Census Linking:
Confidence in the Absence of Ground Truth

Hillary Vipond [1]

Abstract

Linking historical census datasets faces significant challenges. The first order issue is the risk of false positives, which can distort key downstream coefficients of interest, particularly for migration and occupation. These errors then result in incorrect conclusions being drawn in new research in economic history. In this paper, I propose an approach which makes it possible to assess the share of false positives in panel datasets generated by any given linking algorithms. It does not require ground truth data. Instead, it exploits systematic biases induced by false positives in outcomes of interest to construct a litmus testing process. Using this framework, I take a multi-model comparison approach to explore the impacts of model specification choices on the linked panel datasets. Three findings emerge. Firstly, historical British is substantially less exposed to false positives than American census data. Secondly, linking full census datasets yields more accurate results than using population subsets, especially for American data. Finally, a case study of immigrant tailors from Russia demonstrates how this approach facilitates linking challenging segments of the population. I propose a generalizable falsification test derived from this methodology, which can serve as a data quality check for any census linking algorithm.

[1]Particular thanks to James Fenske for a discussion on how to communicate that the process of census linking is salient to outcomes of interest. Thanks to Eric Protzer, Christian Vedel, Patrick Wallis, Chris Minns, James Feigenbaum for comments and discussions. This draft is an extract from one of the papers in my PhD thesis

1 Introduction

With a windfall of newly digitized, individual level, cross-sectional census data for multiple countries having been recently made available, historians have set about assembling it into panel data in order to address a wide array of questions. These include questions on inter-generational social mobility (ABE, 2014; Abramitzky et al., 2012; Connor, 2019; Connor and Storper, 2020; Pérez, 2017), the causes and consequences of migration (Boustan, 2016; Boustan et al., 2010; Abramitzky et al., 2020a), and the impacts of racial discrimination over time (Dahis et al., 2019).

Recent debates have centered on the methodological challenges of linking historical census data, particularly for full-count datasets like those from the United States.The most well-known, and widely used, approach to census linking is the approach created by Ran Abramitzky, Leah Boustan, and Katherine Eriksson (hereafter ABE) (Abramitzky et al., 2012). This is a straightforward deterministic linking algorithm, and the ABE team opened up new frontiers in the field when they made it available as a package in R and Stata. The linking process has been at the center of a methods debate for the last several years. Several papers have claimed that the share of false positives in the panel datasets constructed from the ABE deterministic linking algorithm is high. Most recently, a set of papers working with "ground truth" hand-linked samples have made similar findings. This has then led into a wider discussion on census linking, and a range of approaches have been proposed. It is possible to conduct census linking with deterministic, probabilistic, and machine learning methods. The problem is how to validate them - what criteria should adjudicate the performance of a linking algorithm?

I argue that, while there are a set of challenges with census linking, false positives should be considered of first importance. A false positive, sometimes called a "bad match", occurs when we incorrectly identify a match between the "John Smith" we have in one census cross-section and a "John Smith" in the second census. False positives pose a particular issue because they drive results. For example, a panel dataset with a high share of false positives will always find that a disproportionately large share of people "migrated". This is the result of finding your original "John Smith" in a different geographical location than your original one, which is unsurprising given that it is not the same person. The other two major concerns in census linking are false negatives and bias. False negatives occur when a match is not accepted even though it is in fact a true match. This results in unnecessarily low match rates, and, therefore, smaller panel datasets. As statistical power is always valuable, avoiding false negatives is important. However, high match rates should be pursued only subject to a quality requirement, which means that identifying false positives must take precedence. Finally, there is concern as to whether different linking specifications impact differentially on bias in the linked dataset. In this paper I will not directly address either false negatives or bias [1].

This paper proposes an approach which enables an assessment of the share of false positives generated by different census linking algorithms. It is an approach which circumvents a requirement for costly "ground truth" data. It exploits the findings from (Helgertz et al., 2022), that false positives will necessarily bias outcomes of interest, to construct a litmus test for false positives in any linked dataset. This framework is then twinned with a one-at-a-time model comparison approach, run for a set of 45 census linking models. This is an exercise in comparative statics, which helps identify which choices in linking specifications drive the introduction of false positives.

This results in three key findings. Firstly, British data is considerably less exposed to false positives than American data, across any range of linking algorithms.This is primarily the result of additional information on place of birth, which the historical British census data has at the parish

[1]However, I include an appendix on both, showing the impacts of model choice on false negatives and on representativeness.

level. Secondly, linking the full census is shown to be best practice. Attempting to link subsets of the population —for instance, focusing only on individuals in a specific occupation or geographic region — introduces additional false positives. This is exacerbated with American level data, where the absence of parish-level birthplace information compounds the risk. Full-census linking, followed by filtering the linked dataset to the subset of interest, produces a more reliable dataset. Finally, the paper presents a case study, linking Russian immigrants in historical British census data. This is a particularly challenging group to match, as information on birthplace is often available only at the country level. The case study illustrates how insights gained from the litmus testing approach can help address specific and unique challenges in linking difficult subgroups.

In the discussion I consider whether census linking can be conceptualized as an optimization problem, where the goal is to maximize the size of the constructed panel dataset while maintaining strict quality standards. I suggest that the share of false positives should serve as the primary constraint, and that the litmus test put forward in this paper can serve to do so. This approach would make it possible to run a data quality check on any given census linking algorithm, thereby making it possible to maximize sample size without loss of quality.

The paper is organized as follows. Section 1 reviews the literature on census linking. Section 2 describes the data. Section 3 walks through the process of deterministic census linking, and discusses the time invariant characteristics on which all linking relies. Section 4 introduces the methodology. I introduce the litmus testing approach, and explain how it will be employed in the following sections. I explain how I will take a one-at-a-time-model comparison approach, which is an exercise in comparative statics, to understand the impacts of choices made in census linking specifications on the panel datasets constructed. The litmus test will work in tandem with the set of 45 generated panel datasets, providing an assessment of the share of false positives in those datasets even in the absence of ground truth information. In section 5 I turn to the results of making use of this approach. I have three findings: that British data is substantially less exposed to false positives, that sub-samples of the population should not be linked by themselves (particularly with American level census data), and new approaches to linking particularly difficult groups. Section 6 discusses implications of the findings and next steps. Section 8 concludes.

2 Literature

3 Data

This paper exclusively makes use of the full-count deanonymized British census data, digitized by The Integrated Census Microdata Project (I-CeM). ICeM, based at Essex University, has published a standardised digital dataset of the decennial English Census data between 1851-1911, excluding the census year 1871. This has made deanonymised, individual level English census data available for the first time. It is not panel data: no unique key is assigned to each individual in the dataset. This is cross-sectional data, providing individual level observations for each census year. In total, it provides nearly 180 million individual records (Higgs et al., 2013).

The ICeM data provides data on a set of time invariant information, which is critical to the matching process. This includes first, middle, and last name, age, place of birth, and family members residing at the address. In the remainder of this paper, I focus on linking the 1861 and 1881 census records [2]. The exception to this is the case study which is linking Russian immigrants to the UK.

[2] I am assuming that my findings on model choice generalize to matches in other years. It will be possible to check this assumption

The data used for that exercise is from 1881-1891. This is because the first wave of mass immigration from Russia took place during that period.

The census linking process is 'finding' the people who were bootmakers in 1861 amongst the full population of 1881. Once this matching process is completed, it enables researchers to observe the same individuals over two decades. This provides information on changes in residence and occupation over time. These linked datasets have applications for studying the labor market impacts of technological change, as well as questions related to social mobility, migration, health, and inequality, amongst many others.

4 Census Linking Process

The thought experiment which is the subject of this paper requires an understanding of the process of census linking. I simplify this process to its most basic components. Firstly, all census linking processes rely on time invariant characteristics in the data. I describe these below. Secondly, non-unique names must at some point be dropped. I discuss the different definitions of uniqueness, and explain the importance of distinct names in linking. Finally, I provide a flow diagram which illustrates the basic deterministic linking process, step by step.

4.0.1 Time Invariant Characteristics

The matching process requires a set of time-invariant variables. In the British data these are: county of birth, parish of birth, full name, date of birth, and family information (particularly spouse). Multiple specifications of each of these variables are available. I will briefly discuss these.

Firstly, the name of the individual. First and last names are available in the ICeM deanonymized data. Linking is often referred to as "nominal linking", which reflects that the names of individuals are often selected as the main variable for identifying links [3] Unsurprisingly, many approaches to matching on names have been discussed. There is widespread agreement that minor misspellings of names was a fairly common event in the 19th century, and that, therefore, a requirement for an exact match on name could be expected to disqualify a large number of potentially good matches (the extent of this is seldom, if ever, quantified). In this paper, I will focus on three different approaches to matching on name. The first is matching on the exact name, the second is matching on the ABE cleaned name, and the third is matching on a phonetic version of the name. It is also possible to calculate string distances and run Jaro-Winkler or Levenshtein distance calculations. More promising approaches are available through making use of sets of n-grams and cosine similarities. Names make linking women somewhat more challenging. Women in 19th century Britain changed their last names when they married. This makes the surname for women a time varying feature of the data. This is often viewed as a major issue, and leads many researchers to not link women at all.

Secondly, date of birth. Date of birth is not given in the ICeM census records. However, the age of the individual at the time of the census taking is recorded, and birth year can therefore be derived. Although year of birth does not change over time, individuals in the 19th century did not always recall their age with great accuracy. The age variable could therefore be set to require an exact match, or a window of ± 1 year, or of ± 2 years, or indeed beyond. Another consideration is that age data is sometimes not recorded, and therefore any algorithm which requires a match on date of birth will drop some potential matches.However, age is almost always given in the census

[3] There are alternative approaches, which do not center the match on name. These are beyond the scope of this paper

data; see Appendix A on missing data for a breakdown of the share of observations missing age information by census year.

Thirdly, county of birth. This is given in the data, and is very seldom missing. It is a particularly important variable, as the linking process "blocks" on county of birth. This means that, when linking takes place, we seek for potential matches only amongst the set of people born in the same county. Little research has been done on how accurately individuals remembered and recorded their county of birth, but it is not unreasonable to assume that this is somewhat more memorable than parish of birth.

Fourthly, parish of birth. ICeM includes information on parish of birth. This is extremely helpful. However, there are two challenges. Firstly, as with age, it is sometimes not reported accurately. Several teams are in the process of geo-coding ICeM addresses, and this will allow for a more fuzzy match, by distance, as is done for age. However, the more immediate concern is that the data is often missing or inaccurate. [4]

Fifthly, and finally, ICeM data includes information on family members, and the composition of the household. A unique household identifier is given for family groups. This is unique only within each parish. Information on family members is somewhat different in nature to the other time-invariant characteristics described above. Firstly, children grow up and often leave the home. So, while the presence of the same children in one ten year period can be helpful additional information, it is also the case that the absence of children is not indicative of a false match. Spouses, on the other hand, are usually members of the same household for many decades, until death. However, spouses may be visiting friends or family during the census taking, or have died, and therefore the absence of a spouse does not indicate a false match. On the other hand, the presence of the children, or of a spouse, does indicate a true match. Equally, the presence of the wrong spouse indicates a false match. While missing data is not a concern in the same way it is for the other time invariant characteristics, it is worth noting that a large number of people are not living in their own family groups, and therefore the addition of family information to the linking process will not improve their chance of being matched. This will result in some bias in the final sample [5].

4.0.2 Distinctness

The amount of information available on an individual determines how feasible it is to differentiate between people with the same name. In 1861, the most common male name is John, accounting for about 14% of the men, or 1.43 million observations. The most common female name is Mary, accounting for about 16% of women, or 1.65 million observations. Fortunately, full names are much more distinct. When I concatenate the first and last names together, the most common male and female names are now: Mary Smith, with 22 048 instances, and John Jones, with 21 509, and both account for about 0.22% - less than 1% - of all names of their gender. See the histogram in Figure 5 below, which shows the distribution of frequency for full male and female names in 1861 and 1881 at the national level.

[4] See Appendix A for a summary of missing data on parish of birth

[5] The extent of this bias has not yet been assessed.

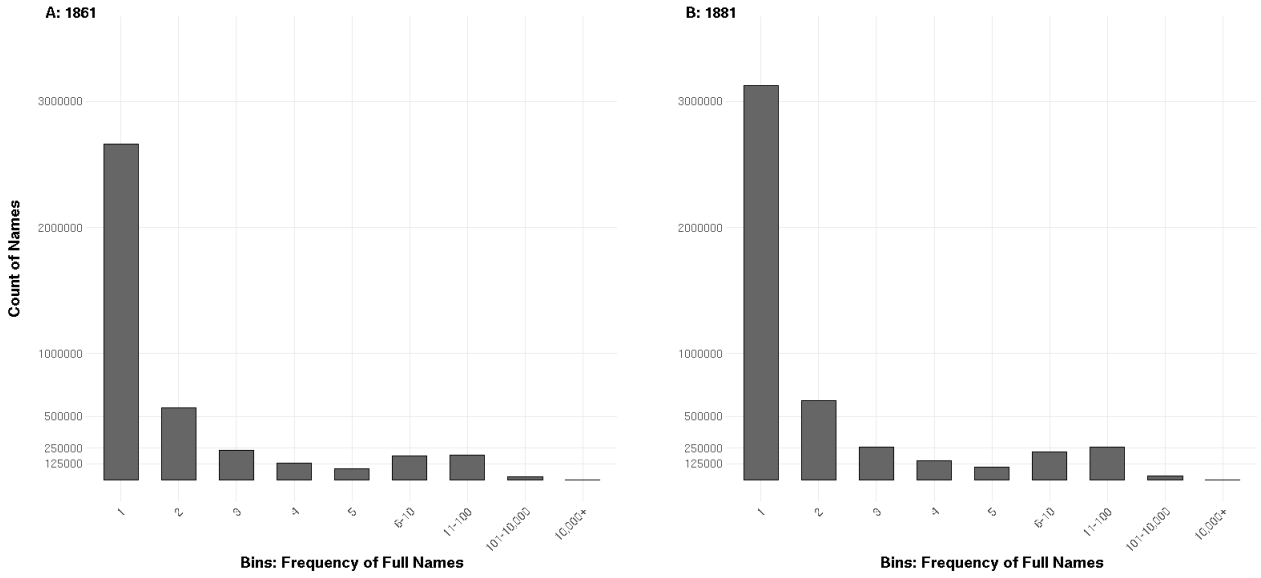


Figure 1: Frequency of Distinct Full Names, National Level
Note: Deanonymized data from the ICeM Project

While a substantial number of individuals have names that are fully distinct at the national level, many do not. However, with more information these people can be differentiated. The first and most important of these additional pieces of information is more granularity in the geographical level. In Table 1, below, I show the share of people who have unique names at the national, county, and parish level of birth. So, while there may be 21 000 John Jones born in Britain, there are only 1074 born in the largest county - Lancashire, and 561 born in the the largest parish, Liverpool. Table In Table 1 shows the share of people with unique names at each level of geography. Adding information on age makes further improvements.

Table 1: Distinct Individuals: Geography and Age

	Census 1861		Census 1881	
Names	Number	Share (%)	Number	Share (%)
National	2,201,813	14.48%	2,678,789	12.39%
National with Age	8,244,434	54.24%	10,785,790	49.90%
County	5,237,753	34.45%	6,124,968	28.33%
County with Age	12,494,510	82.22%	17,105,318	79.15%
Parish	10,989,048	72.32%	14,715,781	68.06%
Parish with Age	14,859,722	97.78%	21,002,869	97.15%
Total Observations	15,196,759	100%	21,619,745	100%

Note: The shares are calculated based on the total number of observations for each census year in the cleaned data. This means that missing data has already been dropped. Uniqueness is calculated on cleaned name.

4.0.3 Flow Diagram

The diagram in Figure 2 depicts the linking process for a simplified deterministic approach to census linking historical census data. The process has two stages: a cleaning process, and then the linking process. Note that I block on county of birth. This is a fairly standard approach, and means that I perform this process for one county of birth at a time. Blocking on county of birth means that only people born in the same county are considered as possible linking matches. This substantially reduces the computational power needed to run the linking process, which is important in this context, as I am running dozens of models.

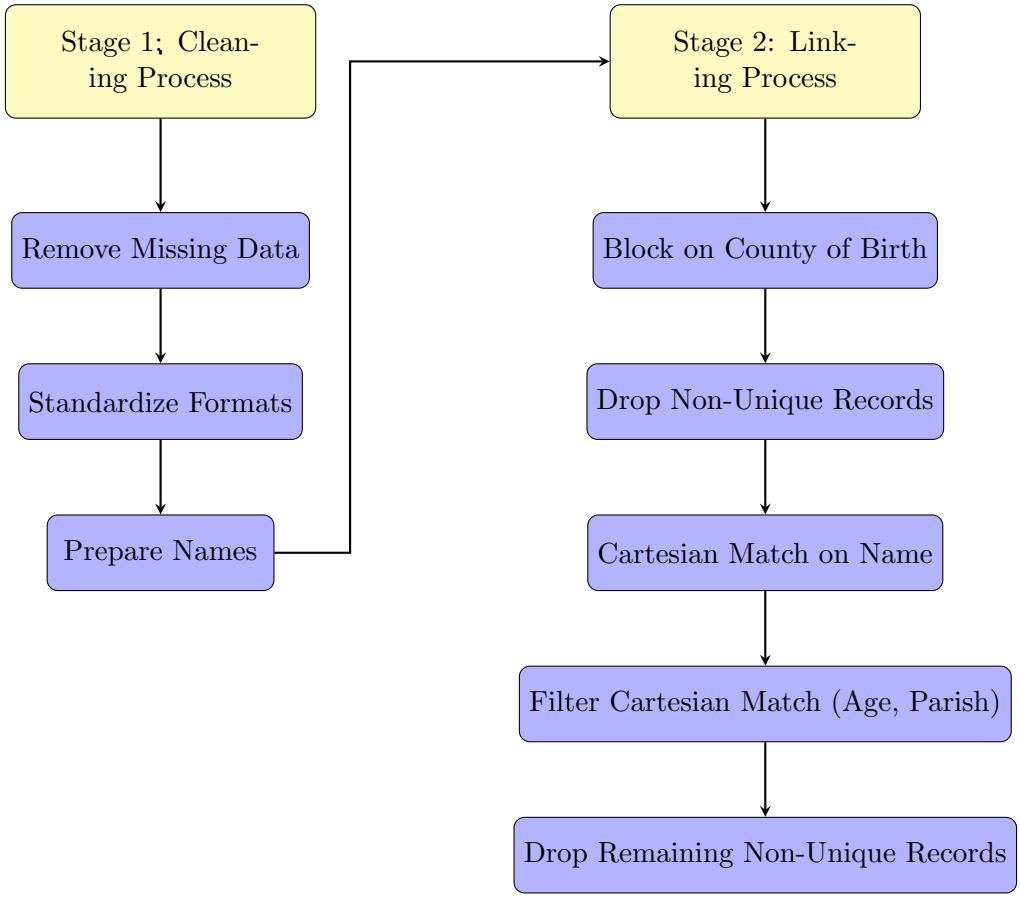


Figure 2: Census Linking Process

In the first stage the cleaning includes the following steps. Firstly, I drop missing date. I drop all observations for which there is no first name or no last name [6]. I also drop all observations for which either name has two characters or less, age is not given, or county of birth is missing. I drop rows where parish of birth is not given. Secondly, I standardize formats. I convert names and addresses to strings, convert to lowercase, and trim white spaces. Thirdly, I prepare the names. I concatenate first and last names to construct an exact name string. I concatenate the cleaned

[6]This is recorded differently across census years in the ICeM data, please see Appendix A on missing data

first and last names to construct a clean name string. I run a phonetic conversation on the exact first name, and the exact last name, to produce a phonetic matching option. I concatenate this to produce the exact phonetic name.

In the second stage I link. This involves five key steps in the deterministic linking approach. Firstly, I block on county of birth. Secondly, I drop all individuals whose information is an exact duplicate. Thirdly, I join on the basis of a name match alone. This creates a large Cartesian join, consisting of dyads of all possible matches. Fourthly I filter: any potential match that does not match on parish of birth or the age window is dropped. Finally, any exact duplicates which remain in the panel dataset are dropped.

More detailed information on each step in the linking stage will clarify this process. The first step is to block on county of birth, meaning I will only look for matches within the same county of birth. This increases computational efficiency. If I do not impose this limitation, I am searching for potential matches for an individual across the entire country, creating arrays of hundreds if not thousands of potential matches. By restricting the search to those born in the same county, the array of possibilities is significantly reduced, making the process more efficient. Although blocking on the county of birth will potential matches if the county of birth is missing, this is a minor issue, as most records include data on county of birth (see Appendix A on missing data.

The second step is to drop all names that are not differentiable. As discussed, individuals are distinguishable to varying degrees depending on the granularity of the place of birth data available. Names that cannot be differentiated within the specified county of birth are excluded from further consideration, ensuring that only those with unique or clearly distinguishable information proceed to the next stage of the linking process. One of the - several - advantages of formally including this step in the linking process is that it makes visible how many individuals are being dropped from the match simply because they are non-unique. The options I will include in methodology are: unique records at the county level ("Unique 1"), unique records at county level when combined with data on age ("Unique 2"), and unique records at the parish level when combined with data on age ("Unique 3").

The third step is to match on name. I require an exact match. The options are as discussed in the section on time invariant characteristics above. The choices are: exact name, ABE cleaned name [7], or phonetic name. It is possible to fuzzy match on names, and many census linking algorithms do so. Fuzzy matching is generally a matter of calculating string distance between names: for instance, rib and and jib are only one letter apart. Distance can be calculated in various ways. Options include: Levenshtein Distance, Jaro-Winkler Distance, Hamming Distance, Cosine Similarity (working with the vectors of n-grams), Damerau-Levenshtein Distance.

In this paper I choose not to make use of string distance approaches, for three reasons. Firstly, computational power: most of these approaches operate on a pairwise basis and compare the reference string (one name) to the full set of other possible names, creating a dyad pair for every possible match and calculating a string distance for those. Even though I am blocking on county of birth, a county of a million people implies creating a matrix of one trillion distance calculations. Given that this is done for each county, it makes the linking process very slow. Secondly, it is unclear that any of the fuzzy linking processes significantly improve linking rates. Many of the names in the census data are not misspelled, so this approach might not be beneficial, and could introduce false positives. For example, "rib" and "jib," while only one character apart, are very different words. Thirdly, clarity is valuable: if names are pre-cleaned and then matched exactly, the transformations they have undergone are clear. This is not the case when a distance linking process is wrapped into

[7] the ABE name cleaning process is largely a matter of separating out the initial from the first name, and replacing nicknames with full first names. If I use my own name cleaning process the results are very similar

the full linking process [8].

Once matching has been done on name, each individual record from the first dataset then has a set of possible matches. The fourth step is to drop potential pairs in which age and parish of birth is not a match. Finally, in the fifth step, I drop any name for which there still remains multiple possible matches. This is the most basic approach to deterministic record linking with historical census data. Having presented the steps to a basic deterministic linking process, I turn to discuss the model comparison approach this paper will follow.

Note that, by definition, an individual must be unique on some set of variables in order to be matched. Even in Liverpool, John Jones will never be matched: there are too many others with the same name. For instance, if we have two entries of Mary Smith, born in Devon, 25 years old, in 1861, there is no way at all of understanding what is the correct match for these people in 1881. The table of distinctness therefore shows the absolute upper bound of possibility for match rates with different amounts of information. Note the extent to which the inclusion of age is vital in being able to distinguish individuals from one another.

As discussed here, individuals must be distinct in order to be matched. However, distinctness can be set at different levels. For example, a name can be distinct nationally, or nationally when combined with information on date of birth. It can be distinct at the county level, or at the country level when combined with information on date of birth. Finally, in the British data, but not in the American data, a name can be unique at the parish level, or at the parish level when combined with information on date of birth. In this paper, I will be making use of three different levels of distinctness, which will be explained in further detail in the methodology section.

5 Methodology: Litmus Testing

The primary objective of this paper is to propose a method for evaluating the quality of data generated by any census linking algorithm, and to do so without relying on a ground truth dataset. The methodology introduced in the below exploits the systematic biases introduced by false matches to construct litmus tests. In this context, a variable can serve as a litmus test for data quality when the probability of a specific outcome for false positives differs significantly from the probability of the same outcome for true positives. By identifying and quantifying this divergence, it becomes possible to estimate the share of false positives in a sample.

The approach involves defining the expected range of outcomes for the litmus variables by constructing two benchmark samples:

- **Known False dataset:** These are deliberately mismatched dyads that do not match on key time-invariant characteristics: age, parish of birth, and family structure. This sample represents the upper bound of the litmus test.
- **Known True dataset:** These are constructed using strict matching criteria that ensure a high level of confidence in the validity of the matches. The criteria include exact matches on name, age, county of birth, and parish of birth, with the additional requirement of requiring a match on spouse.

This dual-sample framework establishes a benchmark for evaluating the accuracy of any census linking algorithm. By comparing the litmus variable outcomes for a given panel against the known true and false ranges, researchers can estimate the share of false positives without relying on costly and time-intensive ground truth datasets.

[8] Once the parameter space has been mapped with exact name matching approaches, it will be valuable to assess how distance linking methods fare in comparison.

5.1 Variable Selection

Some variables can act as a litmus test for false positives in the population, some cannot. To identify false positives in linked census data, we evaluate variables based on their ability to differentiate true matches from false matches. Effective litmus test variables exhibit the following property: the probability of a specific outcome for a false positive is significantly different from the probability of the same outcome for a true positive. In this section, I evaluate four potential litmus test variables — migration, occupation change, hisco score, and marital status — based on their discriminatory power.

5.1.1 Migration

Migration is a natural candidate for a litmus test. False matches are overwhelmingly likely to suggest a change in location. For a county c with population N_c, the probability that a false positive indicates migration is:

$$P(\text{migrated} \mid \text{False Positive}) = 1 - \frac{N_c}{N_{\text{total}}} \tag{1}$$

Since most counties represent a small fraction of the total population, $P(\text{migrated} \mid \text{FP})$ is very high.

In 19th-century England, there are 42 counties. As an illustrative scenario, consider the 1861 and 1881 census returns. Suppose we are trying to link an individual named "John Palmer," who resides in Rutland in 1861. We want to determine the probability that, in the case of a false positive match, the linked individual from 1861 resides in a different county. For example, if N_c as is Rutland, the smallest English county, nearly every false positive will suggest migration. Conversely, true matches reflect actual migration patterns, which are often much lower. For example, in 1861 Rutland had a population of $22,792$ out of $18,000,000$. The probability of a false positive indicating migration is therefore 99.87%.

Migration works as a strong litmus test because false positives are far more likely to suggest migration than true matches. In the results section I will for the most part be applying the litmus test to the entire sample, so the results here will be, effectively, an averaging over likelihoods across all counties.

Occupation Change in occupation is another strong litmus test variable. The probability of a false positive indicating a change in occupation is:

$$P(\text{changed occupation} \mid \text{False Positive}) = 1 - \frac{N_{\text{occ}}}{N_{\text{total}}} \tag{2}$$

Most occupations represent a small proportion of the population, meaning that a false positive match will almost always result in an observed "change of occupation. In contrast, true matches reflect actual occupational mobility, which varies by occupation but is generally much lower. For blacksmiths, who numbered 1331 in a population of $18,000,000$ in 1851, the probability of a false positive indicating a change of occupation is around 99%

5.2 Hisco Score

This is a scoring of the social status of any given occupation. Because most occupations are situated in the middle of the distribution of hisco scores, this will not be an ideal litmus test. However, I include it as a way to show how the impact of the inclusion of false positives will then cascade into our assessment of other questions of interest, such as social mobility.

5.2.1 Marital Status

Marital status serves as an example of a variable that does not work as a litmus test. The probability of being married is similar for both true matches and false positives:

$$P(\text{married} \mid \text{False Positive}) = P(\text{married} \mid \text{True Match}) = p_{\text{married}} \tag{3}$$

When a variable is highly prevalent, false positives are likely to match the general population distribution, which is similar to the true distribution for that variable. This lack of differentiation undermines its utility as a litmus test. For example, among individuals over 30 in 19th-century England, 70% were married. Therefore true and false matches are equally likely to show the individual as married, making marital status an uninformative variable for detecting false positives.

5.2.2 Conclusion

Migration and occupation changes are highly effective litmus test variables because false positives are far more likely to produce these outcomes than true matches. Marital status, in contrast, lacks discriminatory power due to its high prevalence in the population. These results highlight the importance of selecting variables where false positives significantly deviate from true matches, ensuring robust detection of errors in census linking.

5.3 The Range

Having established the relevance of migration and occupational change as litmus test variables, I now present the observed range for these variables using the constructed "known false" and "known true" samples. The graph below illustrates the outcomes for these variables across the two samples.

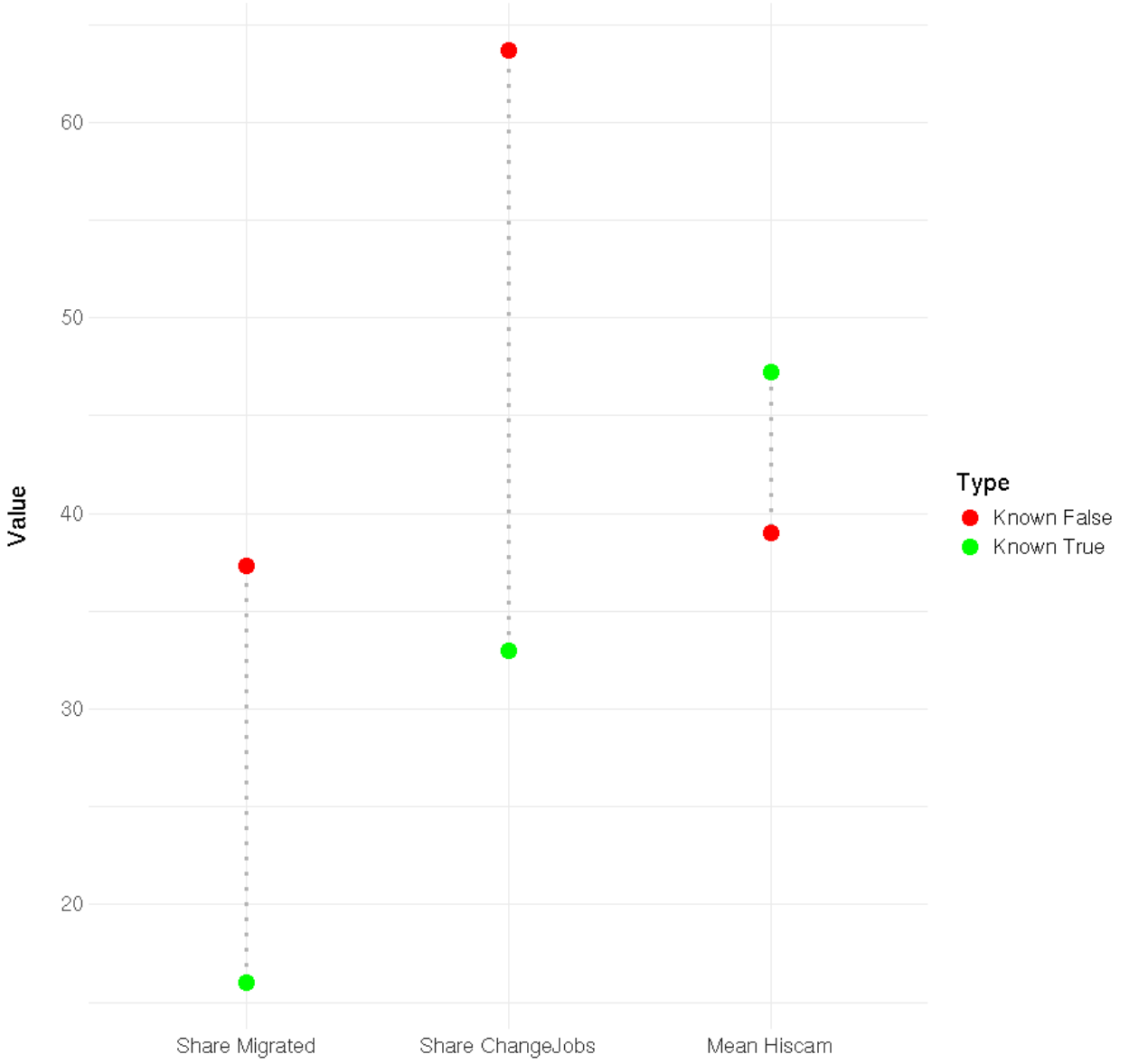


Figure 3: Range Between Benchmark Datasets
Note: Deanonymized data from the ICeM Project

The "known false" samples show up with very high rates for both migration and change in occupation. This is as expected, given the findings from (Helgertz et al., 2022) and discussion on litmus test variable selection. The "known true" samples, on the other hand, report substantially lower numbers.

At first glance, it might seem logical to rely solely on the "known true" sample for analysis. However, the "known true" sample is constructed using highly restrictive criteria, which necessarily result in high rates of false negatives, and therefore a small sample size. While the "known true" sample provides an essential benchmark for the expected rates of migration and occupational change in context with few false positives, it is not a good candidate for an ideal census linking

algorithm.

6 Results

With the litmus test approach now established, I apply it to three different questions. Firstly: is British historical data less exposed to false positives in census linking than American level data? Secondly, to what extent does linking sub-samples introduce additional false positives? Thirdly, how can insights gained from this approach assist in linking particular segments of the population?

6.1 American and British

To compare census linking outcomes at the British and the American level of data, I use British census data to construct a mock-up "American" level models. The main distinction between census linking with British data and American data is that British data includes information on both county and parish of birth, while the American data often records only state of birth.

In the remainder of the paper, the term "American level data" refers to British historical census data which has information on place of birth only at county level. This is taken as a rough equivalent to the historical American census data. It is also the only way to directly compare results from different census linking specifications - any exercise which does this must use the same dataset.

The question in this section is whether the additional information available in the British data is less exposed to false positives. I take a one at a time multi-model approach. The first thing to do is show the base specification, divided at different levels. Let us start by conceptualizing a simplistic deterministic linking model as a function. We can define the linking function as one which determines whether a record matches another based on the given parameters. Here, the county of birth is a fixed requirement, meaning every match must have the same birth county. The remaining parameters — birth date, name, uniqueness, and birth parish — are variable inputs that the function will consider.

$$Y = f_{\text{BirthCounty}}(\text{BirthDate}, \text{Name}, \text{Uniqueness}, \text{BirthParish})$$

Table 2, below, shows the size of the panel data sets that are constructed in this way. For most specifications, we can see that the the sample constructed with American level data is larger than the panel constructed with British level data. A detailed explanation as to why this is the case is beyond the scope of this paper, please see my PhD thesis. However, it is primarily because the American level data relies a great deal on the uniqueness criteria.

Table 2: 45 Deterministic Linking Models - British and American Level

Mode Name	Age	Uniqueness	Name	British	American
model 19	Age 1	1	Exact Name	3818	4811
model 22	Age 1	1	Clean Name	3597	4519
model 25	Age 1	1	Phonetic Name	3027	3870
model 20	Age 1	2	Exact Name	23111	40745
model 23	Age 1	2	Clean Name	24909	42841
model 26	Age 1	2	Phonetic Name	25819	50132
model 21	Age 1	3	Exact Name	29097	
model 24	Age 1	3	Clean Name	31904	
model 27	Age 1	3	Phonetic Name	34969	
model 10	Age 2	1	Exact Name	4278	5559
model 13	Age 2	1	Clean Name	4013	5198
model 16	Age 2	1	Phonetic Name	3378	4482
model 11	Age 2	2	Exact Name	26182	41240
model 14	Age 2	2	Clean Name	28130	42982
model 17	Age 2	2	Phonetic Name	29370	49487
model 12	Age 2	3	Exact Name	32330	
model 15	Age 2	3	Clean Name	35159	
model 18	Age 2	3	Phonetic Name	38553	
model 1	Age 3	1	Exact Name	4480	5991
model 4	Age 3	1	Clean Name	4206	5590
model 7	Age 3	1	Phonetic Name	3564	4874
model 2	Age 3	2	Exact Name	27661	39395
model 5	Age 3	2	Clean Name	29672	40816
model 8	Age 3	2	Phonetic Name	31246	45817
model 3	Age 3	3	Exact Name	33657	
model 6	Age 3	3	Clean Name	36405	
model 9	Age 3	3	Phonetic Name	40066	

[1] *This table shows the size of the sample produced by each of the 27 linking models, for data at both the ■British■ and ■American■ level. Note that, since American data makes no use of parish level place of birth, some linking specifications are not possible for it.*

Now that we have a clear understanding of the 45 different models, I run them through the litmus test. In the graph below, the three panels are for each of the three variables being employed in the litmus testing. Panel one shows the mean hisco (social status) score for occupation. Panel two shows the share which change occupation. Panel 3 shows the share of individuals in the panel dataset who migrated, ie changed location.

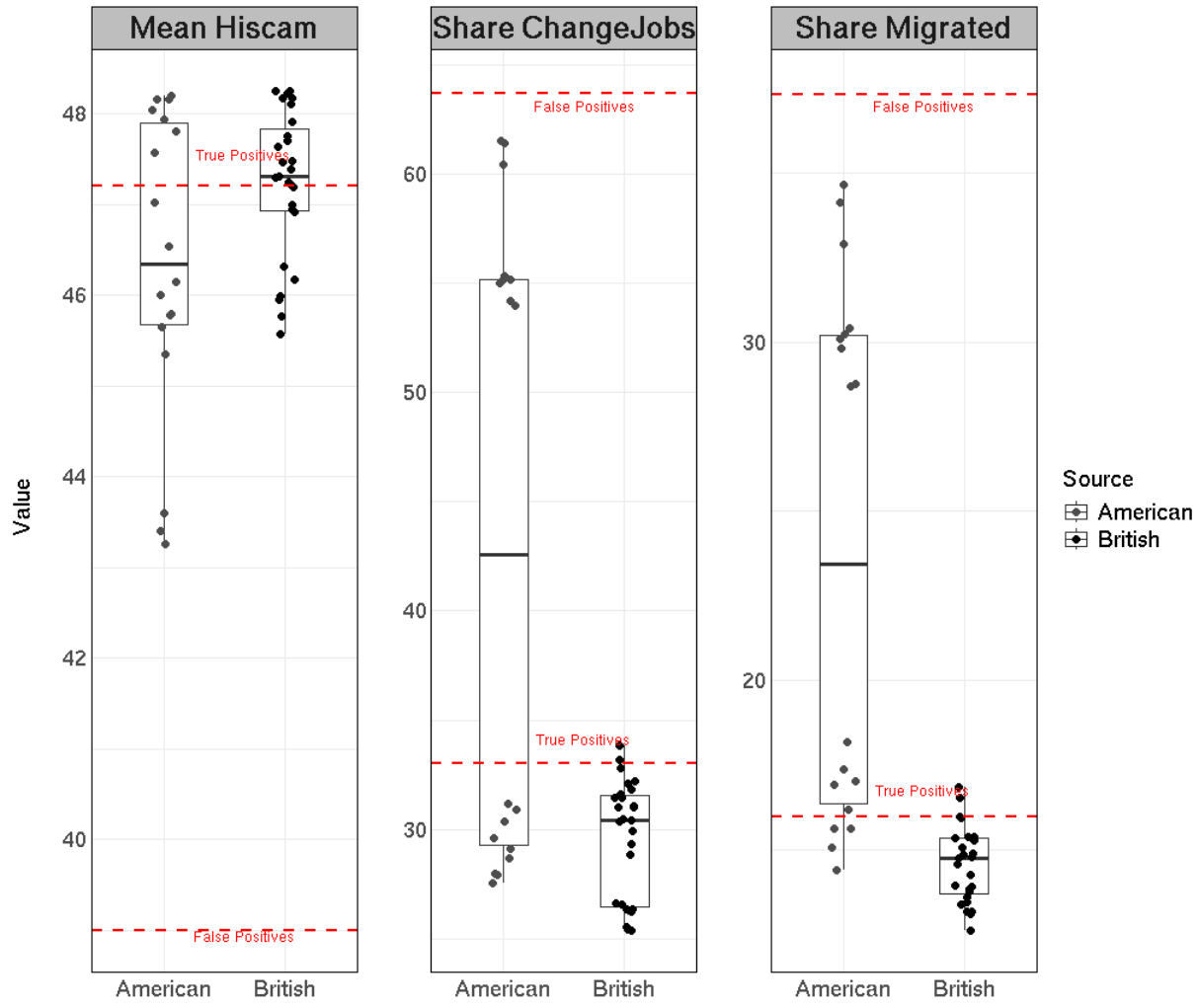


Figure 4: British vs American Level Data
Note: Deanonymized data from the ICeM Project

For both migration and change in occupation we see that the samples constructed with American level data span a very wide range. Half of the samples cluster rather near to the outcomes reported in the "known false" sample. British level data, on the other hand, does not span so wide a range, and all samples are clustered near the "known true" outcomes.

I conclude that British census data is less exposed to false positives irrespective of which linking algorithm is chosen. I attribute this primarily to the additional information on place of birth, which provides considerable power in distinguishing potential matches.

6.2 Linking Sub-samples

The second research question in this paper is to quantify the extent to which linking sub-samples introduces false positives. This involves comparing two approaches: (1) linking the entire census record and subsequently filtering down to bootmakers, and (2) directly linking bootmakers across census records. The comparison aims to determine whether sub-sample linking disproportionately increases the share of false positives.

In the previous section, I simply linked bootmakers to the full census in the next period. In this section I compare the outcomes of these two approaches, and assess the extent to which linking sub-samples introduces additional false positives.

This analysis is conducted for both the British and American level of data. The results from the previous section suggest that linking sub-samples is particularly prone to introducing false positives in American data, where the lack of granular birthplace informationsignificantly limits the distinguishing power of matching variables. Therefore, I hypothesize that the issue will be more pronounced in the American context.

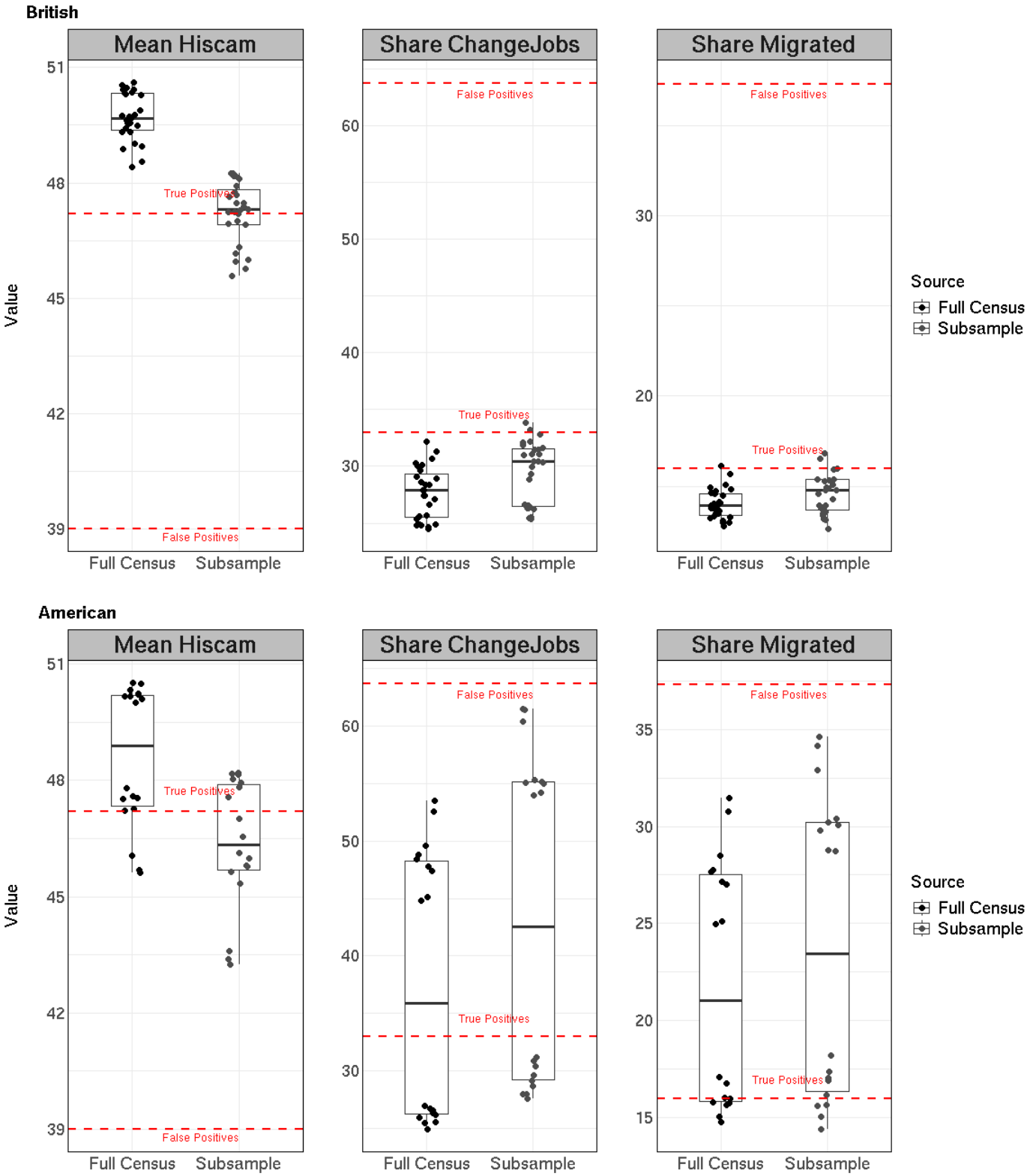


Figure 5: Linking Subsamples vs the Full Census
Note: Deanonymized data from the ICeM Project

In the graph below you see the litmus test conducted on full census versus sample. The top panel is the British Data. The bottom panel is the American data.

When we select a sub-sample to link forward we are vastly increasing the likelihood of false positives and the reliance on the uniqueness of names. Data at the American level, which has less distinguishing information to hand, is much more exposed to this problem. For the British level data, linking sub-samples forward is a bit risky but not entirely outrageous. However, for American level data, linking sub-samples should not be considered unless there is no other option.

6.3 Linking Immigrants

Linking immigrants to the United Kingdom is a very particular case of census linking. In contrast to the approach established for native born British citizens, no information is available on parish of birth. Additionally, there is no information available on county of birth. Finally, while individuals do report their country of birth, there is a lack of consistency in doing so. I find some individuals who report as having immigrated from Russia in one census return, and from Belarus in another. As a result, country of birth cannot be considered a distinctive and time-invariant attribute for these immigrants.

I collect the full set of immigrants from the set of origin countries Belarus, Poland, and Russia, in time period t and t+1. Matching is done on the basis of: full name, age (within a window of 2 years), and membership in the set of immigrants from the Pale of Settlement countries. However, given the lack of information on place of birth, matching is challenging. I therefore include information on family [9], and leverage spouse name and date of birth. This increases the sample size constructed and ensures a lower share of false positive matches. I run a set of extremely conservative matches in order to run the litmus testing and validate the quality of the census linked data.

While the above is the most straightforward way to leverage family information, there are other options. Immigrants often married English citizens, and had children born in England. It is relatively more straightforward to follow these individuals over time, and find the Russian immigrants in the context of their familial links.

Finally, I drop place of birth as a requirement in the matching. This is to account for the fact that immigrants often either change the country they name as their place of birth, or simply register it as "unknown" in census returns. In this case, identifying these individuals in the census has grid search the entire census. In this case, the Jewish names are relatively rare in the English population, and matching on name alone does not yield too many dyads. These can then be adequately distinguished by leveraging information on family and date of birth. Each of these approaches can be stress tested and validated with the litmus testing method for share of false positives in the resulting panel data.

7 Discussion

The various challenges of census linking have been much discussed. Several solutions to census linking challenges have been proposed. Given that false positives are the most pressing issue, the most straightforward approach is to require more narrow - conservative - matching criteria (Abramitzky et al., 2020b). For example, matching pairs can be required to have exactly the same reported year of birth. In the standard ABE process matched pairs are allowed to differ by up to 2 years in reported year of birth. Names can be required to be unique within a 5-year band. In the standard ABE approach observations are dropped if there are multiple potential matches with the same name and year of birth. This modification will also drop those observations that have

[9] see Appendix C

multiple matches with the same name and implied birth year within 5-year interval. Since reported age is an imperfect measure of the year of birth, such an approach can reduce the Type 1 error. While Bailey et al. (2020) and others suggest that this approach may reduce the incidence of false links, there is little data to support this. It is also possible that moving to this more restrictive approach could increase the number of false positives in the sample, by putting undue weight on the uniqueness criteria. For instance, if there is only one Abigail Blizzard in both datasets and the age matches, this person is considered a match. There is ample opportunity, without the availability of additional information, for this match to be a false positive. It is also worth mentioning that restricting the universe of possible matches in this way necessarily creates Type 2 errors, with many potential matches being disqualified. Finally, there is a proposal to make use of the intersection of results from multiple model specifications. This was trialed by Collins and Wanamaker (2022). They conducted matching via approaches proposed by Ferrie (1996) and Abramitzky et al. (2014). Only links that were identified by both algorithms were counted as successful, which they suggested will lead to more conservative results, and a decrease in Type 1 error.

The issue with the representativeness of the sample has less often been explicitly addressed, although researchers are well aware of the problem. A few solutions have been proposed. The first is to make use of inverse-propensity weighting (Collins and Wanamaker, 2022; Bailey et al., 2020). This approach relies on the assumption that if a sample is representative on observables it will also be representative on unobservables. The second is simply to cut the data, and to consider rural and urban sub-samples separately. ABE2010 show that in their setting urban citizens are more likely to be successfully linked, and provide a separate analysis for urban and rural areas.(Collins and Wanamaker, 2022) also present a specification separately for rural areas.

I propose that it would be helpful to have a way to adjudicate on the quality of any given approach to census linking. I argue that false positives should be seen as the most important issue. It is possible to consider the census linking problem as one of constrained optimization. The goal is to construct as large of a sample as possible, subject to the constraint of not having an excessive share of false positives introduced into the sample. In this paper, I have proposed an approach to litmus test the constraint. With an estimate of the constraint now available, it is possible to expand the sample size in various ways.

8 Contribution

This paper provides a falsification framework that enables researchers to make choices between census linking specifications. It does so by providing a litmus testing approach, which allows the quality of the data constructed by a census linking algorithm to be evaluated. This then means that census linking can be approaches as an optimization problem, of maximizing sample size subject to a constraint on data quality.

The analysis in this paper, which made use of the litmus test approach, demonstrates that British census data, with its granular birthplace information, is less prone to false positives than American data. It also shows that linking the full census is not only best practice, but quite important for American level data. Finally, it shows how this approach can be used to link segments of the population which are particularly challenging.

References

ABE. Leaving the Enclave. *Paper Knowledge . Toward a Media History of Documents*, 2014.

R. Abramitzky, L. P. Boustan, and K. Eriksson. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102(5):1832–1856, 2012.

R. Abramitzky, L. P. Boustan, and K. Eriksson. A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(3):467–506, 2014.

R. Abramitzky, L. P. Boustan, and D. Connor. Leaving the enclave: Historical evidence on immigrant mobility from the industrial removal office. Technical report, National Bureau of Economic Research, 2020a.

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

M. J. Bailey, C. Cole, M. Henderson, and C. Massey. How well do automated linking methods perform? lessons from us historical data. *Journal of Economic Literature*, 58(4):997–1044, 2020.

L. P. Boustan. *Competition in the promised land: Black migrants in northern cities and labor markets*, volume 39. Princeton University Press, 2016.

L. P. Boustan, P. V. Fishback, and S. Kantor. The effect of internal migration on local labor markets: American cities during the great depression. *Journal of Labor Economics*, 28(4):719–746, 2010.

W. J. Collins and M. H. Wanamaker. African american intergenerational economic mobility since 1880. *American Economic Journal: Applied Economics*, 14(3):84–117, 2022.

D. S. Connor. The cream of the crop? geography, networks, and irish migrant selection in the age of mass migration. *The Journal of Economic History*, 79(1):139–175, 2019.

D. S. Connor and M. Storper. The changing geography of social mobility in the united states. *Proceedings of the National Academy of Sciences*, 117(48):30309–30317, 2020.

R. Dahis, E. Nix, and N. Qian. Choosing racial identity in the united states, 1880-1940. Technical report, National Bureau of Economic Research, 2019.

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

J. Helgertz, J. Price, J. Wellington, K. J. Thompson, S. Ruggles, and C. A. Fitch. A new strategy for linking us historical censuses: A case study for the ipums multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55(1):12–29, 2022.

E. Higgs, C. Jones, K. Schürer, and A. Wilkinson. The integrated census microdata (i-cem) guide. 2013.

S. Pérez. The (south) american dream: Mobility and economic outcomes of first-and second-generation immigrants in nineteenth-century argentina. *The Journal of Economic History*, 77(4): 971–1006, 2017.

9 Appendices

9.1 Appendix A: Missing Data

Here I provide a table of missing data in 1861 and 1881. Please see Table 3. This includes information on missing data for age, first name, last name, county of birth, and parish of birth. It is impossible to tell whether family information is missing: for the moment we must simply trust the census on this score.

Table 3: Census Data Overview from 1851 to 1911

	1851	**1861**	**1881**	**1891**	**1901**	**1911**
First Name Missing	7147	10120	0	19343	14982	25558
First Name Dots	0	69562	1810	1	0	0
First Name Short	112874	187568	45291	239412	258377	229475
First Name StartsQ	3369	4	0	18847	1199	0
First Name StartsSymbol	0	0	5911	0	0	0
Last Name Missing	51115	5358	15	2535	23152	15963
Last Name Dots	0	62486	0	2	0	0
Last Name Short	63611	60345	15	51805	24279	40049
Last Name StartsQ	5786	15	0	30153	1221	0
Last Name StartsSymbol	0	0	16715	0	0	0
Age Missing	54768	50618	12797	27128	10223	71577
County of Birth Missing	24987	26686	45793	25906	41252	11946
Parish of Birth Missing	555502	976374	460121	599922	438968	346154
Address Missing	847590	686463	804112	636740	521214	66760
Total	15429794	17088934	22954320	25696597	28751024	31363694

Absolute numbers, from ICeM data, etc

Table 4: Census Data Overview from 1851 to 1911

	1851	**1861**	**1881**	**1891**	**1901**	**1911**
First Name Missing (%)	0.046	0.059	0.00	0.075	0.052	0.081
First Name Dots (%)	0.00	0.41	0.008	0.00	0.00	0.00
First Name Short (%)	0.73	1.098	0.198	0.93	0.90	0.73
First Name Start'Q' (%)	0.02	0.00	0.00	0.073	0.004	0.00
First Name StartSymbol (%)	0.00	0.00	0.026	0.000	0.00	0.00
Last Name Missing (%)	0.33	0.031	0.00	0.01	0.08	0.051
Last Name Dots (%)	0.00	0.37	0.00	0.00	0.00	0.00
Last Name Short (%)	0.41	0.35	0.00	0.20	0.084	0.123
Last Name Start'Q' (%)	0.037	0.00	0.00	0.12	0.00	0.00
Last Name StartSymbol (%)	0.00	0.00	0.073	0.00	0.00	0.00
Age Missing (%)	0.35	0.30	0.056	0.11	0.036	0.23
County of Birth Missing (%)	0.16	0.16	0.20	0.10	0.14	0.04
Parish of Birth Missing (%)	3.60	5.7	2.00	2.33	1.53	1.10
Address Missing (%)	5.49	4.02	3.50	2.48	1.81	0.21
Total Rows	100.0	100.0	100.0	100.0	100.0	100.0

All values are percentages of total entries, representing the proportion of each category.

Name Names are classified as "missing" in the census data for any of the following reasons. Firstly, the name is missing. This can occur if the name is completely missing, or given as "...", or "-". This varies between census years in the records. Secondly, the name can start with a question mark or "..." [10]. Thirdly, if the first or last name consists of two characters or fewer [11]

Note that in 1851, the missing names consist of dashes, names with fewer than two characters, and question marks. In 1861 the missing names are made up of: names that are simply left empty, names starting with "..." or only containing "...", and names made up of two characters or fewer. In 1881, most missing names are represented by '@@@'. From 1891 to 1911 there are no new characters representing missing names.

Age As shown in the table, information on age is rarely missing. In the early years of 1851 and 1861 the problem seems a bit more pronounced, with around 50 000 observations missing information on age. From 1881 onwards it is lower. Missing information on age is not likely to be a major issue in census linking. Misreporting, of course, is another question.

County of Birth County of birth is usually not missing, with around 25,000 missing entries in most years. As with age, missing information is not likely to be an impediment to census linking. However, as with age, this tells us nothing about misreporting.

Parish of Birth Parish of birth is very often missing. Around half a million observations are routinely missing information of parish of birth, and this increases to nearly a million in 1861.

[10] the question mark start seems to mostly be limited to the 1851 census

[11] At a later point it will be possible to leverage this information in the matching, however, for the moment names with this minimal information will be counted as missing, and these observations will be dropped

Misreporting may also be high.

Address I am not currently making use of this data, but address is quite often provided. It may be useful in future.

As Table 3 shows, the parish of birth information is quite often missing. Several teams are currently working on geo-coding address data, or deriving information from census enumerator book page number, and this will help. One thing this makes clear is that deterministic matching, as I am doing here, will result in many matches being lost. A probabilistic model is therefore necessary. The work in this paper can be used to inform decisions made in the probabilistic model. Of course I could simply take a machine learning approach, but in this case there are not all that many features involved, and it is nice to have selection of weights clear and understood rather than being inside the black box.

9.2 Appendix B: Finding Family

This is a two stage process. Firstly, I group family members together in blocks. Secondly, I add data on family members to each individual record in the full count ICeM census records. It is worth noting that the majority, but far from all, of the English bootmakers are living in a family unit.

To group family members together I make use of the following two pieces of information. Firstly, the variable 'parid' in the ICeM census data is unique within counties. Each parish has its own 'parid' number. Additionally, the variable 'h', is a unique identifier of households within the parish. I can therefore collect family groups by 'h' and 'parid' within counties. Note that I collect the full family for each individual in the database. A family of four will take up 16 lines in the Family database.

Once the large family database has been constructed, I reshape the data. I construct eight new columns: Wife name, Wife age, Head of Household name, Head of Household age, Son name (*youngest son), Son age, Daughter name (*youngest daughter), Daughter age. This additional data on family members is now available for each individual observation in all of the cross-sectional ICeM census data. It can therefore be integrated into the record linking process.

9.3 Appendix C: Representativeness

A comprehensive discussion of representativeness is beyond the scope of this paper. However, the table below serves to show that, while all census linking models introduce bias, they do so in much the same way. That is, choices made on linking specification, at least within this set of options, does not differentially impact on bias.

Table 5: Bias - British vs American Level, Age Leaves

Model Name	Name	Age	Uniq	Sharel Married		Share Urban		Share Female		Mean Age		Mean Sname Length	
				Britain	USA	Britain	USA	Britain	USA	Britain	USA	Britain	USA
Population				0.55		0.52		0.18		35.8		6.18	
Population				0.55		0.52		0.18		35.8		6.18	
model 22	Clean	1	1.00	0.59	0.58	0.46	0.48	0.17	0.18	31.10	30.81	6.46	6.47
model 13	Clean	2	1.00	0.58	0.57	0.46	0.48	0.17	0.19	31.06	30.72	6.46	6.45
model 4	Clean	3	1.00	0.58	0.57	0.46	0.48	0.17	0.19	31.13	30.73	6.46	6.46
model 23	Clean	1	2.00	0.59	0.56	0.46	0.51	0.13	0.16	31.50	31.79	6.23	6.17
model 14	Clean	2	2.00	0.59	0.57	0.47	0.51	0.13	0.17	31.55	32.19	6.22	6.19
model 5	Clean	3	2.00	0.58	0.57	0.48	0.51	0.14	0.17	31.60	32.58	6.21	6.21
model 24	Clean	1	3.00	0.57		0.47		0.12		30.83		6.17	
model 15	Clean	2	3.00	0.57		0.47		0.13		30.97		6.17	
model 6	Clean	3	3.00	0.57		0.48		0.13		31.13		6.18	
model 19	Exact	1	1.00	0.60	0.59	0.45	0.47	0.16	0.17	31.34	31.12	6.42	6.44
model 10	Exact	2	1.00	0.59	0.58	0.45	0.47	0.17	0.18	31.30	31.01	6.42	6.42
model 1	Exact	3	1.00	0.59	0.57	0.45	0.48	0.17	0.18	31.37	31.01	6.43	6.43
model 20	Exact	1	2.00	0.60	0.57	0.46	0.50	0.12	0.15	31.79	31.93	6.22	6.16
model 11	Exact	2	2.00	0.59	0.57	0.47	0.50	0.13	0.16	31.81	32.28	6.21	6.17
model 2	Exact	3	2.00	0.59	0.57	0.47	0.50	0.13	0.17	31.85	32.63	6.21	6.19
model 21	Exact	1	3.00	0.58		0.46		0.12		31.18		6.17	
model 12	Exact	2	3.00	0.58		0.47		0.12		31.27		6.17	
model 3	Exact	3	3.00	0.58		0.47		0.13		31.40		6.17	
model 25	Phonetic	1	1.00	0.58	0.57	0.43	0.46	0.16	0.17	30.57	30.39	6.73	6.72
model 16	Phonetic	2	1.00	0.57	0.56	0.44	0.46	0.16	0.18	30.58	30.33	6.74	6.72
model 7	Phonetic	3	1.00	0.58	0.56	0.44	0.47	0.17	0.18	30.72	30.36	6.74	6.72
model 26	Phonetic	1	2.00	0.59	0.56	0.47	0.51	0.13	0.17	31.83	32.14	6.32	6.17
model 17	Phonetic	2	2.00	0.59	0.56	0.48	0.52	0.14	0.18	31.87	32.56	6.31	6.20
model 8	Phonetic	3	2.00	0.59	0.57	0.49	0.52	0.15	0.18	31.88	33.01	6.30	6.25
model 27	Phonetic	1	3.00	0.57		0.47		0.13		31.04		6.21	
model 18	Phonetic	2	3.00	0.57		0.48		0.14		31.19		6.21	
model 9	Phonetic	3	3.00	0.57		0.48		0.14		31.34		6.22	

Note: The table shows population statistics in the top row of the table, and then sample statistics for all 45 models, allowing for comparison between the ■British■ and ■American■ level of data. Data arranged by ■Age■ leaves.