# **Appendix**

## **A England And Wales Decennial Census**

Nineteenth- and early-twentieth-century censuses are an invaluable source of quantitative information into the lives of people living in Victorian and Edwardian England, and an alternative primary resource for the study of occupational mobility in the past. The act of census taking began in 1801, although it was not until 1841 that names and details of individuals were collected, and information on birth places and occupations remained limited until the 1851 census (Higgs, 1989). An awareness of the procedures involved in census taking from 1851 onwards may be required to understand the limits and reliability of the information obtained from the census returns.

A simple explanation of how the census was taken is as follows. The country was first divided into enumeration districts, each containing roughly 200 households and one enumerator. The enumerators delivered a “household schedule” and written instructions to each household on the night of the census – normally in March or April to avoid the distortions caused by seasonal movements in the summer by some sections of the population – which had to be filled out and returned by the household head. On collection day, the enumerators would collect and check the schedules, and help the household heads to complete the schedule if they could not do so. Up until 1911, the enumerators would then standardize and copy the information onto the Census Enumerator’s Book (CEB). Both the schedules and the books were submitted for checking to the district registrars before they were sent to the Census Office, where they were checked again by the clerks. The household returns were then destroyed. For the 1911 census, the original schedules were used for the tabulation of statistics, so there was no standardization of the raw data by the enumerators (Higgs, 2005).

One concern that scholars may have with the use of nineteenth-century censuses for historical research is the quality of census enumerators. Enumerators were hired on a temporary basis by local registrars, and anyone can be hired as long as they satisfied the basic requirements (Higgs et al., 2013).[[1]](#footnote-2) In urban areas, the enumerators were often local government officers and schoolteachers, but in the countryside the registrars may have had to depend on the farmers and their kin (Arkell, 1994). Unsurprisingly, there is a lot of variation in the abilities of enumerators – they differed in their ability to read and write, and in their ability to comprehend lengthy instructions given to them by the registrars (Tillott, 1968). Fortunately, the enumerators generally appear to be of a satisfactory standard. In an area sampled by Tillott (1972), only six of the ninety enumerators showed evidence of unsuitability for their task. This may be especially true for the towns, where enumerators were more likely to be men of clerkly habits employed in occupations that require a certain degree of literacy.

Another source of inaccuracies may come from the householders who inadvertently give out the wrong information, mostly due to ignorance or ambiguity in the instructions. Insofar as people’s intentions to answer the questions truthfully were concerned, there is little evidence to suggest that this is a huge issue (Tillott, 1972). With information on name, sex, occupation, and birthplace, there is generally little room for falsification, though inconsistencies may occur as a result of spelling variants with names, ambiguous definitions and instructions given to the recording of occupations, and geographical ignorance (Tillott, 1972; Higgs, 2005). In cases where the householder was illiterate, the enumerators were responsible for filling the schedules. The proportion of schedules filled out by enumerators varied widely across regions – for example, in the six enumeration districts of Great Missenden in Buckinghamshire, this proportion ranged from 5.3 to 64.7 per cent (Higgs et al., 2013). Thus, there may be cases where the wrong information was recorded due to miscommunications between the enumerator and the household. With the introduction of compulsory education after 1870, one would expect the ability to read and fill the schedules improved for both the householder and the enumerator.

## **B ABE Matching Algorithm**

The ABE algorithm matches individuals over time by names (string distances or phonetic names), places of birth (in this case parish), and inferred birth year from age (Abramitzky et al., 2020).[[2]](#footnote-3) Matching via string distances is the preferred method in this paper. The procedure for both string distances and phonetic names versions are as follows.

**Using Jaro-Winkler String Distances – Preferred Linkage Method**

1. The raw strings for first and last names in dataset A (i.e., all men in 1851) and dataset B (i.e., all men in 1881) are cleaned, which removes non-alphabetic characters and accounts for shortened names such as “Ben” for Benjamin and spelling variants.
2. The data is then split into smaller blocks by initial letters of first and last names, age, and birthplace. The string distances of all names within plus and minus 5 years of reported age between dataset A and B are calculated, and only pairs of individuals in A and B with string distances of less than 0.1 in both first and last names are kept.
3. There are three potential outcomes in the matching procedure:
	1. No potential match could be found for a given individual in dataset A, so this observation is dropped from the data.
	2. There may be only one potential match for an individual in dataset A, and the corresponding match in dataset B has no other potential matches in dataset A. This is determined to be a successful match.
	3. In cases where they are more than one potential match by name in dataset B, the individual (let us call him B1) closest in inferred birth year to the observation in dataset A is matched only if the second closest observation in B is more than 2 years apart in reported age to B1.
4. To minimize Type I errors, this paper adopts the conservative approach where matches are also required to be unique within a 5-year band (plus or minus 2 years in age) and to differ in reported age by no more than 2 years.

**Using NYSIIS Phonetic Names – Alternative Linkage Strategy for Additional Results**

1. The raw strings for first and last names are cleaned.
2. Names are then converted into their phonetic names using the New York State Identification and Intelligence System (NYSIIS) Code.
3. The sample from the initial year is restricted to those who are unique by first and last name, age, and parish of birth, since it is impossible to distinguish between which non-unique individuals should be linked to the potential match.
4. Following from this, matches can be identified based on their vital information through an iterative procedure:
	1. If a unique match – same name, birth year, and birth parish – is found, the individual is “matched.”
	2. If there are multiple matches for the same birth year, the observation is discarded.
	3. If no matches are found for the same birth year, the process is expanded to matching within a one-year band (older or younger), and then within a two-year band around the inferred birth year. Again, only unique matches are accepted.
5. To reduce the likelihood of false positives, matches are required to have unique names within a five-year band (plus or minus two years) around the birth year.

## **C Estimating False Positive Rate**

When estimating the rates of intergenerational mobility from linked census data, it has now become commonplace to emphasize the importance of reducing Type I errors (false positives) since a sample with a high share of false positives may attenuate the IGE estimated and therefore overstate the extent of social mobility.

The standard procedure for calculating the rate of false positives associated with a census linking process in the literature is to benchmark the linked dataset with a high-quality reference dataset. For instance, Bailey et al. (2020)’s extensive review of some of the most widely known census linkage algorithms used three different reference datasets, two of which are hand-linked samples and one is a “ground truth” sample with some noise added in to mimic errors in historical data. Abramitzky et al. (2020) also reviewed their own linkage methods, where they compared their linkages to hand-linked family tree data. Such hand-linked samples are rare to find, and one can only make assumptions about their reliability.

Though there are reference samples that are not hand-linked, those are equally difficult to obtain. Anbinder et al. (2021)’s survey on matching Irish immigrants in the US used the Emigrant Industrial Savings Bank records, which contain information about customers that are much more detailed than those found in a census. Massey (2017) circumvents the issue of finding high-quality historical datasets by using modern data instead, where she can guarantee the reliability of her reference sample through unique identifiers (such as Social Security Number). She then conducts record linkage on the same datasets and compares the linkage results with the true links. These exercises are incredibly useful for showing us the potential pitfalls of automated census linkage, but they cannot be replicated in a different historical context (such as for the purpose of estimating false positives in this paper).

In the absence of a high-quality reference dataset, I have devised a method of checking the rate of Type I errors associated with the census linkage process using double-linked samples instead.[[3]](#footnote-4) The procedure for estimating the rate of false positives is as follows. Taking the 1881-1911 sample as an example, I first select sons whose relationship status as reported in the census is “son” in both 1881 and 1891, indicating that they are living with their families in both years. I then check if the fathers they are living with in 1891 are the same individuals that I identified when I linked their fathers from the 1881 to the 1891 census. This is a valid test because fathers and sons are linked across census years independently. I can then calculate the percentage (*γ*) of sons whose actual fathers they are living with in 1861 are different to the fathers that I linked.

 It is important to note here that this is only an upper-bound estimate of the false positive rate (*α*) associated with the linking algorithm. This is because the linkage process entails running the algorithm twice – once for matching sons from 1881 to 1891, and once for matching fathers from 1881 to 1891. Thus, γ is an outcome of these four scenarios:

1. $P\left(E\_{F}=1\right)=α\*α$, where ES = 1 denotes a Type I error in the linkage of sons and EF = 1 denotes a type I error in the linkage of fathers.
2. $P\left(E\_{F}=0\right)=α\*(1-α)$, meaning fathers are correctly matched between 1881 and 1891 but sons are false matches.
3. $P\left(E\_{F}=1\right)=(1-α)\*α$, meaning that sons are matched correctly between 1881 and 1891 but fathers are false matches.
4. An unknown percentage *β* that represents the share of false positives eliminated by the requirement for sons to have a match in every census year within the 30-year interval. In other words, sons who can be falsely matched between 1881 and 1911 but not between 1881 and 1891 or between 1881 and 1901.

Combining these scenarios produce the following equation:

|  |  |
| --- | --- |
| $$γ = 2α-α^{2}-β$$ | (C1) |

Solving the quadratic would reveal the true rate of false positive rate associated with the linkage algorithm:

|  |  |
| --- | --- |
| $$2α-α^{2}-β-γ = 0$$ | (C2) |

However, since we do not know the exact value of *β*, we can only derive a lower-bound estimate of the false positive rate by assuming *β* = 0.

Table C1: False Positive Rate of Linkage

|  |  |  |
| --- | --- | --- |
|  | 1851-1881 | 1881-1911 |
|  | Numbers | Percentage | Numbers | Percentage |
| Correct Match | 37,965 | 82.84 | 99,641 | 83.37 |
| Wrong Match | 7,865 | 17.16 | 19,869 | 16.63 |
| UB False Positive Rate |  | 17.16 |  | 16.63 |
| LB False Positive Rate |  | 8.98 |  | 8.69 |
|  |  |  |  |  |
| Total | 45,830 |  | 119,510 |  |

*Notes*: “UB” = Upper Bound; “LB” = Lower Bound.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Table C1 shows the upper and lower bound estimates for the rate of false positives. For both 1851-1881 and 1881-1911, the false positive rate lies between 8 and 17 per cent. This compares quite favorably to the performance of various prominent linkage algorithms when linking US censuses. For instance, Bailey et al. (2020) found that the most conservative version of ABE-NYSIIS produces a false positive rate of 17 to 23 per cent; Ferrie (1996), when using exact names, produces a false positive rate of 20 to 23 per cent; Feigenbaum (2018) produces a false positive rate of 16 to 29 per cent. Only the Expectation-Maximisation (EM) algorithm constructed by Abramitzky, Mill, and Pérez (2020) performs to a similar or better standard (false positive rate of 10 to 15 per cent). Evidently, the availability of more precise birthplace information makes a huge difference to how well automated census linking performs.

 We can also check the validity of this test for false positives by running this process on a linked sample with less conservative matches. For example, the less conservative version of Abramitzky et al. (2019)’s linkage algorithm only requires matches to be unique by their exact birth years, rather than within a five-year band. Table C2 shows the rates of false positives for this version of the matching algorithm. As the results show, using a less conservative matching algorithm does produce a higher match rate, but also higher incidence of false positives. Nonetheless, because the restriction on similarity of string distances have not been relaxed, we should not expect the false positive rate to be significantly higher, and that is the case here.

Table C2: False Positive Rate of Linkage, Less Conservative Matches

|  |  |  |
| --- | --- | --- |
|  | 1851-1881 | 1881-1911 |
|  | Numbers | Percentage | Numbers | Percentage |
| Correct Match | 49,300 | 81.10 | 126,860 | 81.77 |
| Wrong Match | 11,488 | 18.90 | 28,277 | 18.23 |
| UB False Positive Rate |  | 18.90 |  | 18.23 |
| LB False Positive Rate |  | 9.94 |  | 9.57 |
|  |  |  |  |  |
| Total | 60,788 |  | 155,137 |  |

*Notes*: “UB” = Upper Bound; “LB”.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **D Inverse Probability Weight**

Table D1 shows the representativeness of the weighted sample in comparison to the unweighted sample. The weighted sample is more representative in almost all variables except in over-representing “son” in relationship status and Yorkshire in birth and residence counties.

To address the issue of non-representative sample, I ran probit regressions of linkage outcomes (a dummy variable with value of 1 if the observation has been successfully linked) for sons on first name length, last name length, combined name length, first name commonness, last name commonness, age and its quadratic term, total male population in the parish and county of residence in the final census year, and occupational sector defined by the HISCO major groups (0 to 9). Name commonness is defined as the share of people aged 5 to 15 with the same name living in the same parish in 1851 for the 1851-1881 sample, 1861 for the 1861-1891 sample, and 1881 for the 1881-1911 sample. I then assign inverse probability weights based on the following equation:

|  |  |
| --- | --- |
| $$Weight=\frac{1-P\_{i}(L\_{i}=1|X\_{i})}{P\_{i}(L\_{i}=1|X\_{i})∙q(1-q)}$$ | (D1) |

Where $P\_{i}(L\_{i}=1|X\_{i})$ denotes the probability of being linked, and *q* is the share of people linked.

Table D1: Summary Statistics of Linkage Results (Weighted vs Unweighted), 1851-1911

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1851-1881 | 1861-1891 | 1881-1911 |
|  | Population | Unweighted | Weighted | Population | Unweighted | Weighted | Population | Unweighted | Weighted |
| **Characteristics (Son) in 1881/1891/1911** |
| Final Age (mean) | 39.68 | 39.12 | 39.58 | 39.65 | 39.53 | 39.54 | 39.67 | 39.40 | 39.60 |
| HISCAM (mean) | 54.38 | 53.95 | 54.81 | 54.42 | 54.40 | 55.02 | 55.29 | 55.43 | 55.74 |
| CCC (mean) | 53.13 | 53.51 | 54.43 | 52.72 | 53.66 | 54.18 | 53.39 | 54.41 | 54.67 |
| First name length (mean) | 6.27 | 6.30 | 6.28 | 7.80 | 7.84 | 7.79 | 7.88 | 7.98 | 7.86 |
| Surname length (mean) | 6.33 | 6.40 | 6.35 | 8.31 | 8.41 | 8.32 | 8.34 | 8.43 | 8.36 |
| Kids (mean) | 2.85 | 3.03 | 2.84 | 2.46 | 2.64 | 2.45 | 2.00 | 1.98 | 1.87 |
| Servants (mean) | 0.22 | 0.22 | 0.22 | 0.19 | 0.21 | 0.20 | 0.13 | 0.13 | 0.12 |
| **Characteristics (Father) in 1851/1861/1881** |
| Initial Age (mean) | 40.81 | 41.42 | 41.64 | 40.96 | 42.05 | 42.07 | 40.82 | 41.26 | 41.28 |
| HISCAM (mean) | 53.35 | 52.07 | 52.77 | 53.52 | 52.51 | 53.11 | 54.31 | 53.35 | 53.82 |
| CCC (mean) | 53.38 | 52.61 | 53.36 | 53.06 | 52.61 | 53.31 | 53.13 | 52.41 | 52.85 |
| **Relationship Status (Son) in 1881/1891/1911** |
| Head | 82.75 | 86.03 | 83.27 | 84.63 | 88.08 | 84.43 | 80.73 | 84.56 | 80.11 |
| Son | 4.26 | 6.61 | 7.02 | 4.40 | 5.98 | 7.07 | 5.75 | 8.25 | 9.78 |
| Visitor | 0.63 | 0.38 | 0.54 | 0.56 | 0.30 | 0.42 | 0.71 | 0.40 | 0.58 |
| Lodger | 3.30 | 1.54 | 2.10 | 2.84 | 1.16 | 1.69 | 1.10 | 0.31 | 0.45 |
| Boarder | 2.39 | 1.27 | 1.96 | 2.29 | 1.06 | 1.73 | 4.22 | 2.29 | 3.44 |
| **Marital Status (Son) in 1881/1891/1911** |
| Single | 12.33 | 11.17 | 13.40 | 12.44 | 10.35 | 13.55 | 15.53 | 13.40 | 16.82 |
| Married | 83.57 | 85.50 | 82.48 | 84.08 | 86.75 | 82.75 | 81.46 | 84.33 | 80.00 |
| **Occupational Structure (Son) in 1881/1891/1911** |
| Agriculture | 15.71 | 23.76 | 14.39 | 13.26 | 20.25 | 12.35 | 10.30 | 16.48 | 9.71 |
| Manufacturing | 60.19 | 56.29 | 60.94 | 62.03 | 57.52 | 62.01 | 59.74 | 56.02 | 59.86 |
| Services | 24.10 | 19.95 | 24.67 | 24.71 | 22.23 | 25.64 | 29.96 | 27.50 | 30.43 |
| **Residential Region (Son) in 1881/1891/1911** |
| London | 15.91 | 7.74 | 18.06 | 15.61 | 6.91 | 15.98 | 12.77 | 4.53 | 6.43 |
| Extra London | 8.90 | 8.06 | 6.66 | 9.95 | 10.79 | 8.72 | 13.41 | 12.56 | 12.16 |
| Lancashire | 13.32 | 8.17 | 13.37 | 13.92 | 8.96 | 15.39 | 13.62 | 11.49 | 16.66 |
| Yorkshire | 12.68 | 15.22 | 16.44 | 12.85 | 15.26 | 16.41 | 13.25 | 14.51 | 16.01 |
| **Birth County (Son)** |
| Greater London | 18.15 | 10.07 | 13.48 | 19.75 | 12.64 | 14.88 | 22.13 | 12.34 | 12.16 |
| Lancashire | 10.96 | 6.53 | 9.94 | 11.29 | 6.89 | 11.19 | 12.12 | 9.85 | 13.77 |
| Yorkshire | 12.15 | 14.42 | 15.92 | 11.52 | 14.40 | 15.77 | 12.52 | 14.26 | 15.88 |
| Observations (*N*) | 1,291,487 | 68,329 | 68,329 | 1,445,779 | 86,884 | 86,884 | 2,148,480 | 164,318 | 164,318 |
| Match Rate (%) |  | 5.29 | 5.29 |  | 6.01 | 6.01 |  | 7.65 | 7.65 |

*Notes*: “Population” includes all men aged 35-45 in 1881, 1891, and 1911 when comparing with the sons in the linked sample and all men aged 30-55 in 1851, 1861, and 1881 when comparing with the fathers. “ML” refers to the double- or triple-linked sample. “Manufacturing” in Occupational Structure also includes Mining and Transport sectors. “Extra London” refers to the regions of Middlesex, Kent, Essex, and Surrey that are not included in “London.” “Greater London” refers to the entire regions of Middlesex, Kent, Essex, and Surrey. “Yorkshire” includes all Ridings of Yorkshire. All numbers are in percentages unless stated otherwise.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **E Binscatter Plots**

Figures E1, E2, and E3 show the binscatter plots for 1851-1881, 1861-1891, and 1881-1911. The relationship between fathers’ and sons’ outcomes is clearly linear.

Figure E1**:** Binscatter plot for 1851-1881



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Figure E2**:** Binscatter plot for 1861-1891



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Figure E3**:** Binscatter plot for 1881-1911



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **F Rank-Rank Correlations**

Figure F1**:** Father-son rank-rank correlation, 1851-1881



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Figure F2**:** Father-son rank-rank correlation, 1861-1891



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Figure F3**:** Father-son rank-rank correlation, 1881-1911



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

**G Computing Correlations from IGE**

In the literature on intergenerational earnings mobility, a standard alternative to IGE (*β*) is the intergenerational correlation (*ρ*), which can be calculated by multiplying *β* with the ratio of the children’s and parents’ standard deviation (*σ*) of log earnings (Black and Devereux, 2010):

|  |  |
| --- | --- |
| $$ρ=β({σ\_{1}}/{σ\_{0}})$$ | (G1) |

Table G1 shows the intergenerational correlations in occupational status calculated using the same formula.

Table G1: Intergenerational Occupational Correlations Computed from Elasticities

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1851-1881 | 1861-1891 | 1881-1911 |
|  | (1) | (2) | (3) | (4) | (5) | (6) |
|  | OLS | OLS | OLS | OLS | OLS | OLS |
| Elasticity | 0.402 | 0.414 | 0.384 | 0.405 | 0.391 | 0.408 |
| Correlations | 0.456 | 0.480 | 0.431 | 0.460 | 0.432 | 0.457 |
| Multiple Links | NO | YES | NO | YES | NO | YES |
|  |  |  |  |  |  |  |
| *N* | 257,844 | 66,965 | 267,089 | 84,097 | 597,517 | 161,568 |

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **H Life-Cycle Effects on IGE**

Additional checks were conducted to determine whether life-cycle effects had an impact on the IGE estimated. The first set of checks – “age controls” – involve running the OLS and IV regressions with the sons’ and fathers’ age and their square terms as controls. This had virtually no impact on the size of *β*. The same was true after narrowing down the age range of the fathers, which meant using only father-son pairs where the fathers were aged 35 to 45 at the start of the period (hence in the similar age range as the sons when their occupational statuses are taken).

 Since sons are linked across multiple censuses, it is also possible to estimate an IGE for each stage of their occupational trajectory. Taking the 1881-1911 sample as an example, the “early-career” *β* is estimated based on the sons’ occupational scores 10 years after their first census year (1891); the “mid-career” *β* is estimated 20 years after their first census year; the “peak” *β* is the benchmark chosen for this paper – 30 years after their first census year, when the sons are aged 35 to 45.

The results in Table H1 suggests that there may be some modest life-cycle effects depending on the sons’ age. Existing findings on life-cycle bias in intergenerational (permanent) income elasticities suggest that using annual incomes from sons at younger ages will lead to an attenuation bias on *β* while using annual incomes from sons at older ages will lead to an amplification bias (Haider and Solon, 2006). On the contrary, my results suggest that for occupational status, there may be an amplification bias from using the occupations of sons at younger ages, if we take the *β* estimated at around age 40 as the true level. In any case, any life-cycle bias observed here appears to be modest and there is no indication that my preferred estimates are under-estimating intergenerational mobility due to life-cycle effects.

Table H1: Life-Cycle Effects on β

|  |  |  |
| --- | --- | --- |
|  | OLS | IV |
|  | *β* | SE | *N* | *β* | SE | *N* |
| **1851-1881** |  |  |  |  |  |  |
| Age Controls | 0.414 | (0.004) | 66,965 | 0.679 | (0.007) | 65,700 |
| Narrower Father Age Range | 0.411 | (0.006) | 38,317 | 0.669 | (0.010) | 37,611 |
| Early-Career | 0.392 | (0.004) | 60,512 | 0.716 | (0.007) | 59,526 |
| Mid-Career | No Data – 1871 Census not available |
| Peak | 0.379 | (0.005) | 60,512 | 0.652 | (0.008) | 59.526 |
| **1861-1881** |  |  |  |  |  |  |
| Age Controls | 0.405 | (0.004) | 84,097 | 0.647 | (0.006) | 83,095 |
| Narrower Father Age Range | 0.405 | (0.005) | 47,549 | 0.659 | (0.008) | 47,067 |
| Early-Career | No Data – 1871 Census not available |
| Mid-Career | 0.424 | (0.004) | 83,163 | 0.677 | (0.006) | 82,181 |
| Peak | 0.402 | (0.004) | 83,163 | 0.646 | (0.006) | 82,181 |
| **1881-1911** |  |  |  |  |  |  |
| Age Controls | 0.408 | (0.003) | 161,568 | 0.624 | (0.004) | 159,723 |
| Narrower Father Age Range | 0.406 | (0.003) | 92,768 | 0.626 | (0.005) | 91,988 |
| Early-Career | 0.411 | (0.002) | 151,864 | 0.666 | (0.004) | 150,219 |
| Mid-Career | 0.385 | (0.003) | 151,864 | 0.608 | (0.004) | 150,219 |
| Peak | 0.378 | (0.003) | 151,864 | 0.604 | (0.004) | 150,219 |
|  |  |  |  |  |  |  |

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **I Quantile Regression Results**

Figure I1 shows the father-son association in occupational status from quantile regressions at the tenth, twenty-fifth, fiftieth, seventy-fifth, and ninetieth percentiles. The results suggest that the transmission of status is stronger between high-status fathers and sons than between their low-status counterparts. This may be explained by the fact that high-status families have more resources and avenues to protect the socioeconomic status of their future generations.

Figure I1**:** Quantile regression results, 1851-1911



*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **J Simulation of Minimum IGE**

Table J1 shows the mean *β* and standard errors from 1,000 OLS regressions on samples of randomly matched fathers and sons. For each period, the samples used are the same pool of fathers and sons as the ML linked sample but with random matching of fathers and sons. A total of 1,000 random samples were constructed for each period using this method. The mean *β* is therefore the minimum level of father-son association possible, and is very close to zero.

Table J1**:** IGE estimates from randomly matching fathers and sons

|  |  |  |  |
| --- | --- | --- | --- |
|   | 1851-1881 | 1861-1891 | 1881-1911 |
| *β* | 0.000086 | 0.000012 | 0.000092 |
| Standard Error | (0.004561) | (0.003968) | (0.002825) |
| *N* | 66,965 | 84,097 | 161,568 |
| *Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856). |

## **K Transition Matrices for Altham Statistics**

Table K1: Intergenerational Mobility Table for 1851-1881

|  |  |  |  |
| --- | --- | --- | --- |
| Son's Class | Father's Class in 1851 | Total | *N* |
| 1881 | W | S | F | U |  |  |
| W | 45.39 | 16.81 | 17.97 | 11.73 | 22.97 | 12,291 |
| S | 37.85 | 69.55 | 19.01 | 32.99 | 39.85 | 32,598 |
| F | 4.19 | 2.11 | 40.59 | 2.20 | 12.27 | 3,952 |
| U | 12.57 | 11.53 | 22.43 | 53.08 | 24.90 | 18,124 |
| Total | 100.00 | 100.00 | 100.00 | 100.00 |  |  |
| *N* | 7,501 | 30,086 | 6,139 | 23,239 |  | 66,965 |

*Notes*: W = white collar; S = skilled and semi-skilled; F = farmer; U = unskilled.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Table K2: Intergenerational Mobility Table for 1881-1911

|  |  |  |  |
| --- | --- | --- | --- |
| Son's Class | Father's Class in 1881 | Total | *N* |
| 1911911 | W | S | F | U |  |  |
| W | 53.76 | 21.51 | 20.02 | 15.68 | 27.74 | 40,417 |
| S | 33.67 | 64.98 | 18.95 | 36.36 | 38.49 | 79,544 |
| F | 2.33 | 1.04 | 38.43 | 2.36 | 11.04 | 6,137 |
| U | 10.24 | 12.46 | 22.60 | 45.61 | 22.73 | 35,470 |
| Total | 100.00 | 100.00 | 100.00 | 100.00 |  |  |
| *N* | 26,153 | 80,855 | 9,420 | 45,140 |  | 161,568 |

*Notes*: W = white collar; S = skilled and semi-skilled; F = farmer; U = unskilled.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Table K3: Intergenerational Mobility Table (Corrected) for 1851-1881

|  |  |  |  |
| --- | --- | --- | --- |
| Son's Class | Father's Class in 1851 | Total | *N* |
| 1881 | W | S | F | U |  |  |
| W | 55.23 | 15.49 | 16.85 | 10.70 | 24.57 | 9,206 |
| S | 31.79 | 72.41 | 15.20 | 30.44 | 37.46 | 25,776 |
| F | 3.02 | 1.58 | 45.96 | 1.55 | 13.03 | 3,147 |
| U | 9.96 | 10.52 | 21.99 | 57.31 | 24.94 | 14,963 |
| Total | 100.00 | 100.00 | 100.00 | 100.00 |  |  |
| *N* | 4,568 | 24,599 | 5,067 | 18,858 |  | 53,092 |

*Notes*: W = white collar; S = skilled and semi-skilled; F = farmer; U = unskilled.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

Table K4: Intergenerational Mobility Table (Corrected) for 1881-1911

|  |  |  |  |
| --- | --- | --- | --- |
| Son's Class | Father's Class in 1881 | Total | *N* |
| 1911 | W | S | F | U |  |  |
| W | 61.35 | 20.48 | 17.02 | 14.51 | 28.34 | 31,795 |
| S | 29.25 | 67.73 | 13.19 | 33.69 | 35.96 | 63,844 |
| F | 1.51 | 0.77 | 46.26 | 1.72 | 12.56 | 4,803 |
| U | 7.89 | 11.02 | 23.53 | 50.08 | 23.13 | 28,355 |
| Total | 100.00 | 100.00 | 100.00 | 100.00 |  |  |
| *N* | 19,045 | 66,983 | 7,331 | 35,438 |  | 128,797 |

*Notes*: W = white collar; S = skilled and semi-skilled; F = farmer; U = unskilled.

*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856).

## **L False Positives Arising from Using a Sample for Census Linkage**

A crucial part of census linkage is the restriction on uniqueness of matches – in order for person A from the first census to be matched to person B from the second census, there can be no other potential matches for both person A and B. The issue with using a 2 per cent sample in the 1851 census is that we can only ensure that there are no other potential matches for person A in the 1881 census, but we cannot be sure that there are no other potential matches for person B in the 1851 census because 98 per cent of the 1851 census has been cut off. To demonstrate how this may affect our linkage, I have generated my own two per cent sample of the 1851 census and tried to match people from this sample to the full 1881 census. This can then be checked for false positives by seeing how many of these matches can be found when linkage is conducted using the full census instead – if the use of a two per cent sample does not generate higher number of false positives, then we would expect all matches to be found in the full linked sample. Table L1 shows the results of this exercise.

Table L1: Test for Additional False Positives from Using 2% Sample, 1851-1881

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Matched with 2% Sample** | **Found in Full Linked Sample** | **Rate** |
| **±2 Years** | 6,245 | 5,395 | 86.39 |
| **±5 Years** | 6,557 | 5,235 | 79.84 |
| *Notes*: results for “±2 Years” are produced by allowing matches to differ by up to 2 years in inferred birth year while requiring matches to be unique within the same age band, and results for “±2 Years” are produced by allowing matches to differ by up to 5 years in inferred birth year while requiring matches to be unique within the same age band; “Matched with 2% Sample” refers to the linked sample obtained while using a 2% sample of the 1851 census; the same matching algorithm is then run with the use of the full 1851 census, and individuals who have been successfully matched while using the 2% sample and who can also be found in the linked sample with the full census are shown in the column “Found in Full Linked Sample”; the “Rate” is calculated as Matched with 2% Sample / Found in Full Linked Sample \* 100.*Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856). |

The results clearly suggest that using a two per cent sample can lead to additional false positives than using the full census. When linkage is conducted using a stricter requirement of only allowing birth years to differ by at most two years (the approach taken in this paper), only 86 per cent of the successful matches from the two per cent sample can be found in the full linked sample – in other words, potentially 14 per cent of matches could be additional false positives. When linkage is conducted using a more relaxed requirement of allowing birth years to differ by five years (as Long (2013) did), the additional rate of potential false positives caused by the use of the two per cent sample is moderately higher – over 20 per cent. This is because in the second linkage specification, the restriction on uniqueness is widened – matches have to be unique within a five-year band. Therefore, the removal of the other 98 per cent of the population from the census means that more people would be incorrectly identified as “unique.”[[4]](#footnote-5)

## **M Changes in Father’s HISCAM Score over 10-Year Period**

Table M1 shows how the father’s occupational status may change over a 10-year period. Despite the relatively short amount of time, only around 60 per cent of fathers’ HISCAM scores stay constant. The majority of the changes in HISCAM scores are relatively small in scale (less than 10). These changes could have plausibly been caused by either a misreporting or miscoding of occupations, or by temporary shocks to the father’s occupational status. However, there is still a rather sizeable amount of changes (slightly above 10 per cent of the sample) which are larger in magnitude (20 scores or above) that do occur in just 10 years. These are perhaps more likely to have been caused by data errors rather than actual shocks to a person’s status. Unfortunately, it is not possible to definitely conclude whether these changes occur because of transitory shocks to occupational status or errors in the reporting, recording, or coding or data, but the results should suggest that these changes are at least symmetrical – they are just as likely to move up as they are to move down.

Table M1**:** Changes in Father’s HISCAM Score across 10-Year Period

|  |  |  |  |
| --- | --- | --- | --- |
| **HISCAM Change** | **1851** | **1861** | **1881** |
| **-60** | 0.03 | 0.02 | 0.02 |
| **-50** | 0.11 | 0.10 | 0.09 |
| **-40** | 0.38 | 0.36 | 0.41 |
| **-30** | 1.20 | 1.01 | 1.21 |
| **-20** | 2.85 | 2.87 | 3.23 |
| **-10** | 11.43 | 12.38 | 12.96 |
| **0** | 63.56 | 59.96 | 61.42 |
| **10** | 14.49 | 16.80 | 14.08 |
| **20** | 3.69 | 4.04 | 4.12 |
| **30** | 1.54 | 1.70 | 1.70 |
| **40** | 0.54 | 0.57 | 0.62 |
| **50** | 0.15 | 0.18 | 0.11 |
| **60** | 0.03 | 0.02 | 0.01 |
| *Notes*: for 1851 and 1861, changes are calculated as (father’s score in 1861 - father’s score in 1851), and for 1881 it is (father’s score in 1891 - father’s score in 1881); all figures have been rounded down to the nearest decile if they are lower than 0 and rounded up to the nearest decile if they are above 0; if there are no changes in HISCAM scores, they are kept as 0. *Sources*: author’s analysis of I-CeM (UKDA, SN 7481) and I-CeM Names and Addresses (UKDA, SN 7856). |

**References for Appendix**

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. “To the New World and Back Again: Return Migrants in the Age of Mass Migration.” *Industrial and Labor Relations Review* 72,no. 2 (March 2019): pp. 300-322.

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. “Automated Linking of Historical Data.” NBER Working PaperNo. 25825, June 2020.

Anbinder, Tyler, Dylan Connor, Cormac Ó Gráda, and Simone Wegge. “The Problem of False Positives in Automated Census Linking: Evidence from Nineteenth-Century New York’s Irish Immigrants.” CAGE Working Paper No. 568, June 2021.

Arkell, Tom. “Identifying the Census Enumerators – Cornwall in 1851.” *Local Population Studies* 53 (Autumn 1994): pp. 70-75.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. “How Well Do Automated Linking Methods Perform? Lessons from US Historical Data.” *Journal of Economic Literature* 58, no. 4 (December 2020): pp. 997-1044.

Feigenbaum, James J. “Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940.” *Economic Journal* 128, no. 612 (July 2018): pp. F446-F481.

Ferrie, Joseph P. “A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules.” *Historical Methods* 29, no. 4 (1996): pp. 141-156.

Haider, Steven, and Gary Solon. “Life-Cycle Variation in the Association between Current and Lifetime Earnings.” *American Economic Review* 96, no.4 (September 2006): pp. 1308-1320.

Higgs, Edward. *Making Sense of the Census Revisited: The Manuscript Returns for England and Wales, 1801-1901*. London: HMSO, 1989.

Higgs, Edward. *Making Sense of the Census Revisited: Census Records for England and Wales, 1801-1901 – a Handbook for Historical Researchers*. London: The National Archives and Institute of Historical Research, 2005.

Higgs, Edward, Christine Jones, Kevin Schürer, and Amanda Wilkinson. “Integrated Census Microdata (I-CeM) Guide.” Online document, University of Essex, Department of History (September 2013). https://www.essex.ac.uk/-/media/documents/research/‌icem\_guide.pdf?la=en.

Long, Jason. “The Surprising Social Mobility of Victorian Britain.” *European Review of Economic History* 17, no. 1 (February 2013): pp. 1-23.

Massey, Catherine G. “Playing with Matches: An Assessment of Accuracy in Linked Historical Data.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 50, no. 3 (2017): pp. 129-143.

Schürer, Kevin, Edward Higgs, and FINDMYPAST LIMITED. *Integrated Census Microdata (I-CeM), 1851-1911*. 4 April 2014. Distributed by UK Data Service. SN 7481, http://doi.org/10.5255/UKDA-SN-7481-2.

Schürer, Kevin, Edward Higgs, and FINDMYPAST LIMITED. *Integrated Census Microdata (I-CeM) Names and Addresses, 1851-1911: Special License Access*. 2nd Edition. 21 December 2015. Distributed by UK Data Service. SN 7856, http://doi.org/10.5255/UKDA-SN-7856-2.

Tillott, P. M. “The Analysis of Census Returns.” *Local Historian* 8, no. 1 (1968): pp. 2-9.

Tillott, P. “Sources of Inaccuracy in the 1851 and 1861 Censuses.” in *Nineteenth-Century Society: Essays in the Use of Quantitative Methods for the Study of Social History*,edited by E. A. Wrigley, pp. 82-133. London: Cambridge University Press, 1972.

Van Leeuwen, Marco H. D., and Ineke Maas. *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven University Press, 2011.

1. The requirements for an enumerator were: a person of intelligence and activity; able to read, write, and have some arithmetic knowledge; able to undertake the requisite physical exertion involved; must not be younger than 18 or older than 65; must be temperate, orderly, and respectable, conduct himself with strict propriety, and have the goodwill on the inhabitants of his district (women were allowed to become enumerators after 1891). [↑](#footnote-ref-2)
2. Linking procedures for both methods can be found in Abramitzky et al. (2020). [↑](#footnote-ref-3)
3. Though my review of the literature on census linkage is by no means exhaustive, I have yet to come across any work that have used this method to check for false positives. [↑](#footnote-ref-4)
4. This is not to say that allowing birth years to differ by five years would entail a significantly higher rate of false positives in itself, only that it would cause a significantly higher number of false positives when linkage is conducted using a sample rather than the full census. [↑](#footnote-ref-5)