verbose, with many middle names) but empirically proves sufficient. We apply a simple augmentation to 30% of names by dropping a random character. We then split the universe of census name-sex pairs in an 80/20 train and test set, ensuring that the test split is fully name disjoint (that is, there are no names shared between the train and the test split). We train the model on the main split and test it against the unseen sample where it correctly assigns sex in 96.81% of the cases.

25. First at district level, then at county level. We pin the 'all else' level of comparison such that we do not negatively impact migratory couples.

The main index linking process is made up of three stages (4a, 4b and 4c), following the approach used by Diduch (2024). For two adjacent censuses, potential brides are identified in the earlier Census A (selection criteria explained above), matched to possible entries for brides in the Marriage Index, then paired with grooms on the same register page. These synthetic entities (census single woman, index bride, index groom) are then compared to married couples in the later Census B across a range of nominal and non-nominal comparisons. As is often the case with census linking, the challenge of this approach lies in the amount of blocking required. In the absence of blocking based on age at marriage and geography, any single woman from Census A could be linked to all index records with the same combination of first and last names nationwide and, in turn, linked to every male name featured on the relevant index pages.

As this results in an unprocessable amount of potential pairs of index and census couples to compare, we use similar assumptions to those made by Diduch (2024) to narrow down our list of candidates. We restrict our processing to first marriages that occurred between ages 15 and 35 (for both bride and groom, implied by the year of marriage and the reported age of a candidate in the census), and have a shifting threshold of geographic proximity between Census A, Index, and Census B districts. This usually results in several dozen million potential index couples and several dozen billion comparisons between index couples and marriage couples, near the limit of our computing resources. Further research on marriage and migration is needed to address this challenge in order to refine restrictions on plausible couples and to better account for the variety of marriage patterns.

Our comparisons between pairs of index and census couples are similar to our intercensal comparisons: the cleaning of names and pre-processing of census entries are identical, and our comparisons between characteristics are very similar. In the first stage (4a), we require the single woman's enumeration district in Census A and the district of marriage registration to match. We also require the marriage to have been registered in the same county as the county of enumeration of the couple in Census B. This reflects the prevalence of local marriages (though a degree of migration can have occurred between different places in the same district) and allows us to narrowly identify likely marriages to link and remove them from the candidates' pool before subsequently loosening blocking requirements. These couples are compared and scored on the groom/husband's first and last name, the single woman/wife's first name, age and birthplace, and if present, the middle name initials of bride/wife and groom/husband.

In a second stage (4b), we relax the geographic requirement: the enumeration of the single woman in Census A and her marriage registration are only required to be in the same registration county. There are no longer spatial requirements between the place of marriage and the place of enumeration of the couple in Census B, but we add a further comparison, checking if the district of marriage and second census enumeration match, or if their centroids are within 30 km of each other. We prevent the mismatch level of comparison from carrying any negative weight.

The third and final stage (4c) occurs with no geographic blocking at all: single women can be matched to an index entry from anywhere in the country and then matched to a census couple anywhere. This allows us to link individuals who underwent inter-county migration around marriage.

Training FSEM on this dataset is somewhat challenging: our approach creates a lot of synthetic entities by merging single women and index brides on names alone (with varying and sometimes no spatial blocking). Highly common name combinations end up being multiplied in great numbers, such that the top 5 most frequent combinations (Mary Jones, Mary Smith,

Elizabeth Jones, Mary Williams, and Mary Davies) end up accounting for over 25% of all 'left' entities (Census A single women + index bride and groom). The multiplicative problem is also worsened by the within-register page multiplication issue. Most census/bride pairs are also doubled to create an entity containing each groom on the same register page. To deal with this multiplicity challenge, and the underlying problem of using FSEM over synthetic entities, we restrict FSEM training to a sample of rows with little multiplication.

This approach, through its gradual loosening of blocking criteria, does increase the risk of biasing the linked population against movers. Despite this, we believe that linking civil registration data is still a worthwhile exercise. What may be lost because of an increased non-migratory bias is, however, more than offset by the significantly reduced gender bias of the final sample. We provide researchers with our linked data for both the foundational (Stages 1 to 3) and the index-linked (Stage 4) model, so that they can use the most suitable version depending on their research needs. The civil registration linking results for each census pair are presented in Table 2.

Table 2: **Linking rates for women across marriage**

	51–61	61–81	81–91	91–01	01–11	11–21
Linked in Core Method	180,214	239,705	242,978	304,988	340,572	360,817
Confirmed by Marriage Index	42,651	78,452	80,465	112,191	137,344	129,724
Removed, as no support found in Marriage Index	-137,563	-161,253	-162,513	-192,797	-203,228	-231,093
Other explicit invalidations	0	0	0	-2	0	0
Stage 4a	197,449	429,138	348,470	450,441	552,980	621,668
Stage 4b	72,576	196,641	173,145	221,096	253,154	274,497
Stage 4c	131,251	308,247	152,456	163,550	171,680	196,060
Net Links Made in Marriage Index phase	**263,713**	**772,773**	**511,558**	**642,288**	**774,586**	**861,132**
Total Marriage Links	**443,927**	**1,012,478**	**754,536**	**947,278**	**1,115,158**	**1,221,949**
Brides Linked through marriages (in% of total marriages)	**27.71**	**27.14**	**36.38**	**39.59**	**42.21**	**39.73**
Registered marriages in the period (England and Wales, m)	1.602	3.730	2.074	2.393	2.642	3.076

Intercensal marriage totals are from (Mitchell 2011)

4 VERIFICATION STRATEGY

Verifying the reliability of the linked data produced is a crucial step in ensuring their quality and suitability for further studies. Ruggles, Fitch, and Roberts (2018) divide verification between commission errors and omission errors. In census linking, an error of commission stands for an incorrectly established link, i.e. a false positive, while errors of omission cover individuals who are not linked, even though they are enumerated in the target census, i.e. false negatives. As there are well-known methods in social sciences to deal with biased samples (the consequence of omission errors), the crux of the verification stage is to get an approximation of the nature and level of commission errors in our linked data. What follows is a description of our approach to do this.

The main challenge in our case is the absence of ground truth for the nineteenth and first half of the twentieth century (Bailey et al. 2023). This is not simply a practical question (i.e., the fact that such data are not readily available), but rather a near-epistemological point that such a ground truth is impossible to create. The necessary condition for the possibility of such data is the existence of a source that would reconstitute with certainty the life course of a large and unbiased sample of individuals across the period, and to our knowledge no such document exists. Some have attempted to circumvent the issue using crowd-sourced genealogical data as near-ground truth material (Price et al. 2021), allowing one to compare linked data to a large number of existing links. Some doubts about the quality and bias of links made by amateur genealogists could be raised, even if extensive research on raw data can extract the meaningful signal and produce a useful set of links (Price et al. 2021). Regardless, the unavailability of such data in the British context prevents us from relying on this approach altogether.

Another possible avenue is to base verification on independent manual linking. However, if someone is given the task of linking a random individual entirely independently of the algorithmic solution, what will this person do to find the matching individual among 25 million other individuals in the previous or subsequent census? If the said individual consistently resided in the same parish, had a stable household, and/or a relatively uncommon name, there is no doubt the reviewer will make the link. But what if the person is no longer present in the parish where he or she was previously listed? What will the reviewer do? In the best scenario (supposing a person with data manipulation expertise and some demographic or family history training), they will perform a series of multivariate searches to identify possible matches, likely relying heavily on names to pick the best one, or simply conclude that no obvious single match can be identified. Is this not asking a person to do what the model does, but with more limited means? Whilst the human linker can bring a more nuanced understanding of similarities and differences between names and other variables, they will lack the model's ability to account for the frequency of various levels of similarity and will be more limited in which pairs and how many pairs they can evaluate (the equivalent of blocking). In our view, unless the reviewer has access to other material not included in the model (e.g., registration certificates indicating filiation, individual biographies, press sources, etc), this approach verges on the tautological and has a limited interest.

To illustrate this point, it may be useful to share an example from our manual review process. A link which our algorithm made with high confidence between a single woman in the 1901 census and a married woman in the 1911 census, via a record of marriage occurring in 1903, struck the human reviewer as unlikely because the records disagree on the woman's surname in the first census and maiden name in the Marriage Index - Florence Telling vs Florence Alice

Hinder. A quick search of other 1901 census records for the same district reveals a promising alternate candidate in Alice Hinder, of similar age, raising the question of which is more likely: that a woman would go by a middle name and slightly misstate her age in the census, or that she would have a completely different surname in records two years apart?

The linking algorithm made its choice by contextualising these characteristics within larger population distributions, but the human researcher can go further down the rabbit hole. Having ordered and patiently waited for Royal Mail delivery of the original extract from the Marriage Register, we discovered that Florence Alice Hinder's recorded age at marriage agrees exactly with Florence Telling's age in the 1901 census and that the entry for her father's occupation in the Marriage Register was left blank, suggesting an illegitimate birth. Through online genealogical records, Florence Hinder's mother can be identified in a baptism register and then traced to a record in the Marriage Index, together with one John Telling. Further exploration of the family tree reveals that the other candidate, Alice Hinder, was most likely a cousin, also born to an unmarried mother from the Hinder clan before being recorded in subsequent censuses under her stepfather's surname. Reassuringly, the result of these hours of historical detective work confirmed the seemingly unlikely match made by the algorithm. This case also illustrates the perfect storm of data discrepancies, omissions, and quirks that our algorithm has been designed to tackle. Of course, not all such difficult cases may be resolved in favour of the machine, but this example is a telling case of the power of multivariate comparisons in census linking.

Because it would be enormously costly and time-consuming to produce this quality of review for each case - at least £11 per individual to cover the cost of ordering the birth certificate, not accounting for the researcher's time - the other most common verification strategy involves providing a human reviewer with a set of links created by the algorithm and asking them to assess their veracity. This process is not straightforward either. For example, if only the final link is presented to the reviewer, without information on competing links, the review process will greatly overestimate the precision for cases of high commonality. To remedy this, it is possible to provide reviewers instead with a set of candidates generated by the same blocking rules as the linking method itself, asking the reviewer to select what they consider to be the most likely link.

This approach is also not entirely satisfactory. First, blocking rules are still likely to exclude true matches, but they can also generate an absurdly large number of candidates for individuals with common names, making the review process both tedious and highly time-consuming. Second, it remains an open question whether a verification process can truly rely on the existing OCR/HTR output. Given the (uncertain degree of) error introduced by the transcription (Ghosh, Hwang, and Squires 2024), a verification process based on the original images would surely further increase the precision. Furthermore, even if link verification allows for a relatively good measure of false positives (wrong links), it does not in any way capture false negatives (missed links). Some ways to mitigate this problem would be to complement a manual review of links with a set of family reconstitutions with stochastic selection of starting individuals, compare linking profiles with crowd-sourced biographies, or compare linked data against external records, such as a death register or even full population registers for subsamples of linked data (Ó Gráda et al. 2023).

Due to these conceptual issues related to the nature of verification itself, we are of the view that the best verification method is to combine several of them. Here, we adopted three complementary verification methods. First, we conducted a standard human review of 3,600 links, asking a reviewer to qualify the pairs as true, uncertain, or false. In order to do this,

we built a customised web-based interface for verification work, which greatly increases the verification efficiency. Second, we presented a reviewer with a random set of individuals and, for each individual, several potential matches in an adjacent census, asking them to select a link, if any was deemed possible. Third, we used the death register index to verify that we did not link forward deceased individuals. By comparing all of these data, our article offers a transparent and rigorous analysis of the quality of our links that includes potential biases identified at each stage and for each criterion, including age, place, sex, migration, and marital status. All of these exercises are conducted over the data that stem from the core linking method (Stages 1 to 3, that is, before the register linking).

As for the specific verification of the linking with the Marriage Index (Stage 4), we adopted a similar binary verification approach, presenting a reviewer with a sample of links containing the single woman, wife, husband and the entries to the Marriage Index used to establish that link. We also compared the distribution of the implied average age at first marriage per district with the Singulate Mean Age at Marriage (SMAM) obtained from (Reid et al. 2018).

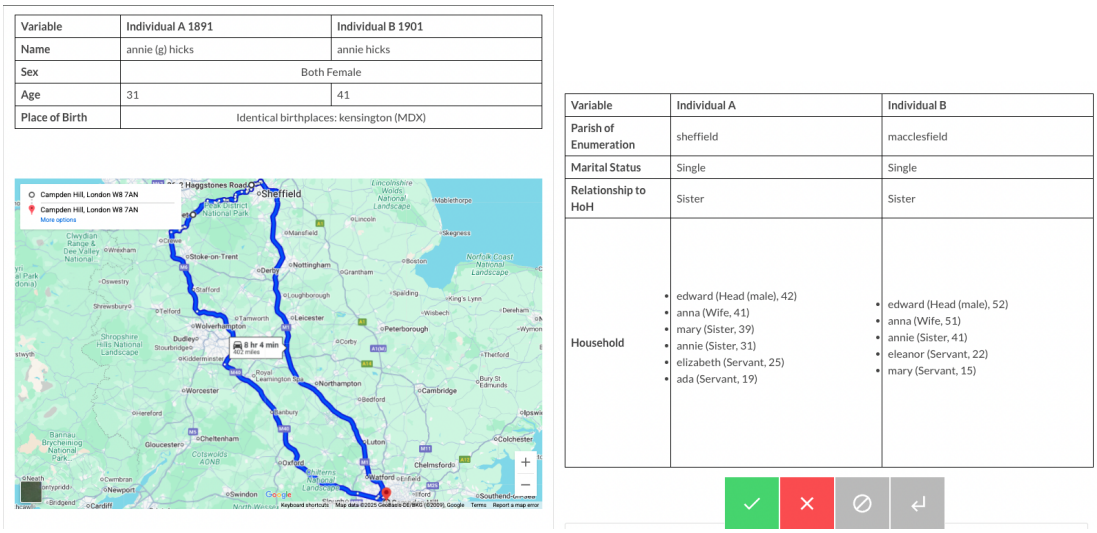
4.0.1 Standard review

We built a verification interface for a human reviewer that presents all the characteristics of both linked individuals, shown in **Figure 5**. We present our human reviewer with data for the first, middle, and last names, sex, age, marital status, and relationship to the HoH. When available, we present both the raw occupational string and the coded value in HISCO. We also show both places of birth and of enumeration for each individual, and where any two of these differ, a map with all four points. We also list all members of the households, with their names, age, and relationship to the HoH. By so doing, we adopted a maximalist approach, and thus present richer information than the linking algorithm would have seen. This is to make the best use of our reviewers' professional understanding of nineteenth-century lifecycles, history of migrations, and occupational change. As a number of cases were neither obviously true nor false, we included an 'uncertain' option. We used this approach to review 600 links per census pair, i.e., 3,600 links in total. **Table 3** shows the results of this process.

Table 3: **Standard manual review** (All%)

	51–61	61–81	81–91	91–01	01–11	11–21	Overall
True	85.97	84.75	91.73	91.43	92.62	93.25	89.95
Uncertain	3.87	6.39	3.86	3.49	1.51	3.04	3.69
False	10.16	8.85	4.41	5.08	5.87	3.71	6.34

Figure 5: **Binary Review Interface**



In the standard review model, our reviewer found four common patterns for false positive links. The first involved links between siblings who were close in age. Taking into account similar ages, identical places of birth, identical surnames, and similar or identical family structures, the algorithm occasionally mistakenly linked same-sex siblings, particularly when one sibling is absent from the household in the next census. The second type of false links regarded individuals with identical birthplaces, positions within the household, HoH's name and identical ages, but with different names and family structures. Third, there were links between sisters-in-law with similar ages and places of birth, some of whom even occasionally shared first names. Our reviewer rejected these due to the differences in household structure and places of birth. The final and least common, albeit still identifiable, pattern of false linking was individuals who shared first names and places of birth but with severe differences in age and different household structures. In most cases these 'false' links were made by the algorithm because unique values for name or birthplace were taken as stronger signals in favour of a match, but the human reviewer disagreed on the relative strength of signals against the match based on discrepancies in other characteristics, including some that were not included in the original algorithm.

Improvements in match rates from one census pair to the next are likely due to greater accuracy and legibility during transcription, due to improving literacy as well as refinement of enumeration practices; weaknesses of this kind are particularly prevalent in the 1851 census, accounting for the greatest rate of false links in the 1851–1861 census pair. The 'true' rate for 1861–1881 is artificially lower because of the twenty year gap. When the 1871 census becomes available, we expect the linking rate to return in line with the trend.

Breaking down the binary review by linking stage of each pair, some trends emerge: Stage 1 concentrates the highest proportion of pairs labelled as false (11.8%), Stage 2 produces 3% of false pairs, whilst Stages 3A and 3B contain respectively 4.2% and 1.3% of false pairs. There are clear gender tradeoffs between precision and recall: the pipeline establishes more links for men, but this comes at the cost of a higher error rate. We postulate that this also comes from greater

mobility for men and more frequent changes in characteristics, making them more challenging to link. Precision-Recall tradeoffs also occur within gender: for men, Stage 1 links about 3 times more than Stage 3A and carries the highest rate of error, whilst for women, Stage 2 which generates about 25% more links than stage 3B also carries the higher rate of error.

4.0.2 Manual subset linking - Blind review

We created a second version of our web interface specifically for this verification exercise. Here, we seek to compare the precision of the automated linking method against a simili-ground truth. Thus, we selected 1,000 random linked individuals across all censuses and selected potential candidates in the adjacent censuses. To do so, we establish blocking rules to select potential candidates in a similar way to our automated method. For each randomly selected individual, we presented all the candidates identified to the reviewers in a random order (if there are more than 30 candidates, we rank them by Jaro-Winkler similarity of the first and last names and only retain the top 20, then shuffle the order again). We ensured that the set of options shown to the reviewer always included the link selected by our method: crucially, as our objective is to review the method, we did not inform the reviewer of this.

We asked our reviewer to inspect all records and pick the correct link. We also gave our reviewer the option to indicate the absence of a matching individual, to account for under-enumeration, false negatives introduced by blocking rules, or over-commonality. **Table 4** shows the results of this process.

Table 4: **Manual subset linking** (All%)

	51–61	61–81	81–91	91–01	01–11	11–21	Overall
Agreement	64.82	45.45	60.1	70.56	77.16	79.29	66.23
Disagreement	3.52	1.52	3.54	3.55	1.52	2.53	2.69
No link selected	31.66	53.03	36.36	25.89	21.32	18.18	31.07

From this experiment, we draw two insights. First, it appears clearly that there are very few cases in which the reviewer disagrees with the algorithmic solution: accounting for the imperfection of the reviewer's own choice, we take this as a strong sign of confidence in the quality of the linking. Secondly, there is a high proportion of cases in which the reviewer declined to make a choice even though the algorithm's solution was among the candidates. The majority of these cases are due to over-commonality. In the face of many plausible choices, the reviewer was unable to select a clear match. However, in those same cases, the machine established several plausible links, of which only the most likely survived the disambiguation stage and triggered a link.

In our view, this proves the advantage of systematic computation and comparison of the statistical distribution of match levels and specific values for all possible candidates in a way that is often out of reach for a human reviewer. The example of Florence Telling described in the previous section illustrates the difference between the linking process for humans and machines. Only a painstaking analysis of full registration records allows a reviewer to pick with high certainty the right candidate, while the algorithm is able to reach a similar conclusion by weighing instead all the other possible candidates at once. The algorithm is certainly also

making errors in this process, but it will only be possible to measure the rate at which this happens when we are able to construct a very thorough validation set built on these most difficult cases. This is something we leave for future work.

4.0.3 Comparison using the Index to the Death Register

We also used the Index to the Death Register as an additional piece of information in the evaluation of our links, as obviously any person registered as deceased in the ten years (twenty in the case of 1861-1881) after a census should not have been linked to an individual in the subsequent census (this approach is adapted from (Feigenbaum 2016)). To do this, we adapted our methods to link individuals aged over 40 from the census to the Index to the Death Register (infant mortality is harder to link, notably because most young deaths occurring in the intercensal decade would never have been enumerated). Then, we compare these links to the links made by the model to evaluate the proportion of potentially deceased individuals who have been (thus erroneously) linked in the following census. **Table 5** shows the results of this comparison.

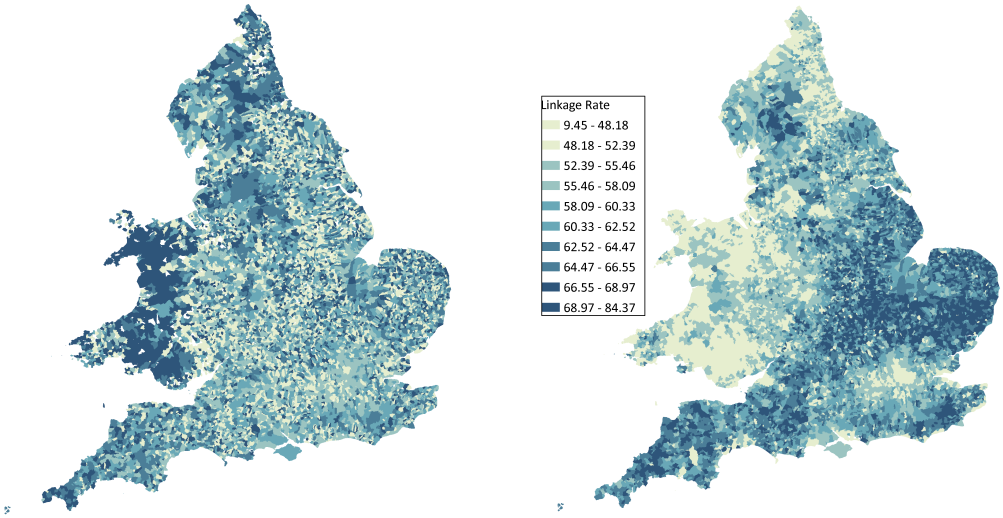
Table 5: **Comparison of Census and Death Register Linking** (in%)

	61–81	81–91	91–01	01–11	11–21
Agreement	86.94	88.72	90.84	92.37	94.35
Potential errors	13.06	11.28	9.16	7.63	5.65
Total number of linked deaths	194,293	197,785	238,240	305,285	405,540

This approach has four main limitations. First, it can provide information on the quality of the linking only for an older sample of the population. Second, it does not provide us with an actual error rate, since the link to a death entry may be incorrect in the first place. Third, the Index to the Death Register, like the Index to the Marriage Register, contains limited individual information, and ages at death are only recorded from 1866, making any attempts to link between the census and the Death Index before 1866 much more speculative, and thus, we do not attempt to link the 1851 census to the Death Index. Finally, we cannot disentangle the errors coming from linking to the Death Index from the errors coming from intercensal linking, which means that we cannot use these data to give us a rate of error, but as an internally consistent, relative proxy.

However, **Figure 6a** shows a very clear spatial distribution of these potential errors, which points to a marked Welsh, and, to a lesser extent, Cornish and Northern English, exception characterised by a much higher rate of questionable links. We believe that this is mainly due to the remarkably higher homogeneity in naming patterns in these areas (see **Figure 7**) compounded, in the case of Wales, by the linguistic variance from the rest of the English data seen by the model. This pattern of over-commonality increases both the likelihood of making a false link to the Death Index and the likelihood of producing false intercensal links.

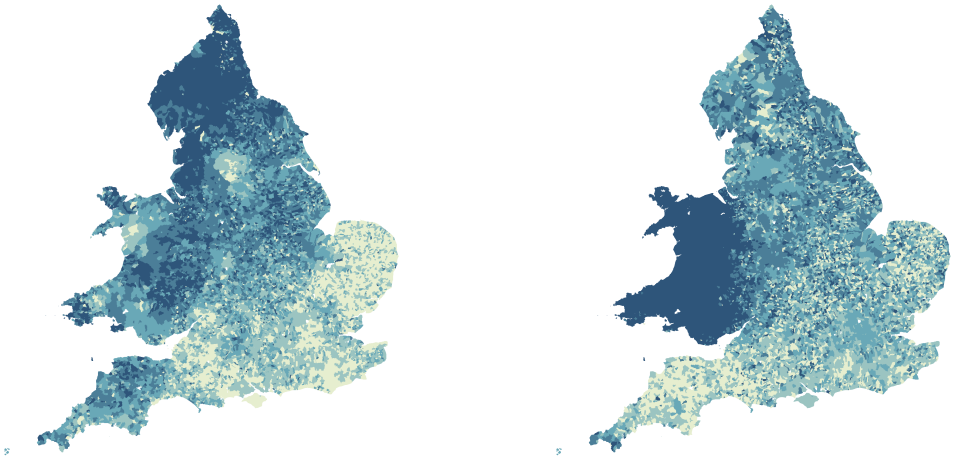
Figure 6: **Spatial distribution of death-register conflicts and mean parish-level linking rates**



(a) **Spatial Distribution of linking error using the Death Register Index 1861–1921**

(b) **Mean Match Rate per Parish 1851–1921**

Figure 7: **Proportion of first names and surnames in the national top-20 in 1891**



(a) **First names**

(b) **Surnames**

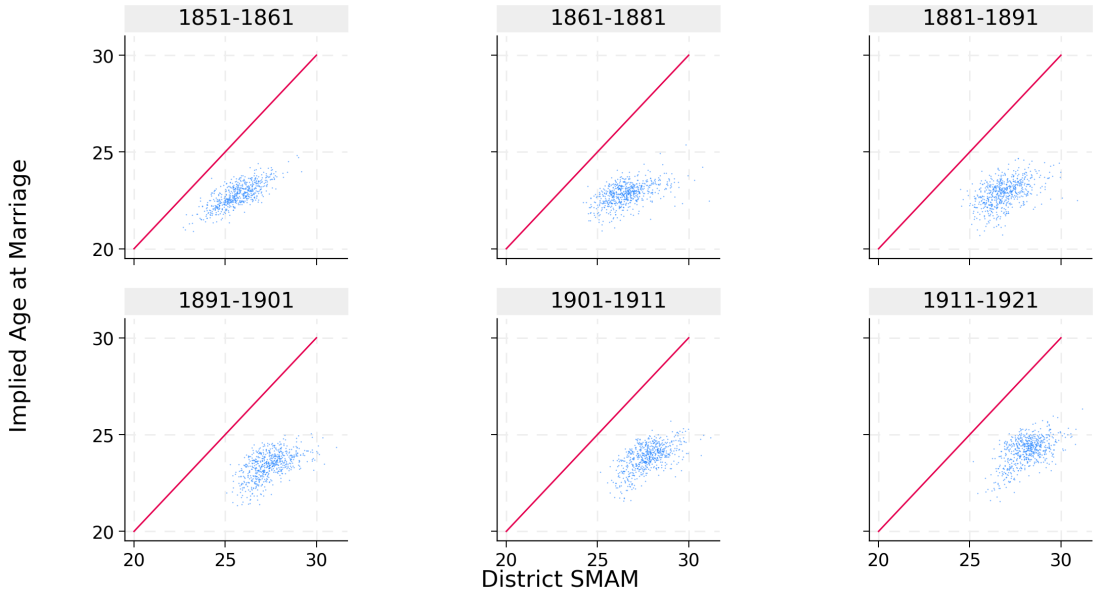
When comparing **Figure 6a** to the one showing our average match rate for all censuses (**Figure 6b**), it is clear that these two phenomena, linguistic variance and name commonality, explain the significant underperformance for Wales. Interestingly, this pattern is not found in London, where the Death Index error rates are low, and match rates are low. We believe that the lower match rates in London most likely result from a much higher level of mobility among the population, rather than linguistic idiosyncrasies or excessive naming commonality.

4.1 Comparison of implied age at marriage and Singulate Mean Age at Marriage

Another way to assess the quality of our links to the Marriage Index is to compare the ages at marriage implied by these links with other estimates of age at marriage in this period. These can be obtained from cross-sectional census data using an indirect estimation technique to calculate Singulate Mean Ages at Marriage (SMAMs) by registration district (Reid et al. 2018). These estimates are derived from the proportion of the population that is married by year of age for a given geographic area and are therefore less precise than individual-level estimates and subject to biases caused by migration and other changes to the underlying age and sex distributions of a population (Schurer 1989). However, it is important to note that our observations of ages at marriage derived from links to the indexes are also imprecise, both due to potential linking errors and to the lack of exact dates of birth and marriage. Nevertheless, the ability to observe age at marriage at the individual level is a great advantage of the linked data for researchers.

Comparison of the average implied age at marriage for brides and female SMAMs by registration district and decade shows that our direct estimates are consistently lower than the indirect ones (**Figure 8**). We refer to our estimates as 'implied' ages at marriage because they are derived from the round age of a woman recorded at the nearest census and the year of the linked Marriage Index entry, and are thus rather fuzzy. The largest difference between these average estimates is eight years for the district of Farnham in Surrey in 1861; the smallest difference is just 0.19 years for Settle in the West Riding of Yorkshire. In districts with high turnover of single women (such as through employment as domestic servants), SMAMs are expected to be biased upwards; a woman leaving the district upon marriage is likely to be replaced by another single woman of similar age, thus maintaining the upper tail of the age distribution of single women. In addition, the age restrictions which we have had to implement in our method for linking to the Marriage Index may favour younger marriages (although they are not unduly restrictive around the expected distribution of age at first marriage). Reassuringly, both estimates track similar shifts in the distribution towards older ages at marriage during this period.

Figure 8: **SMAM and average implied age for linked women by Registration District**



4.2 Manual Verification of Marriage Linking

In our final verification, we reproduce the standard review exercise conducted over intercensal links and tailor it to verify the links established using the Marriage Index. For each link, we present our reviewer with 5 entries: the census bride, the index bride and groom, the census wife and her husband. Presenting these five personas allows our reviewer to verify the coherence of the full 'key' made to establish links, verifying at once the validity of the link between the census single woman and the census wife, the index groom and the census husband, and the index bride and census wife. Our reviewer checked 300 randomly sampled links by census pair, for a total of 2,000 links.

Table 6: **Manual Review of Marriage Linking** (All%)

	1851–1861	1861–1881	1881–1891	1891–1901	1901–1911	1911–1921	Overall
True	67.31	70.95	69.86	70.18	72.99	69.67	70.23
Uncertain	9.39	5.74	6.20	6.60	5.97	6.43	6.67
False	23.30	23.31	23.94	23.22	21.04	23.91	23.10

The results (see **Table 6**) show that the Marriage Index links are substantially noisier than the core intercensal links. On average, 70.23% of reviewed links are verified as true, 6.67% are plausible but uncertain, and 23.1% are judged false. The True rate peaks in 1911, following a similar pattern as recall (and most likely for similar reasons).

The error rate is non-trivial; there are three partial explanations for this. First, each link is made up of three sublinks (as stated in the method section), which means that the reviewer has more fields on which to find a mismatch, disproving the whole link.[26] Second, although the FreeBMD Index to the Marriage Register is fundamental in allowing us to cover the universe of marriages across our period, it is particularly vulnerable to the mistranscription of a single digit in a page or volume number, which corrupts the identification of spouses. Third, the error rate is not even across stages: with only 13.26% false links, stage 4a fares far better than 4b or 4c. As this is the most spatially constrained stage, it is least affected by the issue of multiplicity and synthetic entities that affect other stages more substantially. Future research should thus investigate the use of other linking approaches to incorporate the Index to the Marriage Register in a full linking pipeline. We nonetheless believe that the inclusion of the Marriage Index is a valuable contribution of this work, as it constitutes the first use of a comprehensive marriage source for census linking at this scale. We also release two versions of the links, with and without Marriage Index links, allowing researchers to select the version most appropriate for their research design.

5 RESULTS AND BIASES

This final section describes our results and offers a systematic analysis of the biases embedded in our linked data.

5.1 Headline Results

Table 7 shows the match rates for all census pairs expressed as a percentage of the total linkable population (that is, the population aged 10 or over at the end census date).[27] In this section, all our bias statistics are calculated by comparing the characteristics of the subset of linked individuals with those of the total linkable population at the end census. Overall, we link between 45 and 63% of 'linkable' individuals for a standard census pair, and across the entire period we successfully link 55.62% of all 'linkable' individuals. The rise in match rates over time reflects a variety of factors, including increasing literacy rates, improved enumeration, and better source survival. The low performance for the pair 1861–1881 is an anomaly which is a direct consequence of the absence of a transcription for 1871 in the I-CeM dataset. The twenty-year gap between observations affects mostly Stage 2 of our linking, as household stability over two decades is significantly less likely. The lower match rate for 1911–1921 is partly caused by the low maturity of the currently available version of the 1921 census (the I-CeM team will release an updated version in the coming months), by the mortality impacts of the First World War and Spanish Influenza,[28] and the specific conditions in which the census was taken (delayed

26. Equally, the reviewer has little information on which to gain certainty. For example, whilst an intercensal link for a child aged 2 and 12 could be easily confirmed by matching other household first names, this cannot apply to women across marriages who, almost universally, simultaneously change households

27. A few individuals under the age of 10 are matched, as other characteristics match sufficiently, but this would only happen in the case of a substantially incorrect age. This is a rare occurrence, forming 1.73% of our links

28. Which would not directly impact the match rate as it is calculated backwards from the terminal census. However, it had substantial impact in terms of household dislocation and recomposition, making linking more challenging

to June due to the threat of industrial action, with many people enumerated whilst on holiday, away from home).

Table 7: **Match rates for each census pair**

	51–61	**61–81**	**81–91**	**91–01**	**01–11**	**11–21**
Stage 1 Men with Mutable Chars.	37.45%	25.34%	46.24%	50.14%	51.92%	48.12%
Stage 2 Co-resident Women	16.56%	7.16%	24.41%	28.17%	31.14%	29.39%
Stage 3A Men with Immutable Chars.	15.65%	20.00%	15.94%	16.57%	15.66%	14.39%
Stage 3B Women with Immutable Chars.	17.16%	13.13%	20.42%	22.11%	22.80%	19.79%
Stage 4 Marriage Index Linking	3.48%	10.50%	4.46%	4.89%	5.19%	5.21%
Total Match Rate	**44.85**%	**37.72**%	**55.19**%	**60.35**%	**63.18**%	**58.21**%
Total Links	6,584,138	5,307,505	12,195,636	15,277,231	18,182,810	18,325,373
Linkable Population	14,679,246	14,069,971	22,098,287	25,313,117	28,781,341	31,479,437
Links Stages 1-3	6,320,425	4,534,732	11,684,078	14,634,943	17,408,224	17,464,241
Links Stage 4 (net)	263,713	772,773	511,558	642,288	774,586	861,132

N.b. Stage-wise match rates are sex specific and as such do not add up to the Total Match rate.

5.2 Deterministic comparison

To contextualise our results within the wider literature, we implemented a baseline deterministic linking algorithm following the widely used Abramitzky, Boustan, and Eriksson (2012) (ABE) method. Because the ABE algorithm was originally designed for US historical censuses, where birthplace information is restricted to the state level, we ran two variants on our data. The first variant, "ABE Birth County," restricts geographic blocking to the county of birth, mirroring US censuses. The second variant, "ABE Birthplace," leverages the parish-level birthplace information available in British censuses Day (2022).

Table 8: **Comparison against ABE (2012) deterministic linking**

	This Paper	**ABE Birth County**		**ABE Birthplace**	
Census Pair	**Match Rate**	**Match Rate**	**Agreement**	**Match Rate**	**Agreement**
1851–1861	44.85	25.63	60.68	27.30	77.79
1861–1881	37.72	24.12	52.02	22.66	71.45
1881–1891	55.19	32.80	70.91	35.44	82.93
1891–1901	60.35	35.92	75.09	41.47	84.32
1901–1911	63.18	39.07	78.52	43.91	86.26
1911–1921	58.21	37.06	76.80	33.78	87.38

Notes: Agreement is the percentage of ABE links that match this paper's source–target link.

The results, presented in **Table 8**, unsurprisingly show the advantages of a multi-stage probabilistic approach. By allowing for variance in some variables when agreement is strong in others, using a wider set of information, and using more advanced deduplication, our FSEM model is able to produce millions of additional valid links that strict deterministic rules must discard. This exercise also quantifies the value of granular birthplace information (as opposed to rougher units, such as birth county in our case), supporting both greater recall and precision (using agreement with our links as a proxy for precision.)

5.3 Analysis of biases in the linked data

The most persistent biases in the results for Stages 1 to 3 are sex composition and age. **Table 9** shows that the linked population is consistently more male-skewed than the full census population. In the Core links, the male–female gap is up to 36.57 percentage points larger than in the full census population, for the 1861–1881 pair, corresponding to women accounting for 18.28 percentage points less of the linked sample than they do of the full census population. Our use of the Index to the Marriage Register in Stage 4 substantially reduces this imbalance. In the best case, for the pair 1901–1911, the Core+Index male–female gap increase falls to 5.70 percentage points, corresponding to a female-share shortfall of 2.85 percentage points.

Similarly, as shown in **Figure 9**, match rates also vary significantly by age. A high proportion of children between 10 and 15 are linked, and a substantial proportion of adults older than 35 can also be linked.[29] However, between these age brackets, linking rates are significantly

29. The lower match rates for 10-11 years old is due to children born after the previous census and the under-enumeration of infants.

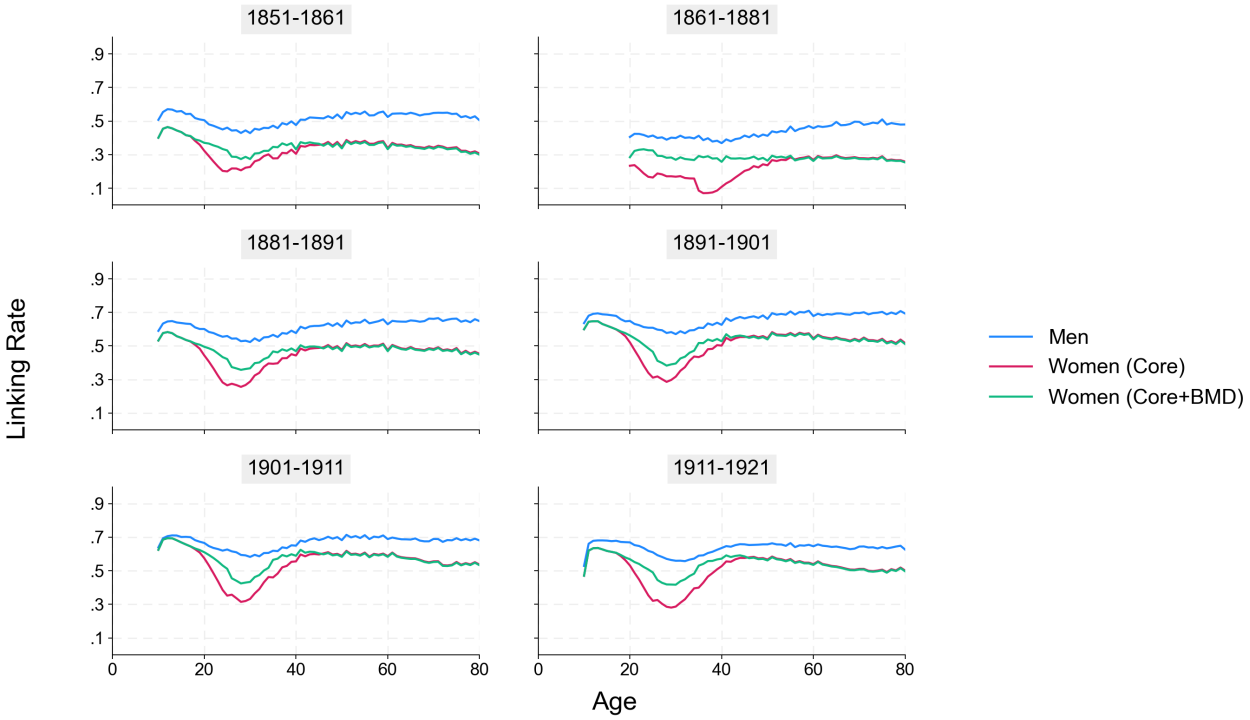
Table 9: **Sex composition and male–female gap in linked populations**

	Full census		Core			Core+Index		
Census Pair	Male	Female	Male	Female	Δ M–F gap (pp)	Male	Female	Δ M–F gap (pp)
51–61	48.35	51.65	58.75	41.25	20.79	56.31	43.69	15.92
61–81	47.68	52.32	65.96	34.04	36.57	55.85	44.15	16.35
81–91	47.87	52.13	55.40	44.60	15.06	53.02	46.98	10.30
91–01	47.84	52.16	54.33	45.67	12.99	52.00	48.00	8.33
01–11	48.13	51.87	53.28	46.72	10.31	50.98	49.02	5.70
11–21	47.43	52.57	53.10	46.90	11.33	50.57	49.43	6.29

The Δ M–F gap is calculated as $(M_{\text{linked}} - F_{\text{linked}}) - (M_{\text{full}} - F_{\text{full}})$.

lower. This corresponds to the average age of leaving the parental household and the most intense stage of migration (Day 2019). This age bracket also covers most marriages (women's average age at first marriage throughout the period was around 25) and therefore changes in female surnames, which explains both the deficit in linked women and the pronounced age-specific linking pattern.

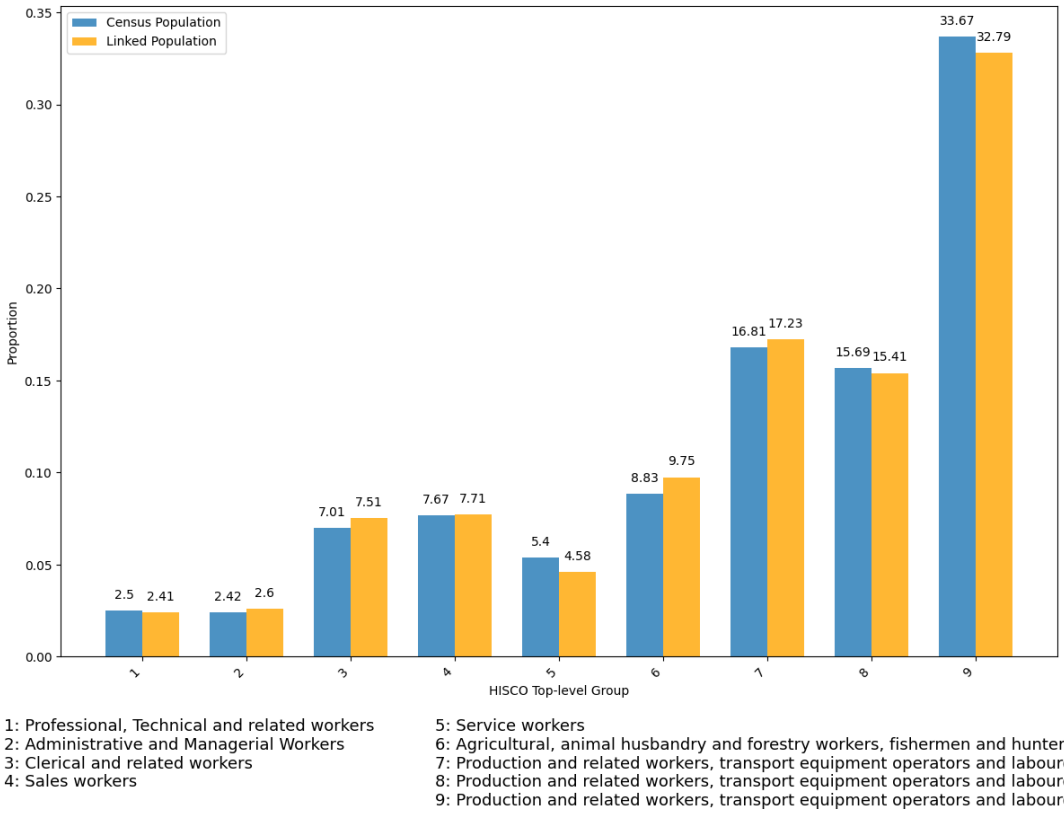
Figure 9: **Backwards linking rate by age**



As discussed above, the distribution of the mean linking rate per parish throughout the period (**Figure 6b**) shows that areas with the lowest linking rates are those whose linguistic characteristics may have made the enumeration more challenging (especially Wales) and London and its periphery due to the very high population turnover.

We also measure the occupational bias introduced in our linked sample. Occupation is an important determinant of migration and thus some occupations may be disproportionately represented in our linked sample if they are more associated with residential stability. **Figure 10** shows that occupational bias is mainly against people who do not report an occupation. Some particular occupations suffer from a very low linking rate: 'servants, maids and housekeepers' (54020) are only linked at 35.25%, 'cooks' (53101) at 30.24%, 'charworkers' (55200) at 26.44%. The negative bias for linking domestic servants is an interesting result, both because occupation is probably interacting with sex and age (as domestic service was the largest occupational category for single women in this period) and because the census enumeration of servants may have been affected by errors introduced by heads of household recording information on behalf of their residential employees. On the other hand, as many as 68.73% of 'watch, clocks, and precision instruments manufacturers' (84290) are linked, 63.83% of farmers, and 59.18% of 'office clerks' (39310). Thus, the use of these linked records for purposes that are highly dependent on representativeness over a few occupations may require careful reweighing of the data.

Figure 10: **Occupational Bias by HISCO top-level group**



Note: This figure is calculated by aggregating all census pairs, as the effect is comparable.

Finally, we use the occupational status indices (CCC-HISCO score) produced by Clark, Cummins, and Curtis (2024) to study the social representativeness of our linked sample from a near-class perspective. We study the population retrospectively: occupations are taken from the second census in a pair for any individual linked to the previous census. **Table 10** shows that the linked sample is relatively similar to the total linkable population.

Table 10: **CCC-HISCO scores**

Census Pair	Whole population	Linked Sample	Difference (% of SD)
1851–1861	35.69	35.66	−0.21
1861–1881	37.10	37.14	0.31
1881–1891	36.84	37.23	2.67
1891–1901	38.29	38.82	3.71
1901–1911	38.69	39.25	3.83
1911–1921	39.30	39.65	2.27

6 Conclusion

In this article, we detail a full pipeline for the linking of full-count British historical censuses and offer the first set of links covering the period 1851–1921. We hope that our method may also be transferable to other national contexts and uses. By staggering the linking over several stages and choosing to incorporate both immutable and mutable characteristics, our method offers a better balance to deal with the two primary challenges of census linking: excessive variance and excessive commonality. Furthermore, incorporating the full Index to the Marriage Register allows us to significantly increase match rates for women aged between 15 and 30, where traditional census linking generally fails due to surname change. We link an average of 55.62% of the total linkable census population with a rigorously measured error rate of around 7%.

We also provide analyses of possible biases in our linked sample to help researchers understand and use our data. Our linked sample is most significantly biased in gender and age due to the remaining deficit in linking rates among women aged 15–30. Our linking is not homogeneous across space, and two areas, London and Wales, underperformed compared to the rest of the country. We do not detect any major bias in terms of class.

We do not claim that our linked population is free of further unobserved bias: a number of lifecycle events and method design choices are likely to bias our population towards individuals maintaining similar characteristics across censuses (spatially, occupationally, in terms of co-residence or marital status). We therefore encourage any users of these linked data to remain aware of that limitation in their use of our data.

We would like to conclude by offering some thoughts regarding several developments we are pursuing as part of our recently funded research project `GeoLinking the Census`, which will build future versions of our dataset. Since submitting this article, we have investigated and implemented a range of new features that have increased the quality of our links. The details of these new approaches will be published separately and updated versions of the linked data will be made available on our Zenodo repository. Areas of ongoing research include: non-adjacent census linking; linking methods to be used in Marriage Index linking that could grasp the challenge of the implicit internal links better than FSEM does; embeddings to block and compare names rather than relying only on the usual set of phonetic and string distance functions; incorporating the Index to the Birth and Death registers into the linking itself, allowing for more longitudinal reconstruction; better ways to use occupation and occupational change as a linking criterion rather than the implicit distance provided by the nested HISCO typology.

This paper is a first step in this direction, and by providing both a method and a new set of links for British historical censuses, we hope it will open an area of research and debate.

Acknowledgments

This work used resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). We also benefited from funding from the Large Bid Support Fund

from the University of Cambridge, which allowed us to purchase the BMD data from FreeBMD and produce our verification work. We also acknowledge FindMyPast's support as original providers of I-CeM's underlying data.

Data availability

All linked tables are available on our public Zenodo repository, and our code is available on the linked GitHub repository.

- https://github.com/alexislitvine/geolinking
- https://doi.org/10.5281/zenodo.15727456

References

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration" [in en]. *American Economic Review* 102, no. 5 (August): 1832–1856. ISSN: 0002-8282, accessed April 25, 2024. https://doi.org/10.1257/aer.102.5.1832. https://pubs.aeaweb.org/doi/10.1257/aer.102.5.1832.

———. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration" [in en]. *Journal of Political Economy* 122 (3).

———. 2017. "To the New World and Back Again: Return Migrants in the Age of Mass Migration" [in en]. *ILR Review* 72 (2): 300–322. ISSN: 0019-7939, 2162-271X, accessed April 25, 2024. https://doi.org/10.1177/0019793917726981. http://journals.sagepub.com/doi/10.1177/0019793917726981.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data" [in en]. *Journal of Economic Literature* 59, no. 3 (September): 865–918. ISSN: 0022-0515, accessed April 28, 2024. https://doi.org/10.1257/jel.20201599. https://www.aeaweb.org/articles?id=10.1257/jel.20201599.

Abramitzky, Ran, Leah Platt Boustan, Harriet M. Brookes Gray, Katherine Eriksson, Santiago Pérez, Hannah M. Postel, Myera Rashid, and Noah Simon. 2025. *Finding John Smith: Using Extra Information for Historical Record Linkage.* Working Paper, July. Accessed July 14, 2025. https://doi.org/10.3386/w33999. https://www.nber.org/papers/w33999.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2020. "Linking Individuals Across Historical Sources: A fully Automated Approach" [in en]. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, no. 2 (April): 94–111. ISSN: 0161-5440, 1940-1906, accessed May 1, 2024. https://doi.org/10.1080/01615440.2018.1543034. https://www.tandfonline.com/doi/full/10.1080/01615440.2018.1543034.

Antonie, Luiza, Kris Inwood, Chris Minns, and Fraser Summerfield. 2020. "Selection Bias Encountered in the Systematic Linking of Historical Census Records" [in en]. *Social Science History* 44 (3): 555–570. ISSN: 0145-5532, 1527-8034, accessed November 14, 2024. https://doi.org/10.1017/ssh.2020.15. https://www.cambridge.org/core/product/identifier/S0145553220000152/type/journal_article.

Bailey, Martha, Peter Lin, Shariq Mohammed, Paul Mohnen, Jared Murray, Mengying Zhang, and Alexa Prettyman. 2023. "The creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database project." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56, no. 3 (July): 138–159. ISSN: 0161-5440, accessed July 28, 2024. https://doi.org/10.1080/01615440.2023.2239699. https://doi.org/10.1080/01615440.2023.2239699.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data" [in en]. *Journal of Economic Literature* 58, no. 4 (December): 997–1044. ISSN: 0022-0515, accessed October 8, 2024. https://doi.org/10.1257/jel.20191526. https://pubs.aeaweb.org/doi/10.1257/jel.20191526.

Beach, Brian, Joseph Ferrie, Martin Saavedra, and Werner Troesken. 2016. "Typhoid Fever, Water Quality, and Human Capital Formation" [in en]. *The Journal of Economic History* 76, no. 1 (March): 41–75. ISSN: 0022-0507, 1471-6372, accessed November 14, 2024. https://doi.org/10.1017/S0022050716000413. https://www.cambridge.org/core/product/identifier/S0022050716000413/type/journal_article.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley Wilbert. 2025. "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project." *Explorations in Economic History,* NBER Working Paper Series, 98. Accessed March 23, 2024. https://doi.org/https://doi.org/10.1016/j.eeh.2025.101717.

Calderón-Bernal, Liliana P., Diego Alburez-Gutierrez, and Emilio Zagheni. 2023. *Analysing biases in genealogies using demographic microsimulation* [in en]. Technical report WP-2023-034. Edition: 0. Rostock: Max Planck Institute for Demographic Research, August. Accessed November 18, 2024. https://doi.org/10.4054/MPIDR-WP-2023-034. https://www.demogr.mpg.de/en/publications_databases_6118/publications_1904/mpidr_working_papers/analysing_biases_in_genealogies_using_demographic_microsimulation_7875.

Chong, Michael, Diego Alburez-Gutierrez, Monica J. Alexander, and Emilio Zagheni. 2022. *Correcting historical mortality rate bias in big crowd-sourced online genealogies* [in en]. Edition: 0. Rostock, January. Accessed November 18, 2024. https://doi.org/10.4054/MPIDR-WP-2022-005. https://www.demogr.mpg.de/en/publications_databases_6118/publications_1904/mpidr_working_papers/correcting_historical_mortality_rate_bias_in_big_crowd_sourced_online_genealogies_7161.

Clark, Gregory, Neil Cummins, and Matthew Curtis. 2024. "Three new occupational status indices for England and Wales, 1800–1939." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 57, no. 1 (January): 41–66. ISSN: 0161-5440, accessed November 18, 2024. https://doi.org/10.1080/01615440.2024.2368458. https://doi.org/10.1080/01615440.2024.2368458.

Colasurdo, Andrea, and Riccardo Omenti. 2024. "Using online genealogical data for demographic research: An empirical examination of the FamiLinx database" [in en]. *Demographic Research* 51 (November): 1299–1350. ISSN: 1435-9871, accessed May 14, 2026. https://doi.org/10.4054/DemRes.2024.51.41. https://www.demographic-research.org/articles/volume/51/41.

Collins, William J., and Marianne H. Wanamaker. 2015. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants" [in en]. *The Journal of Economic History* 75, no. 4 (December): 947–992. ISSN: 0022-0507, 1471-6372, accessed November 14, 2024. https://doi.org/10.1017/S0022050715001527. https://www.cambridge.org/core/product/identifier/S0022050715001527/type/journal_article.

Dahl, Christian Møller, Torben Johansen, and Christian Vedel. 2024. *Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE.* ArXiv:2402.13604 [cs], April. Accessed November 14, 2024. http://arxiv.org/abs/2402.13604.

Day, Joe. 2019. "The Process of Internal Migration in England and Wales, 1851-1911: Updating Ravenstein and the Step-Migration Hypothesis" [in en]. *Comparative Population Studies* 44. ISSN: 1869-8999, accessed April 28, 2024. https://doi.org/10.12765/CPoS-2020-13. https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/374.

Day, Joseph. 2022. "The accuracy and precision of birthplace reporting in the 1851–1911 censuses: Place as a component of identity in nineteenth-century England and Wales" [in en]. _Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/psp.2537, *Population, Space and Place* 28 (5): e2537. ISSN: 1544-8452, accessed December 5, 2024. https://doi.org/10.1002/psp.2537. https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.2537.

Diduch, Emma. 2024. "The Record Linking Glass Ceiling. Applying Automated Methods to the Census and Women's Marriage Records, 1881–1911." *Historical Life Course Studies* 14 (November): 126–143. ISSN: 2352-6343, accessed November 14, 2024. https://doi.org/10.51964/hlcs19189. https://hlcs.nl/article/view/19189.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records" [in en]. *American Political Science Review* 113, no. 2 (May): 353–371. ISSN: 0003-0554, 1537-5943, accessed August 10, 2024. https://doi.org/10.1017/S0003055418000783. https://www.cambridge.org/core/journals/american-political-science-review/article/using-a-probabilistic-model-to-assist-merging-of-largescale-administrative-records/DB2955F64A1F4E262C5B9B26C6D7552E.

Feigenbaum, James. 2016. *A Machine Learning Approach to Census Record Linking* [in en]. https://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking.

Feigenbaum, James J, Jonas Helgertz, and Joseph Price. 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 (April): 101656. ISSN: 0014-4983, accessed April 25, 2026. https://doi.org/10.1016/j.eeh.2025.101656. https://www.sciencedirect.com/science/article/pii/S0014498325000038.

Fellegi, Ivan, and Alan Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64.

Ferrie, Joseph P. 2005. "History Lessons: The End of American Exceptionalism? Mobility in the United States Since 1850" [in en]. *Journal of Economic Perspectives* 19, no. 3 (August): 199–215. ISSN: 0895-3309, accessed November 14, 2024. https://doi.org/10.1257/089533005774357824. https://pubs.aeaweb.org/doi/10.1257/089533005774357824.

———. 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" [in en]. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29, no. 4 (October): 141–156. ISSN: 0161-5440, 1940-1906, accessed November 14, 2024. https://doi.org/10.1080/01615440.1996.10112735. https://www.tandfonline.com/doi/full/10.1080/01615440.1996.10112735.

Fu, Zhichun, H.M. Boot, Peter Christen, and Jun Zhou. 2014. "Automatic Record Linkage of Individuals and Households in Historical Census Data" [in en]. *International Journal of Humanities and Arts Computing* 8, no. 2 (October): 204–225. ISSN: 1753-8548, 1755-1706, accessed November 14, 2024. https://doi.org/10.3366/ijhac.2014.0130. https://www.euppublishing.com/doi/10.3366/ijhac.2014.0130.

Ghosh, Arkadev, Sam Il Myoung Hwang, and Munir Squires. 2024. "Links and Legibility: Making Sense of Historical U.S. Census Automated Linking Methods" [in en]. *Journal of Business & Economic Statistics* 42, no. 2 (April): 579–590. ISSN: 0735-0015, 1537-2707, accessed October 9, 2024. https://doi.org/10.1080/07350015.2023.2205918. https://www.tandfonline.com/doi/full/10.1080/07350015.2023.2205918.

Gross, Matthew, and Michael Mueller-Smith. 2021. *Modernizing Person-Level Entity Resolution with Biometrically Linked Records* [in en].

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles, and Catherine A. Fitch. 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, no. 1 (January): 12–29. ISSN: 0161-5440, accessed October 9, 2024. https://doi.org/10.1080/01615440.2021.1985027. https://doi.org/10.1080/01615440.2021.1985027.

Hollingsworth, T.H. 1976. "Genealogy and Historical Demography." *Annales de démographie historique,* 167–170. ISSN: 0066-2062, accessed November 18, 2024. https://www.jstor.org/stable/44384526.

Kates, Sean, Eric Manning, Tali Mendelberg, and Omar Wasow. 2025. "Plutopopulism: Wealth and Trump's Financial Base" [in en]. *American Political Science Review* (April): 1–18. ISSN: 0003-0554, 1537-5943, accessed June 6, 2025. https://doi.org/10.1017/S0003055425000061. https://www.cambridge.org/core/journals/american-political-science-review/article/plutopopulism-wealth-and-trumps-financial-base/04A0071D849FBC3E1DEEB2962A6B977F.

Leeuwen, Marco van, Ineke Maas, and Andrew Miles. 2002. *HISCO: Historical International Standard Classification of Occupations* [in en]. Google-Books-ID: EMPtAAAAIAAJ. Leuven University Press. ISBN: 978-90-5867-196-7.

Linacre, Robin. 2026. *Training rationale - Splink.* Accessed July 17, 2025. https://moj-analytical-services.github.io/splink/topic_guides/training/training_rationale.html.

Linacre, Robin, Sam Lindsay, Theodore Manassis, Zoe Slade, and Tom Hepworth. 2022. "Splink: Free software for probabilistic record linkage at scale." *International Journal of Population Data Science* 7, no. 3 (August). ISSN: 2399-4908, accessed August 10, 2024. https://doi.org/10.23889/ijpds.v7i3.1794. https://ijpds.org/article/view/1794.

Long, Jason, and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850" [in en]. *American Economic Review* 103, no. 4 (June): 1109–1137. ISSN: 0002-8282, accessed November 26, 2024. https://doi.org/10.1257/aer.103.4.1109. https://www.aeaweb.org/articles?id=10.1257/aer.103.4.1109.

Malin, JC. 1935. "The turnover of farm population in Kansas." *Kansas Historical Quarterly* 4 (4): 339–372.

Mattheis, Ross. 2024. *Spurious Mobility in Imperfectly Linked Historical Data* [in en], May.

Mitchell, Brian R. 2011. *British Historical Statistics* [in eng]. Cambridge: Cambridge University Press. ISBN: 978-1-107-40244-7.

Newton, Gill, and Robert Bennett. 2020. "Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM" [in en] (February). Accessed August 18, 2025. https://doi.org/10.17863/CAM.50178. https://www.repository.cam.ac.uk/handle/1810/303101.

Ó Gráda, Cormac, Tyler Anbinder, Dylan Connor, and Simone A. Wegge. 2023. "The problem of false positives in automated census linking: Nineteenth-century New York's Irish immigrants as a case study." _Eprint: https://doi.org/10.1080/01615440.2024.2312293, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56, no. 4 (October): 240–259. ISSN: 0161-5440, accessed April 27, 2024. https://doi.org/10.1080/01615440.2024.2312293. https://doi.org/10.1080/01615440.2024.2312293.

Pérez, Santiago. 2019. "Intergenerational Occupational Mobility across Three Continents." *The Journal of Economic History* 79 (2): 383–416. ISSN: 0022-0507, accessed April 25, 2026. https://www.jstor.org/stable/26787191.

Philips, Lawrence. 2009. *Metaphone 3.* http://www.amorphics.com/buy_metaphone3.html.

Pooley, Colin G. 1979. "Residential Mobility in the Victorian City." *Transactions of the Institute of British Geographers* 4 (2): 258. ISSN: 00202754, accessed November 19, 2024. https://doi.org/10.2307/622038. https://www.jstor.org/stable/622038?origin=crossref.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. "Combining family history and machine learning to link historical records: The Census Tree data set." *Explorations in Economic History* 80 (April): 101391. ISSN: 0014-4983, accessed October 9, 2024. https://doi.org/10.1016/j.eeh.2021.101391. https://www.sciencedirect.com/science/article/pii/S0014498321000024.

Proffit, Guillaume. 2024. "Geocoding British Censuses for England and Wales (1851-1911)" [in en] (August). Accessed November 17, 2024. https://doi.org/10.17863/CAM.111637. https://www.repository.cam.ac.uk/handle/1810/373069.

Reid, Alice, Sophy Arulanantham, Joe Day, Eilidh Garrett, Hannaliis Jaadla, and Martin Lucas-Smith. 2018. *Populations Past: Atlas of Victorian and Edwardian Population.* www.populationspast.org/.

Reid, Alice, Ros Davies, and Eilidh Garrett. 2002. "Nineteenth-Century Scottish Demography From Linked Censuses and Civil Registers: A 'Sets of Related Individuals' Approach" [in en]. *History and Computing* 14, nos. 1-2 (March): 61–86. ISSN: 0957-0144, 2053-7832, accessed August 18, 2025. https://doi.org/10.3366/hac.2002.14.1-2.61. https://www.euppublishing.com/doi/10.3366/hac.2002.14.1-2.61.

Reid, Alice, Eilidh Garrett, and Simon Szreter. 2016. "Residential Mobility and Child Mortality in Early Twentieth Century Belfast" [in en]. In *New Approaches to Death in Cities during the Health Transition,* edited by Diego Ramiro Fariñas and Michel Oris, 55–76. Cham: Springer International Publishing. ISBN: 978-3-319-43001-0 978-3-319-43002-7, accessed August 30, 2025. https://doi.org/10.1007/978-3-319-43002-7_4. http://link.springer.com/10.1007/978-3-319-43002-7_4.

Rijpma, Auke, Jeanne Cilliers, and Johan Fourie. 2020. "Record linkage in the Cape of Good Hope Panel" [in en]. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, no. 2 (April): 112–129. ISSN: 0161-5440, 1940-1906, accessed November 14, 2024. https://doi.org/10.1080/01615440.2018.1517030. https://www.tandfonline.com/doi/full/10.1080/01615440.2018.1517030.

Robin Linacre. 2021. *Probabilistic Record Linkage.* https://www.robinlinacre.com/probabilistic_linkage/.

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. "Historical Census Record Linkage" [in en]. *Annual Review of Sociology* 44 (July): 19–37. ISSN: 0360-0572, 1545-2115, accessed May 24, 2024. https://doi.org/10.1146/annurev-soc-073117-041447. https://www.annualreviews.org/content/journals/10.1146/annurev-soc-073117-041447.

Schurer, Kevin. 1989. "A Note Concerning the Calculation of the Singulate Mean Age at Marriage." *Local Population Studies* 43. https://www.localpopulationstudies.org.uk/PDF/LPS43/LPS43_1989_67-70.pdf.

Schürer, Kevin. 2024. *I-CeM Lookup Table - Occupation code,* June. https://doi.org/10.17863/CAM.106514.

Schürer, Kevin, Eddy Higgs, and FINDMYPAST LIMITED. 2024. *Integrated Census Microdata (I-CeM) Names and Addresses, 1851-1911: Special Licence Access.* SN: 7856. https://doi.org/http://doi.org/10.5255/UKDA-SN-7481-3. http://doi.org/10.5255/UKDA-SN-7481-3.

Schürer, Kevin, Tatiana Penkova, and Yanshan Shi. 2015. "Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911." _Eprint: https://doi.org/10.1080/01615440.2015.1010028, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 48, no. 4 (October): 195–213. ISSN: 0161-5440, accessed April 28, 2024. https://doi.org/10.1080/01615440.2015.1010028. https://doi.org/10.1080/01615440.2015.1010028.

Schürer, Kevin, Alexander Wakelam, and FINDMYPAST LIMITED. 2024. *Integrated Census Microdata (I-CeM) Names and Addresses, 1921: Special Licence Access.* SN: 7481, June. https://doi.org/http://doi.org/10.5255/UKDA-SN-9281-1.

Solon, G. 1992. "Intergenerational income mobility in the United States." *American Economic Review* 82 (3): 393–408.

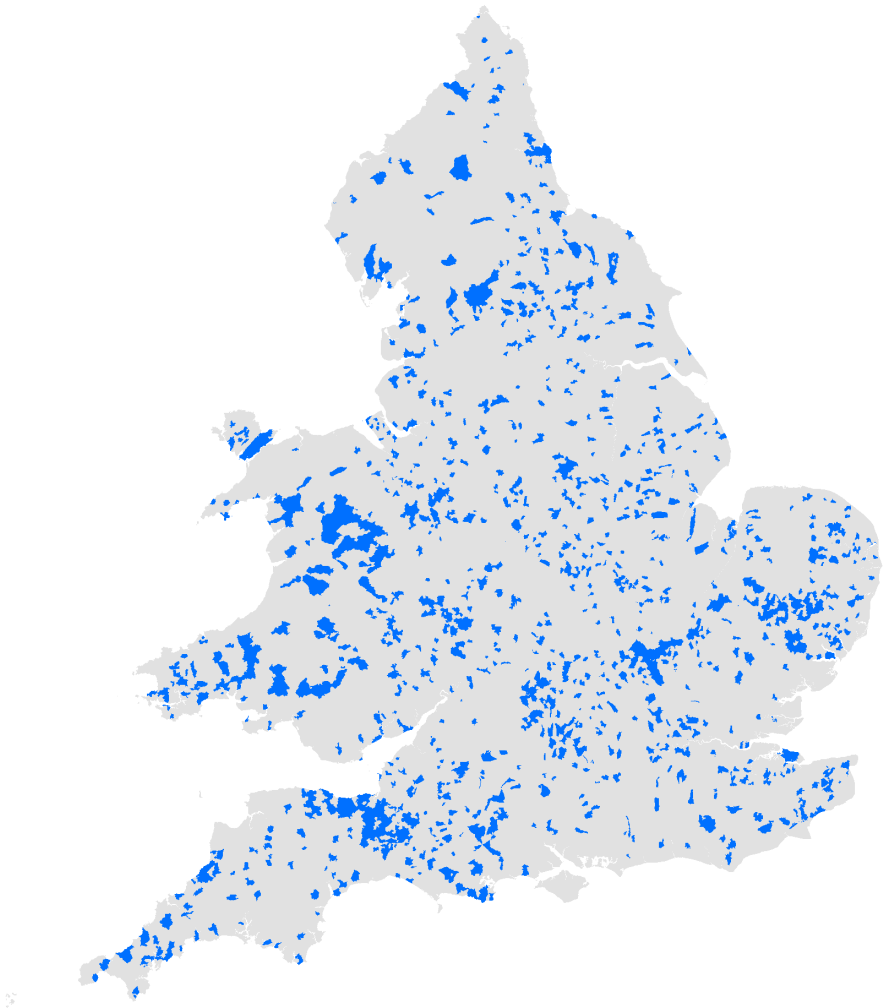
Stelter, Robert, and Diego Alburez-Gutierrez. 2022. "Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics" [in en]. *Proceedings of the National Academy of Sciences* 119, no. 10 (March): e2120455119. ISSN: 0027-8424, 1091-6490, accessed November 18, 2024. https://doi.org/10.1073/pnas.2120455119. https://pnas.org/doi/full/10.1073/pnas.2120455119.

Thernstrom, Stephan, ed. 1964. *Poverty and progress: social mobility in a nineteenth century city*. Cambridge: Harvard University Press. ISBN: 978-0-674-69501-6 978-0-674-04431-9.

Appendix

Alternative Birthplaces

Figure 11: **Mapping of alternative Birthplaces**



Verification Outcomes Distribution

Table 11: **Error and uncertainty rate by age and sex (%, with uncertainty in parentheses)**

Age band	Female	Male	Overall
10–19	2.15(1.91)	13.97(3.85)	8.55(2.96)
20–29	2.18(6.18)	7.76 (4.31)	5.68(5.01)
30–39	2.01(6.02)	5.65 (3.57)	4.1 (4.62)
40–49	0.9 (2.24)	2.66 (5.32)	1.85(3.91)
50–59	1.2 (1.2)	4.93 (4.93)	3.25(3.25)
60–100	0.52 (0)	2.92 (0.83)	1.85(0.46)
Overall	2.21(3.25)	9.54 (3.99)	6.4 (3.67)

Table 12: **Error and uncertainty rate by marital status and sex (%, with uncertainty in parentheses)**

Marital status	Female	Male	Overall
Married	1.6 (4.27)	4.12 (3.84)	3.09(4.02)
Not married	2.82 (2.3)	15.66(4.18)	9.89(3.34)
Overall	2.21(3.25)	9.54 (3.99)	6.4 (3.67)

Table 13: **Error and uncertainty rate by HISCO group (%, with uncertainty in parentheses)**

HISCO group	Overall
1	4.62(1.54)
2	2.08(4.17)
3	4.76(1.36)
4	3.65(4.17)
5	1.09(0.36)
6	7.46(1.49)
7	7.05(4.89)
8	9.71 (3.6)
9	5.67(4.38)
N/A	7.63(4.21)
Overall	6.4 (3.67)

Blocking Rules

- **Stage 1:**
  - NYSIIS first name, NYSIIS surname, age
  - NYSIIS first name, NYSIIS surname, county of birth

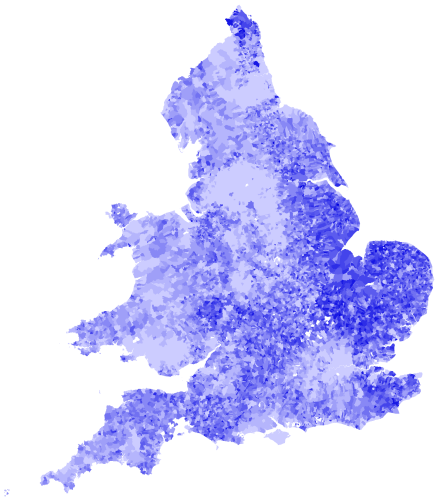
  - NYSIIS first name, NYSIIS surname, HISCO top-level group
  - NYSIIS first name, NYSIIS surname, HoH first name
  - NYSIIS first name, birthplace, age
  - NYSIIS surname, birthplace, age
  - NYSIIS first name, surname initial, age, county of birth
  - NYSIIS surname, first name initial, age, county of birth
  - HoH first name, birthplace, age
  - HoH first name, place of birth, first name initial
  - Geocoded within 60 km, NYSIIS surname, age, HoH first name

- **Stage 2**
  - co-resident with a man (linked in Stage 1) across two censuses

- **Stages 3A and 3B**
  - NYSIIS first name, NYSIIS surname, age
  - NYSIIS first name, NYSIIS surname, county of birth
  - NYSIIS first name, age, birthplace
  - NYSIIS surname, age, birthplace
  - Birthplace, age, surname initial
  - Birthplace, age, first name initial
  - First name initial, NYSIIS surname, birthplace
  - Surname initial, NYSIIS first name, birthplace

Distribution of Agricultural Labourers and General Labourers in 1901

Figure 12: **Distribution of Agricultural Labourers and General Labourers in 1901**



(a) **Distribution of Agricultural Labourers - 1901**



(b) **Distribution of General Labourers - 1901**