Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM

Gill Newton and Robert J. Bennett

ghn22@cam.ac.uk rjb7@cam.ac.uk

Working Paper 24:
Working paper series from ESRC project ES/M010953:
Drivers of Entrepreneurship and Small Businesses

University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure, Downing Place, Cambridge, CB2 3EN, UK.

February 2020

Comments are welcomed on this paper: contact the authors as above.

> Newton, Gill and Bennett, Robert J. (2020) *Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM*. Working Paper 24: ESRC project ES/M010953: 'Drivers of Entrepreneurship and Small Businesses', University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure.

Keywords: Entrepreneurship, Employers, Self-employment, Small businesses, Census

JEL Codes: L26, L25, D13, D22

Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM

Gill Newton and Robert J. Bennett

Working Paper 24: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge

1. Introduction

This paper outlines a methodology for algorithmic record linkage to track individuals between census years focusing on employer and own account business proprietors contained in the population censuses for 1851-1891. The data used derive from the data deposit of the *British Business Census of Entrepreneurs (BBCE)* at UK Data Archive/Service (UKDS).[1] BBCE uses the transcripts of the censuses, and coding of individuals, mostly derived from the UKDS data deposit of *The Integrated Census Microdata* (I-CeM),[2] and adds additional transcripts for 1871 not in I-CeM using S&N as an additional source.[3] The BBCE extends I-CeM by identification and coding of entrepreneurs, data enrichment, and various I-CeM coding corrections and infills of those detected as missing.[4] The addition of 1871 to the BBCE data base is crucial to extend the scope to track people between census years.[5] The BBCE and I-CeM can be linked through the individual identifiers for each entrepreneur identified in the censuses to provide cross sections of census information. This paper

[1] Bennett, Robert J., Smith, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2020), *The British Business Census of Entrepreneurs 1851-1911 (BBCE)* [data collection]. UK Data Service, SN: pending.

[2] Schurer, K., Higgs, E. (2014). Integrated Census Microdata (I-CeM): 1851-1911. [data collection]. UK Data Service. SN: 7481, http://doi.org/10.5255/UKDA-SN-7481-1; see also Higgs, E., Jones, C., Schürer, K. and Wilkinson, A. (2015) *Integrated Census Microdata, 1851-1911, User Guide version v. 2 (I-CeM.2),* Second edition, Colchester: Department of History, University of Essex. https://www1.essex.ac.uk/history/research/icem/documentation.html

[3] For 1871 data input from S&N see WP 12: https://doi.org/10.17863/CAM.2748

[4] The I-CeM version used in BBCE is enhanced from I-CeM v.1 at UKDS. It is a Provisional version of I-CeM v.2 to be deposited at UKDS developed by Kevin Schürer, with inputs from an I-CeM team at Campop, and corrected codes for entrepreneurs by the BBCE team: see *BBCE User Guide*.

[5] Bennett, Robert J., Smith, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2020) *The British Business Census of Entrepreneurs 1851-1911 (BBCE)*: *User Guide*, https://doi.org/10.17863/CAM.47126

describes how record linkage can be developed from these BBCE/I-CeM cross-sections to track individuals over time. The data were prepared using support from the ESRC-supported project ES/M010953 (PI: Bob Bennett) 'Drivers of Entrepreneurship and Small Businesses'. The aims of this project are summarised in WP 1; the main data BBCE extraction methods are described in WPs 3 and 4; 1871 data assembly is discussed in WP 12; see WP 21 for additional people added to BBCE not in I-CeM. All WPs available are listed at the end of the references with further details and updates at www.bbce.uk.

Opportunities to track individuals between UK historical censuses at the scale of the whole population have opened up with the release I-CeM. Record linkage is developed here for the specific case of entrepreneurs in I-CeM and now in the BBCE with other supplemental data. Entrepreneurs are identified in I-CeM/BBCE from occupational declarations in the England and Wales population censuses of 1851-1881, and from those selecting the employer category for 'employment status' in the 1891 census. The starting point is prior extractions of employers and masters for the 1851-81 censuses.[6] In line with the analysis of these data in Bennett et al. (2019), we refer to the combination of employers and own account as entrepreneurs, which gives all self-employed.[7] In addition to entrepreneurs, the paper also examines a smaller stratified random quota sample of non-entrepreneurs to give an indication of differences in record linkage success and characteristics of entrepreneurs and the general population.

The paper is a pilot of methodology to inform subsequent developments. It investigates for census record linkages of entrepreneurs the use of one of the mainstream algorithms available, the Jaro-Winkler string comparison method, extended by with fuzzy name frequencies, data blocking, and preparatory data standardisation. It uses key five linkage variables: surnames, forenames, birthplaces, ages and sex. The next section of the paper summarises the challenges that historical data record-linkage face. Section 3 outlines the method applied. Section 4 introduces the two aspects used: tracking entrepreneurs, and a comparative sample of non-entrepreneurs. Section 5 assesses the main results.

[6] Bennett, R.J. and Newton, G., 'Employers and the 1881 Population Census of England and Wales', *Local Population Studies* (2015), p.29-49; van Lieshout, C., Bennett, R., Smith, H.J. and Newton, G., 2017. 'Identifying businesses and entrepreneurs in the Censuses 1851-1881'. WP 3 (2017), doi:10.17863/CAM.9640.
[7] Bennett, Robert J., Smith, Harry, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2019) *The Age of Entrepreneurship: Business Proprietors, Self-employment and Corporations Since 1851*, Abingdon: Routledge. https://doi.org/10.4324/9781315160375

The paper is restricted to England and Wales, and to 1851-91, but can be readily extended to the equivalent data for Scotland in BBCE/I-CeM and to other years where data allow. The linkage, as a pilot, had relatively restricted resources so that a high emphasis was placed on achieving an acceptable set of results within a limited scale of computing and operator time. The evaluation suggests possible areas for development in the future.

2. Historical census record linkage: the nature of the challenge

British historical census data contains no unique person identifiers that persist over time, and other identifying information is scant. Hence, individuals have to be identified from the information that is available which is often too unspecific to provide confident matches. As a result it has to be accepted at the outset that it will be difficult to achieve both very high rates of recall (the proportion of all true matches existing that have been found) *and* precision (the proportion of matches that are accurate). Even in manual record linkage by expert human operators, there will always be individuals who cannot be linked because their recorded characteristics are either too unspecific, or are too commonplace to distinguish them with certainty from others. When we have a source such as the census that purports to cover all persons, it is tempting to imagine that every person alive on a given census night must have been enumerated and must therefore be locatable in the digital versions of the census we now have at our disposal. The main impediments are damaged or lost archival records (which are significant in the 1851 and 1861 censuses) , imperfect transcription, and database coding and harmonization, the imperfect recording of some census information, and the inherent ambiguity of popular names and birthplaces(especially in the most populous, typically urban locations).

The extent to which false positive matches are tolerated varies considerably in applications with some recent historical record linkage research proposing methods that, while sophisticated in their execution, achieve very low real-world precision of only 0.1: where true positive matches are outnumbered by false positives by nine to one.[8]

[8] Zhichun Fu, H M Boot, Peter Christen and Jun Zhou: ‘Automatic Record Linkage of Individuals and Households in Historical Census Data, *International Journal of Humanities and Arts Computing* 8.2 (2014), 218, 220.

For most practical research it is necessary to obtain datasets that have higher level of accuracy so that subsequent analysis can interpret an individual's development over time with some confidence. This usually means that it necessary to sacrifice recall to obtain higher precision. A recent overview of census record linkage in favour of this balanced approach is provided by Ruggles et al. (2018).[9] This summarises the plurality of techniques that have emerged in historical research, each fitted to the particular sources used and the specific aims of researchers, with incremental progress resulting from automation, increased sophistication of algorithms, and improvements in computing power. However, on the whole, computational record linkage has reduced the time and effort needed to produce results, without much improvement in accuracy over what human operators can achieve.

Mathematical models of probabilistic record linkage have informed the approach taken by the US Bureau of the Census, often working with richer more recent data which may have a 'truth' data set for comparative purposes.[10] A universal model that is practically useful is a difficult to develop for historical record linkage because the personal information on which it depends is case-specific, often fairly minimal and frequently inconsistent. This is true even when limited to the specifics of historical census record linkage. In Britain the constituent countries (England, Scotland and Wales) may have different languages, dialects, or terminology where the popularity distribution of values of key variables such as names differs, and information is gathered differently in different locations and time periods.

It is now common to favour machine learning approaches that start from manually created training datasets, but these training data are time-consuming to obtain, and in any case there is no guarantee of perfectly accurate and consistent solutions when the record sets to be matched are very large and where there is no 'truth' data that can be used as a starting point. While it is possible to generate artificial training datasets exemplifying 'good matches' by duplicating the input and treating each pair and its duplicate as matches, optionally transforming the copy set slightly, this cannot reflect the range of real variations in the data.

For the current problem, no training set of 'truth' or ideally matched data exists. After some unpromising initial experiments with various approaches the paper adopts the FEBRL (Freely

[9] Steven Ruggles, Catherine A. Fitch and Evan Roberts: 'Historical Census Record Linkage', *Annual Review of Sociology*, 44, forthcoming 2018 (https://doi.org/10.1146/annurev-soc-073117-041447)
[10] For an overview see William E Winkler: 'Matching and Record Linkage', *WIREs Computational Statistics*, 6 (2014), 313–325

Extensible Biomedical Record Linkage) version of the Jaro-Winkler string comparison algorithm, modified in various ways to work with fuzzy name frequencies, and extended to include data blocking, and manipulations of cut-off criteria.[11]

3. Method

The starting point for the record-linkage is employers, masters and others identifiable from occupational declarations in the censuses 1851-1881, and those selecting the employer category for 'employment status' in the 1891 census. We refer to these as entrepreneurs, or 'Ents', further defined below. The whole population is referred to as 'All'.

3.1 Selection of variables

For census record linkage we need personal information that persists over the ten year gap between each census and is ubiquitous across the population, but with a level of specificness to individuals to be an effective discriminator between different people. Surnames, forenames, birthplaces, ages and sex are the only persistent characteristics that meet these criteria. Surnames and birthplaces both yield a large number of distinct values and their distribution is least skewed to a small number of popular values, although it is far from trivial to draw hard-and-fast boundaries between values. Forenames, birth years calculated from stated ages, and sexes are less varied but easier to treat as discrete values.

Other variables considered but not used were occupation and place of residence, because these would introduce bias towards those who did not change jobs or locations, both of which are expected to be frequent over the 10 year gaps between censuses. Forenames of all others in the household is too indiscriminate and biases towards larger households, where many present are in any case servants or lodgers who are transient and are unlikely to be present in the same household in the previous or succeeding census. Marital status is also often transient, with many over the periods examined being unmarried at the start, and/or widowed at the end. Middle names seemed initially useful, but experiment showed them to be

[11] Peter Christen, Tim Churches and Markus Hegland (2005) *Febrl - Freely extensible biomedical record linkage*, originally from Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Springer Lecture Notes in Artificial Intelligence, Volume 3056; Release 0.3.1, http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html (accessed 2018)

infrequently present in the census record and highly abbreviated. For example, in our largest record linkage exercise (1881 entrepreneurs to 1891 all persons) less than a quarter have a middle name value on both sides of the match (48,760 out of 217,552 matches). Only 1.3% of these were longer than two characters. Since the abbreviations are usually initials (and capital letters are particularly prone to mis-transcription), this was deemed insufficiently reliable.

However, we used forenames of children under 5 or over 10 years to account for the elapse of time between censuses, and forenames of spouses in calibrating acceptability thresholds for matches made. Table 1 below summarises the variables used directly by our record linkage algorithm and how these relate to the BBCE/I-CeM datasets.

Description	**Variable name in I-CeM/BBCE**	**Source in BBCE**	**Match method**	**Required?**
Forename string	Pname	I-CeM special licence and S&N	**fuzzy** string matching	Yes
Surname string	Sname	I-CeM special licence and S&N	**fuzzy** string matching	Yes
Standardised birthplace county (with foreign-born category)	Cnti	I-CeM /BBCE adapted from Day (2018) birthplace county look-up table	**exact**	Yes
Standardised birthplace polygon	BPPolygon (ultimately derived from I-CeM Bpstring)	Adapted from Day (2018) coding of I-CeM/BBCE birthplace strings	**exact** – represents most probable quasi-parish geolocation of coded birthplace	No Cnti only if no value
Sex (M, F or U)	Sex	I-CeM and S&N	**exact** allowing U-M or U-F	Yes
Age in whole years	INTage	I-CeM and S&N	**exact** meaning Age ± 10 in next/previous Census	Yes
Entrepreneur	Ent	BBCE variable for extraction GROUP	Extraction via algorithm	Yes at start; not in whole population

Table 1. I-CeM and BBCE UK Census variables used by the matching algorithm.

3.2 Overview of matching algorithm

Matching proceeds pairwise between sets of pre-identified candidate entrepreneurs at a given census date and the *whole* population at the next or preceding census date. To achieve a tractable level of computational efficiency, blocking by birthplace county, sex and exact age is applied. Preliminary investigation of smaller record linkage runs without these constraints indicated that only a very small proportion of true matches crossed these boundaries.

Because we aim to develop a high level of confidence in the record links achieved so that subsequent inferences drawn from the data will be robust, we heavily prioritise minimising the false positive rate. As a result we discarded all potential matches where a minimal confidence threshold was not met, or where multiple competing possibilities of equal strength existed. Rather than calculate all possible matches for each individual and then take the best available, the matching algorithm instead followed an iterative process, finding the best-matching records first and removing them from further consideration, then using the remaining records to test fuzzy potential matches of decreasing strength.

Names were compared using the Jaro-Winkler method to calculate distances between the original strings adapted by using fuzzy name frequencies, so that matches between common strings (such as the surname Smith and its congruents) were penalised in order to ensure they were accepted only where there was a very high degree of agreement between other variables. The Jaro string comparator is a commonly used record linkage method that computes the number of common characters in two strings, the lengths of both strings, and the number of transpositions to compute a similarity measure that ranges 0 – 1.0. The Winkler improvement extends Jaro by taking into account the generally more frequent occurrence of typographical errors towards the end of words by giving an increased weight to characters in agreement at the beginning of the strings.[12]

Birthplace was compared using encoded standardisations of the original string obtained by pre-processing the census data using Day's (2018) standardised birthplace and location

[12] W.E. Winkler and Y. Thibaudeau (1991*) An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census*, Research Report RR91/09, US Bureau of the Census; E.H. Porter and W.E. Winkler (1997) *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, Research Report RR97/02, US Bureau of the Census.

data.[13] This pre-processing increases the level of I-CeM data cleaning and enhancements in I-CeM and BBCE enhancements through further correction and standardisation. Day was also able to apply an equivalent consistent method to code the S&N birthplace and locational data for the 1871 data in BBCE that is not included in I-CeM (see WP 12). This ensures there is full standardisation across the five census years used.

3.3 Procedural steps

(i) Initial blocking and identification of match candidate pool

Blocks for all subsequent record-linkage searches were defined so that only those with the same birthplace county code were ever compared, to reduce the scale of search to a tractable level. Further quasi-exact equivalences were required on sex, age and most-probable birthplace polygon, as detailed below:

Sex:	M=M or F=F or M=U or F=U	*(in all cases)*
Age:	age=age+10 or age=age-10	*(depending on match run)*
Birthplace:	birthplace county code = birthplace county code	*(in all cases)*
	birthplace polygon = birthplace polygon	*(only if birthplace polygon present on both sides)*

Sex, age and birthplace county were as in the I-CeM or S&N source data (Sex, INTage and Cnti variables respectively), with substitutions in **Ents** input for BBCE corrections of age, sex, or joining of split strings between people in some cases (see WP 3). Birthplace polygon was derived from the most probable consistent parish level GIS polygon according to the Day (2018) assignment of I-CeM birthplace strings (Bpstring variable) to birthplace codes (variable BP_CODE), and thence to GIS polygon (variable BP_POLYGON)(see also GIS acknowledgements). Where there was more than one equally probable polygon, which was rare, one was chosen at random, but always the same one for a given value of Bpstring to maintain cross-comparability.

[13] J Day (2018) Enriching I-CeM: Matching Individuals Birthplaces to a GIS, Unpublished Working Paper, University of Cambridge.

This created an initial, broadest possible set of candidates. In earlier iterations of the algorithm, experiments with occupations or residence information were also used, but these were discarded as potentially biasing the result (as noted above), but also because they were not found to be needed.

(ii) Initial name matching constraint set

Fuzzy match similarity score of forename and surname was calculated between all pairs of records constituting the match candidates identified in (i) (also see below). All match candidates with a similarity score lower than 0.88 on either forename or surname were discarded.

(a) Determining a winning match from the candidate set.

Two separate passes were made through all candidate matches that were formed by the blocking and initial fuzzy name matching described in (i) and (ii) above. These passes ensured that each record from both input sets ultimately featured in only one winning match, or was eliminated from the final match output entirely.

In the example given in Figure 1 below, in Pass 1 a match candidate from the input set for census Year 1 (e.g. **1851 Ents**) started out with four potential matches to the input set for census Year 2 (e.g. **1861 All**). Three of these match candidates were eliminated using the name matching method described below, leaving the strongest match candidate for reconsideration. However, it remained possible for this Year 2 candidate to be matched also to one or more further candidates in Year 1. In Pass 2 of this example, the surviving match candidate from census Year 2 was found to have three potential matches in the input set belonging to census Year 1. Two of these match candidates were eliminated by the same name matching methods as before, leaving one winning match, which was the strongest overall.

X	**match candidate**
~~X~~	**eliminated match candidate**

Pass 1:

Year 1	**Year 2**
	~~X~~
	~~X~~
X	**~~X~~**
	X

Pass 2**:**

Year 1	**Year 2**
~~X~~	**X**
~~X~~	
X	

Figure 1. Example of how matches were chosen.

(b) Match calibration: performed manually (see below) and through setting of match acceptability thresholds.

3.4 Name matching method

In large measure, the success of the record linkage depends upon effective name matching and the identification of unambiguous match candidates. Distinguishing ambiguous from unambiguous matches with no truth data where the inputs are fuzzy is a crucial element, and is described in detail below.

Jaro-Winkler edit distance is a metric that produces a value between 0 and 1, representing the degree of similarity between a pair of strings.[14] It is based on the number of character transformations needed to turn one string into the other, boosted towards fewer transforms of

[14] W E Winkler: 'Matching and Record Linkage', in B. G. Cox et al. (eds.) *Business Survey Methods*, New York: Wiley (1995), 355-384.

the beginning sequence of the string (up to the first four characters). It is thus a refinement over the other most commonly used similar method, Levenstein edit distance. It is acknowledged that Jaro-Winkler is by no means a perfect way of evaluating true similarity between names or indeed any other natural language labels/words, especially short ones with few characters where a small number of changes can make a large difference to the match.

Sound matching algorithms (such as Double Metaphone or Soundex) tend to over-group names and, without further manual intervention, have generally been found to produce high numbers false positives. In comparison, Jaro-Winkler distance was developed specifically for use with US Census and other administrative records, and thus is appropriate to the task at hand. It is an expedient choice given that the number of comparisons to be made is far too large for human operators to perform, where a computationally efficient method is needed.

Name matching was performed by calculating the Jaro-Winkler distance between names in a prospective match. The Jaro-Winkler threshold value chosen for eligibility or rejection of the match varied depending on stage of match processing reached. This cascade of progressively higher thresholds aimed to ensure that accepted matches for records with multiple potentially valid match pairings passed progressively higher standards of name similarity. Since the existence of multiple potentially valid pairings is by definition an indicator that the record is similar to several others, this has the effect of exerting more stringent controls on matches between persons with common names, birthplaces and ages in an attempt to limit false positives in the cases where they most frequently occur.

Jaro-Winkler distance ranges over 0 – 1.0. For those matches that passed the initial threshold of Jaro-Winkler eligibility of 0.88 and where the same record featured in other matches (and there was thus competition between matches), three different eligibility thresholds were set for surnames with reference to the name’s popularity (popular, middling, or low). Popularity was defined by fuzzy name frequency value bands as follows:

Popularity band	**Defined as fuzzy name frequency**	**Jaro-Winkler eligibility threshold**
Popular	>=0.004	0.97
Middling	>=0.002<0.004	0.95
Low	<0.002	0.88

These values were obtained by qualitative evaluation of prospective match results, intended to minimise false positives. Note that the above thresholds do not imply that fuzzy name frequencies are defined by the stated Jaro-Winkler eligibility thresholds: the values are the thresholds for the popularity bands only.

3.5 Fuzzy frequencies for name matching

Common names may sometimes be expressed with uncommon variations in spelling that are presumed to be accidental, as a result of misspelling, mis-transcription etc.; for example, “Johnso” for “Johnson”. Fuzzy name frequencies were conceived and implemented to try to ensure that such uncommon variants do not score a much lower frequency than their popular congruents, leading to unwarranted confidence in matches between records that are in reality much weaker than the name frequency scores alone might suggest.

Fuzzy name frequency is a probabilistic calculation intended to go some way towards representing how common a name truly is, setting aside the vagaries of spelling and transcription. It is an augmentation of name frequency (meaning the proportion of the total population having a given name, pre-calculated for all names in the match run) that adds to this a share of the popularity of other popular names that a name closely resembles.

By scrutiny of a reference name dataset (1851 England and Wales all persons forenames and surnames), a cut-off threshold defining what constituted a popular name was obtained as follows. From a histogram of the frequency count of records with each name sorted in descending frequency order, a point of sizeable transition in frequency was identified, such that a reasonable number of names remained to the left of the transition point (important for surnames in particular as there was a steep transition after the single most common name, and using a set threshold would provide a set of popular names with just one member). The popularity cut-off constituted a rounded version of the proportion of all persons having the name at the identified point of transition in the frequency distribution.

Fuzzy name frequencies were calculated for all records in each match-run according to the following definition:

$$fuzzyfrequency(name) = \sum_{\substack{popname \\ \in \\ Popnames}} similarity(name, popname) \times frequency(popname)$$

where:

name = Pname (i.e. forename) or Sname (i.e. surname) of record
popname = **name** whose **frequency** exceeded the **popularity cut-off**
similarity = 0.08 ≤**Jaro-Winkler** ≤1, scaled to 0 – 1, or 0 if **Jaro-Winkler** <0.08
frequency = proportion of all persons in population having this **name**
popularity cut-off = threshold set at **frequency** 0.002 if **name** is a surname or 0.02 if **name** was a forename
Jaro-Winkler = edit distance between two strings, weighted by similarity of the first four characters

3.6 Match calibration

Accepted matches that were output from the record-linkage algorithm were subsequently refined using a reference extended dataset comprising **1851eEnts -> 1861 All matches**, using name similarity and popularity scores augmented with spouse forename on both sides of the match, and forenames of children aged under 5 in 1851ents or aged 10 to 14 in 1861all (listed in age order), for all heads of household. The source for the England and Wales name frequencies for this reference data set was created from the 1851 I-CeM data. This extended match output was repeatedly reordered using weighted geometric means of name similarity, name popularity, and birthplace polygon likelihood, to arrive at a single matchscore for evaluation purposes. Threshold values of acceptability for the key match variables involved in matchscoring were ascertained by repeatedly drawing samples of the newly restricted output and checking them manually for mismatches in spouse or child names, calibrating the initial values upwards or downwards to reduce mismatches to <10% while retaining as many matches as possible.

The aim of this calibration was to further restrict the match output to eliminate remaining false positives and obtain precision for the record linkage exercise of >0.9. The final calibration restrictions on matches used were as follows:

Ent on both sides of match	BP Polygon probability	Name popularity	Name similarity
TRUE or FALSE	>0.3 AND <1	<0.000008	>0.8
TRUE or FALSE	≥0.5 AND <0.8	≥0.000008	>0.9
TRUE or FALSE	>0 AND <0.5	≥0.000008 AND <0.008	>0.96
TRUE or FALSE	>0 AND <0.5	<0.000008	>0.9
TRUE or FALSE	≥0.8 AND <1	≥0.000008	>0.9
TRUE or FALSE	≥0.8 AND <1	<0.000008	>0.8
TRUE or FALSE	1	≥0.000008	>0.9
TRUE or FALSE	1	<0.000008	>0.8
TRUE or FALSE	0	<0.00002	>0.93
TRUE	0	<0.00008	>0.93

Name similarity was here the product of surname similarity and forename similarity (similarity as defined above), and Name popularity was forename frequency + forename fuzzy frequency multiplied by surname frequency + surname fuzzy frequency, in each case from whichever side of the match had the highest value. A birthplace Polygon probability of 0 implies that the match was made on birthplace county alone for those cases where no birthplace polygon was available on either or both of the pair of records.

3.7 Limitations of the record linkage method

The results section below presents the match success rate and a detailed evaluation of the output obtained using the above method, together with preliminary conclusions relating to entrepreneurs that can be drawn from it. However, there are some general restrictions arising from the method that are discussed in brief here, with a view to identifying areas for subsequent improvements and in order to clarify the suitability of the method for other purposes.

The most obvious constraint is that there is no attempt to assess what proportion of the input can be matched or to achieve the highest possible rate of recall. Thus, record-linkage applications where it is important to know who remained in observation in a later source, and who did not remain *through record-linkage itself alone*, will not be solvable with this method. Measurement of mortality rates is an example of this. However, there is potential to improve the rate of recall by pre-processing, searching for clusters of co-resident individuals, and by relaxation of the exact correspondence constraint on age, although the last two possibilities are likely to increase considerably the amount of processing time needed, and the first increases the human input needed before record linkage can begin.

It would have been desirable to repeat the calibration for all match outputs rather than relying on a single reference set, but the time-consuming task of manual identification of mismatches for evaluation of threshold values did not permit this. It might also have been useful to experiment with requiring the winning candidate from the pool of match candidates to have a certain minimal difference from other candidates to be considered successful, as a possible alternative or addition to the present method of calibration.

Performance was a concern throughout, with processing time running to double digits of hours per matchrun. Our sample size for the candidate pool on one side of each matchrun did not exceed 700,000 person records, some 2% of the total population enumerated. However, each candidate pool was matched to a destination pool of the whole population enumerated in an adjacent census, constituting up to 29 million person records, which was essential to avoid missing strong, correct matches and, worse, introducing weaker and incorrect matches restricted to the sub-population of entrepreneurs. As the method, in effect, discards popular name and birthplace combinations that induce multiple or ambiguous identification, and high thresholds of match acceptability are subsequently applied, this approach minimises the risk of misidentification arising from one side of each match run being a sub-population.

Whether the method is fully scalable and the match success rate could be maintained with a larger candidate pool is untested. High Performance Computing hardware facilities might speed up performance considerably, and the algorithm would probably need to be re-coded for multi-threading on HPC clusters, which is a significant task that could be developed in subsequent research.

4. Inputs

Record linkage is developed here for two pilots: first, for the matching with the adjacent years of all Ents identified in 1851-81 and employers in 1891; and second, undertaking the same record linkages for a stratified random quota sample of non-entrepreneurs.

4.1 Scale and characteristics of the inputs to record linkage: Ents

The entrepreneurs (Ents) input data derived from BBCE/I-CeM. It comprised for 1851 to 1881 all persons aged over 14 extracted from census occupational information as 'employing' a stated number of employees (including zero), and also masters/mistresses, farmers, owners of land and owners of other major assets such as mines and ships (see WP 3). For 1891, where the format of the census questions on employment changes, the data comprised those who indicated they had 'employer status'. As a pilot study those in 1891 who said they were working on their own account were not included since the number of employers already greatly exceeded the extractions from previous censuses. In essence this means that Ents in 1891 are defined much as they were conceived in the 1851 to 1881 census questions as employers of others.

Table 2 gives the characteristics of the Ents extracted in BBCE. These extractions range from 300,000 to 700,000 individuals, predominantly of middle-aged men. There were also female Ents, who were older on average than the men and often widowed, and other individuals ranging from 14 years to over 90 years, with retired individuals also identified. Just under half were farmers, except in 1891 where the change in census format promoted a much broader response from those employing others. Note that farmers throughout the discussion here are the single occupation code (I-CeM Occode 173: 'Farmer and grazier'; and non-farmer is all except 173).

Year	All persons enumerated in I-CeM	Entrepreneur GROUP extracted N	% farmer	Mean age	Sex ratio M/F
1851	17,704,457	385,530	61.8	47.3	11.1
1861	19,828,560	373,196	60.0	47.6	10.7
1871	n/a	298,208	57.1	48.0	8.3
1881	25,954,690	414,939	60.3	48.7	8.7
1891	29,050,639	672,395	21.3	45.7	6.6

Table 2. Characteristics of the entrepreneurs (Ents) extractions as used in record linkage.

4.2 Scale and characteristics of the inputs to record linkage: Non-Ents

Samples of the general population on non-Ents were extracted from the 1851, 1861, 1881 and 1891 I-CeM census data. As no full transcriptions for 1871 are not available in I-CeM, and as the BBCE data for 1871 contain only those extracted as ‘employers’, this year could not be used for non-Ents. The Non-ent sample aims to test record linkage methods on workers and own account who may be less stable than Ents. Because of limited resources a constrained sample was necessary, but also to limit the variance between cases this had to be carefully designed to allow a focus on sampled groups that allows comparison of like with like by reducing the range of different cases that would be otherwise included. Stratified random quota samples were drawn, with quotas within two strata: for sectors and geography.

Sectors

Twenty sectors were identified to pilot contrasted experiences of worker and own account characteristics likely to have different levels of switching, range of skill specialisation, different capital requirements to move into employer or own account status, different levels of gender participation, more and less localised in markets, and with a range of growth/decline histories over the period. Selection was based on I-CeM Occodes, after cleaning and correction by the BBCE team. This reduces the effect of spurious Occodes in the origin data, but leaves possible errors in the linked individuals since the BBCE cleaning was usually applied only to the economically active. None of the major textile sectors were included since the total sample size possible and the locations with good data survival coverage would not allow justice to be done to these industries without very constrained sampling within regions. Textiles should be a priority for future attention. Of course business proprietors in textiles and professions giving ‘employer’ information were already included in extractions used for the Ent record linkage. Professions were also excluded as they have a high non-response rate in the census which makes employer status distinctions problematic.

Geography

Counties were chosen as the geographical sampling frame because these give large potential samples within relatively uniform areas, and their boundaries were stable over the period considered. Five counties were selected to give regional diversity, urban-rural, city and town mixes, and to provide appropriate locations for the sector samples. These were: Bedfordshire, Durham, London, Oxfordshire, and Warwickshire. Of the possible candidates several were ruled out because record survival prevents linking of many of the individuals (see WP 23). First, the 1851 data have been truncated in I-CeM and it has not been possible to fully restore all of them (this is mainly an issue for Lancashire and Cheshire, and to a lesser extent parts of Yorkshire). Second, lost data from the records for 1851 and 1861 are highly concentrated spatially in some of the important places in other counties. In addition, Yorkshire is very large to use as a sample frame, with some major data gaps; the decision to exclude textiles also makes Yorkshire less relevant. However, despite data that are lost in the census records for London, this county was included to ensure that its different characteristics were reflected in the record linkage challenge. The final sample constructed prior to record linkage is shown in Table 3.

Sector (I-CeM Occode)	1851	1861	1881	1891
119. Commercial clerks	10,000	10,000	10,000	10,000
196. Coal Miners - underground	7,308	7,220	7,891	7,615
198. Coal Miners – others underground	2,780	2,890	2,339	2,535
246. Tinplate manufacturers	602	723	1,361	1,399
305. Nail manufactures	1,006	998	1,479	857
362. Bicycle makers & repairers	14	22	559	5,932
393. Piano & organ makers	2,618	2,951	4,842	5,740
405. Builders	2,909	3,598	7,306	7,772
412. Bricklayers	11,450	10,824	10,918	11,065
426. Gasfitters	1,307	2,272	4,393	4,775
437. Cabinet makers	7,820	8,755	13,037	13,603
506. Tanners & fellmongers	1,785	1,143	1,879	1,158
646. Straw mat manufacturers	3,910	3,407	7,399	7,287
650. Milliners (not retail)	14,557	13,368	14,674	12,868
652. Milliners (retail)	280	350	485	880
653. Tailors (not merchants)	18,037	15,918	17,248	18,258
663. Shoe & boot makers & repairers	21,207	14,759	17,473	20,378
691. Bakers (dealers)	13,378	12,456	14,717	14,216
693. Sugar Refiners	1,157	1,210	987	785
709. Brewers	3,056	2,698	4,122	4,195
758. General shopkeepers	12,506	12,619	17,251	18,561
765. General labourers	10,000	10,000	10,000	10,000
Total	147,687	138,181	170,360	179,880
Mean age (years)	**34.6**	**35.3**	**35.5**	**35.2**

Table 3. Sample for non-entrepreneur pilot across 5 counties.

The 20 sectors chosen cover 22 I-CeM occupational codes because the I-CeM subdivisions of coal miners underground and milliners were both included. For the two very large occupational groups of commercial clerks and general labourers the sample was limited to a quota of 10,000 by random selection in each year. For bricklayers, bakers, tailors, milliners, and shoe makers all non-London were accepted, and then the rest of the sample topped up to 10,000 from a random sample in London. For the rest of the sectors all individuals were accepted; hence the final sample is a mix of full and random coverage by sector/location. In addition to the worker and own account elements of the sample, which was the vast majority, all those employers identified by the reconstruction method for supplementing census non-respondents were included in the 20 sector strata (see WPs 9, 9.2, and Montebruno et al, 2020). This had the aim of allowing a test of how far the reconstruction methods were successful in matching employers who had been non-respondents. The reconstruction method used was EMPLOYSTATUS_NUM (see WP 9, and BBCE *User Guide*). The differences in sample size, as random selections with quotas, largely reflect the differences in the size of the occupational categories over time, as well as lost data in 1851 and 1861.

5. Assessment of results

5.1 Record linkage success rate

Table 4 presents the results of record linkage using the method outlined in Section 3, giving the number and proportion of persons matched between censuses with a high rate of confidence, for each entrepreneur (Ent) matchrun and the additional four matchruns of the Non-Ent samples.[15] In each case the Ents or Non-ents were matched to the whole population of the chronologically adjacent census. For Non-ent matchruns, only those individual records not forming part of an Ents extraction and thus previously included in an Ent matchrun were used, hence the total number of records deployed for linkage in each of these matchrun was somewhat lower than the total sample sizes given in Table 4.

[15] Note that because the 1871 data do not have access to the whole population (but only Ents) there can be no 1871 match to All.

Matchrun Entrepreneurs	**N matches**	**% matched**
1851 Ents -> 1861 All	72,844	18.9
1861 Ents <- 1851 Al	96,248	25.8
1871 Ents <- 1861 All	65,205	21.9
1871 Ents -> 1881 All	65,624	22.0
1881 Ents -> 1891 All	104,132	25.1
1891 Ents <- 1881 All	217,532	32.4
Non-entrepreneurs		
1851 -> 1861 All	15,142	10.9
1861 <- 1851 All	21,586	16.3
1881 -> 1891 All	28,551	17.5
1891 <- 1881 All	40,325	25.3

Table 4. Record linkage success rate per matchrun (backwards links in grey)

Between one in six and one in three Ents, and between one in ten and one in four Non-ents, were successfully matched with a high degree of confidence. This is a promising result taking into account the simplicity of the method. It represents a substantial gain in recall over what initial explorations indicated might be achieved with no standardisation and no fuzzy string matching of any kind, with improved precision since the most likely sources of equivalence in the attributes of prospective record matches (i.e. semantically identical personal names and birthplaces) have been accommodated, and ambiguous candidates dropped.

A feature of the success rates is a level of improvement over time. There are several factors that may explain this. The data were created through different transcribers not all using the same method, with 1851 known to have generally lower but very variable transcription quality. In addition the underlying sources have varying levels of archival survival, with 1851 and 1861 most adversely affected by missing records, which alone is sufficient to explain much of the lower linkage rates achieved for 1851 and 1861, whilst 1871 Ents are a more restricted database and derived from a different genealogical source (S&N) to those used in I-CeM.

Improvements in literacy, resulting in more accurate name, birthplace and age reporting may also affected match rates. We might expect literate and numerate persons to be better at providing the enumerator with consistently expressed names, birthplaces and ages, all of which are key variables in successful record linkage. Equally, better skilled and more knowledgeable enumerators would be easier to find. However, the relationship between literacy/numeracy of the person enumerated and record linkage success is complicated by the fact that such details are typically provided by the head of household, and the likelihood of being head of household differ between the Ents and Non-ent sample. Some evidence for the effect of improvements in numeracy might be indicated by the extent of age heaping, measuring the extent of reported ages ending in 0 and 5 compared to the total population in a particular age range, as per Whipple's index. Values obtained from the Anderson's (1988) and Anderson (1979) 2% Sample of the 1851 census for Britain suggest continuous improvements in the numeracy of the general population from the cohorts born between 1820 and 1870 (the same cohorts most likely to appear as Ents in the 1851 to 1891 censuses). The Whipple Index values fell from 123 to 110 (where 100 would indicate no age heaping).[16] This brought Britain close to Germany and Sweden, world leaders in numeracy and literacy.[17] There was an upward trend between 1850 and 1913 in the ability to sign names at marriage, which improved for men from just under 60% to 99%, and for women from just under 55% to 99%, with the introduction of compulsory elementary schooling from 1870 increasing the pace of improvements already underway since the early 1800s or before.[18] High levels of human capital have been associated with extensive record-keeping.[19]

Entrepreneurs generally have a higher match success rate compared to the Non-ent general sample, which may reflect higher literacy and numeracy. However, it may also reflect a more thorough approach to recording by enumerators of entrepreneurs who were more prominent and better known by correct name than many of the non-entrepreneurs. In the Non-ent sample

[16] Michael Anderson (1988) Households, Families and Individuals: Some Preliminary Results from the National Sample from the 1851 Census of Great Britain. *Continuity and Change*, 3, 421-38; Anderson, M., B. Collins, and C. Scott (1979) *National Sample from the 1851 Census of Great Britain*. [data collection], UK Data Service, http://doi.org/10.5255/UKDA-SN-1316-1.

[17] Dorothee Crayen and Joerg Baten (2010) 'Global Trends in Numeracy 1820-1949 and its implications for long-term growth', *Explorations in Economic History*, 47, 82-89 Figure 1.

[18] Roger Schofield: 'Dimensions of illiteracy, 1750-1850', *Explorations in Economic History*, 10.4, 1973, 437-455.

[19] e.g. Lars Sandberg 'The Case of the Impoverished Sophisticate: Human Capital and Swedish Economic Growth before World War I', *Journal of Economic History*, 39.1 (1979), 225-241.

the effect of literacy is only weakly confirmed by the variation in the record linkage success rate between occupations. The main occupation consistently associated with high rates of literacy and numeracy that had generally higher match rates was clerks; but high matching was also achieved for a wide range of manufacturers and retailers (Tables 11 and 12).

Backwards matchruns attempt to link a candidate pool of records from one census backwards in time to the whole population of the preceding census. They generally attract the highest rates of success, as can be seen in Table 2 (except for 1871, for reasons noted earlier). This is because in backwards matchruns there is no natural attrition of the candidate pool in the intervening period due to mortality, unlike forwards matchruns where some candidates will die before the next census. In consequence, the match success rate for backwards matchruns most closely approximates the rate of recall for our record linkage exercise (the number of true positive record matches obtained divided by all true positive matches that potentially exist).

Precision is far harder to measure in the absence of truth data with linked records, but thresholds set during match calibration aimed for >0.9, meaning no more than 10% potentially false positives were accepted, and it is hoped that >0.95 has been achieved.

5.2 Success rate by age

Forward matchruns from one census to the next differ from backwards matchruns because all individuals in the candidate match pool are survivors, who have not died or out-migrated since the last census. As all our candidates are aged over 14 years and only ten years will have elapsed since the last census, virtually all of them should, in theory, be detectable in the earlier census, excepting a small proportion who can be expected to have migrated into England and Wales only after this earlier census took place. This contrasts with forward matchruns, where mortality exerts a significant toll on the candidate match pool, especially as our Ent candidates are mostly middle-aged or older. This means the hypothetical proportion that could be matched to the next census is lower, and especially so in the oldest age groups. Of course, if our samples included young children, it is they whose forwards matchrun match success rate would be most affected by mortality, but in their absence the oldest age groups should show the lowest match success rates.

1851 Ent -> 1861 All		1861 Ent <- 1851 All	
age in 1851	% matched	age in 1861	% matched
		14-19	28.9
14-19	16.6	20-29	27.3
20-29	22.2	30-39	25.5
30-39	22.3	40-49	25.8
40-49	21.4	50-59	26.2
50-59	19.1	60-69	25.2
60-69	14.0	70-79	25.3
70-79	6.7	80-89	23.8
80-89	1.6	90-99	16.3
90-99	0.0	100-109	0.0
1861 Ents -> 1871 All		1871 Ents <- 1861 All	
		age in 1871	% matched
No matchrun possible: no data for 1871 ***all*** *persons*		14-19	22.4
		20-29	23.6
		30-39	21.1
		40-49	21.4
		50-59	22.2
		60-69	22.0
		70-79	22.1
		80-89	20.6
		90-99	15.9
		100-109	0.0
1871 Ent -> 1881 All		1881 Ents <- 1871 All	
age in 1871	% matched		
14-19	17.9	*No matchrun possible: no data for 1871* ***all*** *persons*	
20-29	25.2		
30-39	25.4		
40-49	24.4		
50-59	23.0		
60-69	17.8		
70-79	8.9		
80-89	2.0		
90-99	0.0		
1881 Ent -> 1891 All		1891 Ent <- 1881 All	
age in 1881	% matched	age in 1891	% matched
		14-19	32.4
14-19	23.2	20-29	32.6
20-29	28.9	30-39	31.0
30-39	31.1	40-49	32.4
40-49	29.2	50-59	32.8
50-59	25.7	60-69	33.0
60-69	19.6	70-79	35.1
70-79	9.5	80-89	34.0
80-89	2.0	90-99	33.7
90-99	0.2	100-109	0.0

Table 5. Age specific match success rates: comparing backwards and forwards matchruns. Note: persons of unknown age excluded.

Table 5 demonstrates that the expected age specific match rate differences between forwards and backwards matchruns is indeed observed. In the forwards matchruns, considering the success rates for each successive age group, it is clear that at older age groups the match rate declines and then dwindles to zero for the very oldest, all of whom may be assumed to have died between censuses. In contrast, the match success rates in the backwards matchrun remain almost stable across each successive age group, until the very oldest again, but for different reasons. This is likely a consequence of being reported by others, with a few in institutions that have less accurate information about age or birthplace. Age mis-statement or inconsistent reporting of age or birthplace have more serve effects because the initial blocking uses exact age and birthplace. In English historical data such as ours, age misreporting among the oldest age groups occurred because of deliberate exaggeration or debility. In mid-nineteenth century censuses, the size of cohorts over age 80 have been found in some cases to be artificially inflated by 45%, rising to as much as 300% for those over 90.[20] Misrepresentation of ages may also affect the (apparently) older cohorts of the forwards matchrun, but the greater influence on declining match rates in this case is mortality.

5.3. Success rate by marital status

Another test is the proportion of women who apparently transition from single to married or *vice versa*, since the customary surname change on marriage should make it impossible to trace them give the method followed. As no use was made of marital status in record linkage or calibration, this is independent of what was considered to make a match. In fact, less than 1.3 per cent of female matches involve a marital status transition of this type. In a small proportion of the individual cases inspected, from surrounding context such as street addresses, it seems that either the marital status is incorrect, or that contrary to custom, the woman did not change her surname – or perhaps married a cousin or other person of the same surname meaning that no change occurred. Match rates by marital status, shown in Table 6, have systematically lower success rates for women than men, for all female matching as a result of name change on marriage/remarriage, as expected. Also as expected the married have higher match success than single or widowed of both genders, though single men almost achieve the same match rates as married men. There is also improved match success for later compared to earlier matchruns in line with the general trend of matching.

[20] R. Lee and D. Lam (2011) Age distribution adjustments for English censuses, 1821 to 1931, *Population Studies*, 37, 3, 464.

Matchrun	Sex	single	%	married	%	widowed	%	married spouse absent	%	divorced	unknown/not recorded	%
ICEM mar code:		**1**		**2**		**4**		**3**		**5**	**9 or null**	
1851 ents matched to 1861 all	F	602	11.2	533	15.9	2,444	11.7	226	11.4		9	4
1851 ents matched to 1861 all	M	8,471	19.1	54,499	20.5	3,866	14.0	2,016	16.5		66	3.6
1861 ents matched to 1851 all	F	1,047	20.6	758	14.7	4,075	21.1	379	16.8		37	19.1
1861 ents matched to 1851 all	M	10,811	25.4	69,037	26.8	6,486	25.2	3,361	23.6		257	21.3
1871 ents matched to 1861 all	*no data on marital status for 1871*											
1871 ents matched to 1881 all												
1881 ents matched to 1891 all	F	2,484	18.0	668	22.2	4,137	17.0	255	18.3		16	14.3
1881 ents matched to 1891 all	M	11,107	24.6	78,445	27.5	5,104	15.6	1,803	21.5		113	21.1
1891 ents matched to 1881 all	F	7,903	30.0	2,973	20.0	10,810	26.1	926	17.7		11	6.3
1891 ents matched to 1881 all	M	22,750	33.0	153,253	33.8	12,667	32.9	5,668	29.4	1	52	16.0

Table 6. Record linkage success rate by marital status. Note: persons of unknown sex excluded; rate not calculated where fewer than 5 Ents in a category

Middle names provide a third way of evaluating the reliability of the record linkage results, since they were also excluded for matching purposes. It is more time-consuming to make this check because there are many ways of expressing the same name, not least as initials and other highly abbreviated forms. As discussed above, a large proportion of middle names are initials only, which is unsurprising given the small amount of space provided in the Householder's form and the Census Enumerator's Book *pro forma*. For this reason, only differences in the initial letter of middle names were evaluated. Of all 2,843 matches eligible for the middle names comparison in the largest matchrun (1881 ents -> 1891 all), only 146 5.1% do not begin with the same letter. Among these, clerical checks revealed frequent mis-transcription of difficult to read capital letters, so the true rate of error is lower than 5%.

5.3 Geographical variation in record linkage success

Name diversity is not geographically uniform: isolated, especially upland rural communities, tend to have a high proportion of individuals with the same surname (often related to each other), whereas large and densely populated cities have a lot of individuals who share exactly the same name by chance. Both scenarios are potential obstacles to achieving good rates of record linkage success, since in our method the emphasis on avoiding false positives means that the existence of conflicting equally viable potential matches leads to no match being accepted (even though one of the matches will normally be correct). Birthplaces are also geographically uneven in specificity and intelligibility. While many individuals leave their community of birth, Ents are probably relatively more immobile, after achieving mid-adulthood. Insofar as this immobility extends to the whole life course, Ent birthplace and place of residence on census night are more likely to be equivalent, or at least spatially close or related, potentially meaning that same county of birth and residence is more likely among Ents than the general population.

More generally, for the urban-born, birthplaces can be either too specific in that only one parish in a community is named, or too vague in that the entire city is named. Refinements to birthplace coding could improve treatment of instances that fall into the former category, but for this record linkage pilot, only the most likely administrative unit is considered. In any case locational matches are constrained because there will remain a large number of birthplace records in the census that cannot be very precisely attributed.

Other problems that constrain the geography of matching arise from for the Welsh-born, where there are greater problems of translation or transcription due to Welsh language place names. This is exacerbated by a more highly constrained pool of forenames and surnames in common use, perhaps itself also in part a result of transliteration and transcription issues arising from Welsh language names.

In general, record linkage rates were best in predominantly rural English rather than Welsh counties, and south of the Severn to Humber line, as shown in Figure 2 and Table 7. These show high quality match success rates by Ent county of enumeration in 1851 and 1881. In the figure, the scale is arranged in natural breaks, meaning that each match run has a different key, to make the two maps more cross-comparable given the general improvement in match success rate over time. Rutland, Huntingdonshire and Sussex do particularly well, and in the later match run this extends to several other counties in the South East and South West. Wales, in contrast, fares particularly poorly. Urbanised northern counties such as Durham and Lancashire also do relatively badly. London, if visible at this scale would be shaded blue if shown (i.e. the lowest match rates), reflecting its very large size, mobile population, and its coverage of parts of two counties (Middlesex and Surrey) are the least successfully matched counties of the South East.

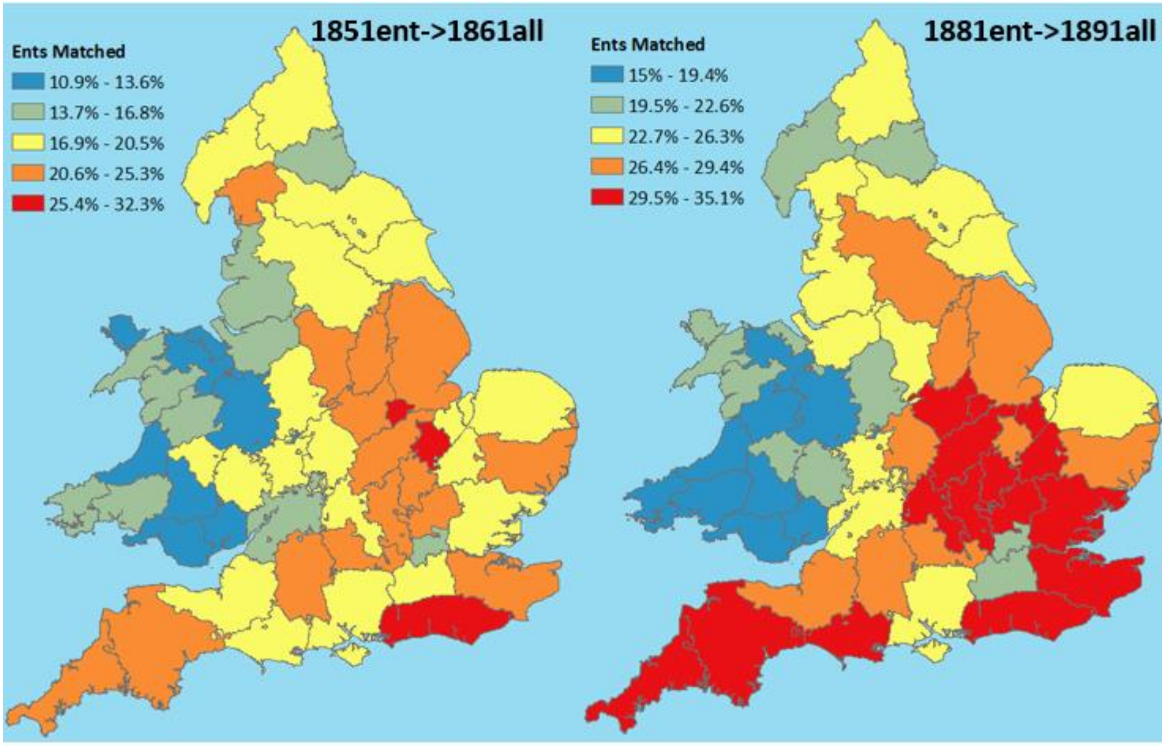

Figure 2. County variation of success rates in the earliest and latest match runs.

1851 Ent to 1861 All			1881 Ent to 1891 All		
1851 enum county	**% matched**	**n**	**1881 enum county**	**% matched**	**n**
Anglesey	10.9	231	Anglesey	20.8	399
Bedfordshire	22.6	578	Bedfordshire	35.1	930
Berkshire	22.1	771	Berkshire	29.0	1079
Brecknockshire	13.6	370	Brecknockshire	19.2	491
Buckinghamshire	24.3	654	Buckinghamshire	31.6	935
Caernarvonshire	16.2	669	Caernarvonshire	20.2	867
Cambridgeshire	20.1	973	Cambridgeshire	30.4	1481
Cardiganshire	12.7	733	Cardiganshire	16.8	1111
Carmarthenshire	15.4	796	Carmarthenshire	18.6	996
Cheshire	14.9	1561	Cheshire	23.5	2774
Cornwall	24.5	2527	Cornwall	30.8	3365
Cumberland	18.6	1381	Cumberland	20.0	1500
Denbighshire	11.5	434	Denbighshire	15.0	587
Derbyshire	22.2	1705	Derbyshire	25.4	2157
Devon	22.7	3822	Devon	30.6	4687
Dorset	20.1	843	Dorset	30.8	1159
Durham	15.6	1086	Durham	21.5	1627
Essex	19.5	1298	Essex	30.3	2076
Flintshire	13.5	153	Flintshire	20.3	200
Glamorganshire	12.9	585	Glamorganshire	19.3	1000
Gloucestershire	16.7	1215	Gloucestershire	25.6	1995
Hampshire	19.7	1304	Hampshire	25.1	1868
Herefordshire	17.9	695	Herefordshire	20.2	847
Hertfordshire	24.6	848	Hertfordshire	31.3	978
Huntingdonshire	27.6	442	Huntingdonshire	29.2	407
Kent	25.3	2498	Kent	30.2	3735
Lancashire	14.9	4209	Lancashire	23.5	8556
Leicestershire	22.6	1318	Leicestershire	32.1	1828
Lincolnshire	22.7	3502	Lincolnshire	27.8	4021
Merionethshire	16.8	564	Merioneth	20.5	690
Middlesex*	16.5	383	Middlesex	21.7	4265
Monmouthshire	11.5	439	Monmouthshire	17.9	638
Montgomeryshire	15.2	698	Montgomeryshire	18.1	807
Norfolk	18.5	2131	Norfolk	25.7	2439
Northamptonshire	23.8	1113	Northamptonshire	31.5	1361
Northumberland	20.3	1179	Northumberland	24.1	1396
Nottinghamshire	23.6	1651	Nottinghamshire	27.0	1786
Oxfordshire	20.3	729	Oxfordshire	30.3	1030
Pembrokeshire	14.8	441	Pembrokeshire	19.1	652
Radnorshire	17.9	327	Radnorshire	21.3	280
Rutland	32.3	329	Rutland	31.2	202
Shropshire	12.8	977	Shropshire	19.4	1527
Somerset	20.5	2249	Somerset	27.9	3267
Staffordshire	17.7	2103	Staffordshire	22.5	2709
Suffolk	23.1	2083	Suffolk	29.4	2287
Surrey	19.9	743	Surrey	22.6	2507
Sussex	27.2	2168	Sussex	33.4	2981

Warwickshire	19.7	1557	Warwickshire	28.5	2382
Westmorland	21.6	691	Westmorland	25.1	871
Wiltshire	22.7	1156	Wiltshire	28.0	1435
Worcestershire	20.4	990	Worcestershire	25.8	1391
Yorkshire E. Riding	18.1	1291	Yorkshire E. Riding	26.3	1883
Yorkshire N. Riding	19.5	1885	Yorkshire N. Riding	24.0	2417
Yorkshire W. Riding	18.7	5695	Yorkshire W. Riding	26.7	9127
*London	11.9	2111			

Table 7. County level variation in Ent record linkage success rates in the earliest and latest match runs

It is possible that literacy rates have some effects on linkage rates, since the geography of literacy improvements was uneven.[21] For example, North and South Wales and Monmouthshire were near the bottom ranked counties for literacy, especially among women, and hence may contribute to the lower record linkage success rates in Wales. Conversely, Sussex and Rutland had high rates of record linkage from an early date and both well have above-average literacy, ranking as the top two counties for female literacy by 1885. However, there is no simple relationship between a county's literacy ranking and its record linkage success rate: London and the northern counties of Cumberland, Westmorland, Durham, and North Yorkshire each have high literacy rates but relatively poor record linkage rates, but most of these counties and Wales have high archival loss for 1851 and 1861.

5.4 Potential firm-size effects on Ent matching

The Ent data from BBCE has the size of the workforce extracted and parsed along with the identification of individuals who were employers. This allows two ways of assessing the size of a business: the size declared as its workforce, and for farming and other land-based businesses the declared acreage of the holding. In assessing whether there were any effects of business size on match success rate, the workforce size is more valuable because it is given by both farming and non-farming employers. Workforce size reports also tend to be less clustered at particular thresholds than acreages.

[21] Literacy comparisons are drawn from the county tables in: *Eighth Annual Report of the Registrar General* (1845), p. lvii-lviii; *Eighteenth Annual Report of the Registrar General* p. vii; *Twenty-Eighth Annual Report of the Registrar General* (1955), p. x; *Thirty-Eighth Annual Report of the Registrar General* (1865), p. xxii, *Forty-Eighth Annual Report of the Registrar General* (1875), p. xxxix.

The general results for matching by firm size for each matchrun are shown in Table 8 for each gender. As with the other matching, there is generally better backwards than forwards matching, with surprisingly low match rates for the proprietors of very large firms in forwards matching for 1851. The matching is again far lower for women than men, especially in the largest firms. This reflects the likelihood that women are returned as proprietors in these cases often as a result of inheriting the business as a widow. The 1881-91 matching of course differs in context because of the different question in 1891 was more all-embracing of employers across all size categories. The lower match rates for large firms in forwards matching for 1851 may be associated with the generally lower transcription quality suspected for that year, but also interrelates with age as many of the largest firm as have older proprietors making the employee return: there are more likely to be match successes for older individuals backwards than for the same age forwards. It may also have arisen from the nature of the 1851 census process, where the smallest businesses were surveyed more thoroughly than any subsequent year, especially those with only one employee (see Bennett et al., 2019, Chapter 5; van Lieshout et al., 2020).

1851 Ents ->1861 All	F	%	M	%	All	%
firm size in 1851	matches	matched	matches	matched	matches	matched
none mentioned	3,133	19.1	44,368	27.3	47,591	26.4
0	13	3.9	497	10.9	511	10.5
1	149	3.7	5,060	10.8	5,220	10.2
2 to 4	294	4.2	10,095	12.8	10,412	12.1
5 to 9	145	5.6	4,759	14.9	4,915	14.1
10 to 19	62	6.1	2,649	16.1	2,727	15.5
20 to 49	16	5.1	1,195	15.7	1,211	15.2
50 to 99	1	1.8	184	11.9	185	11.5
100 to 199	1	5.3	57	8.5	58	8.4
200 to 249	0	0.0	8	5.6	8	5.5
250 to 499	0	0.0	31	10.6	31	10.4
over 500	0	0.0	15	9.6	15	9.4
1861 Ents <-1851 All	F	%	M	%	All	%
firm size in 1861	matches	matched	matches	matched	matches	matched
none mentioned	3,476	18.4	41,530	23.6	45,006	23.1
0	41	21.4	533	25.3	574	24.9
1	536	20.2	8,402	27.5	8,938	26.9
2 to 4	1,229	21.0	19,211	28.9	20,440	28.2
5 to 9	635	22.7	10,457	29.7	11,092	29.2
10 to 19	276	23.0	5,898	31.4	6,174	30.9
20 to 49	88	23.5	2,747	32.5	2,835	32.2
50 to 99	6	13.3	631	33.4	637	33.0
100 to 199	8	50.0	288	32.1	296	32.4
200 to 249	0	0.0	64	32.0	64	30.8
250 to 499	0	0.0	122	30.8	122	30.5
over 500	1	16.7	69	30.4	70	30.0

1871 Ents <- 1861 All	F	%	M	%	All	%
firm size in 1871	matches	matched	matches	matched	matches	matched
none mentioned	2,795	15.4	27,245	20.2	30,373	19.6
0	81	14.4	648	21.7	734	20.3
1	437	17.0	5,470	23.9	5,964	23.1
2 to 4	910	16.7	12,006	24.7	13,060	23.8
5 to 9	499	18.2	6,865	25.9	7,463	25.1
10 to 19	249	17.9	4,087	26.8	4,386	25.9
20 to 49	101	17.7	2,042	28.2	2,173	27.3
50 to 99	27	15.5	515	27.9	548	26.6
100 to 199	11	13.9	256	26.8	270	25.8
200 to 249	1	7.1	66	28.1	67	26.3
250 to 499	5	20.8	98	25.7	103	25.2
over 500	1	4.5	60	26.2	64	24.8

1871 Ents -> 1881 All	F	%	M	%	All	%
firm size in 1871	matches	matched	matches	matched	matches	matched
none mentioned	2,337	12.9	27,352	20.3	30,057	19.4
0	70	12.4	629	21.1	699	19.3
1	387	15.0	5,604	24.5	6,068	23.5
2 to 4	832	15.3	12,121	25.0	13,111	23.9
5 to 9	493	18.0	7,014	26.5	7,624	25.6
10 to 19	268	19.2	4,333	28.5	4,669	27.6
20 to 49	102	17.9	2,161	29.8	2,298	28.8
50 to 99	31	17.8	535	29.0	577	28.0
100 to 199	12	15.2	253	26.5	268	25.6
200 to 249	1	7.1	69	29.4	71	27.8
250 to 499	4	16.7	96	25.1	101	24.7
over 500	4	18.2	73	31.9	81	31.4

1881 Ents -> 1891 All	F	%	M	%	All	%
firm size in 1881	matches	matched	matches	matched	matches	matched
none mentioned	5,103	16.5	44,085	22.1	49,188	21.3
0	2	8.0	66	24.5	68	23.1
1	474	18.3	9,189	28.3	9,663	27.6
2 to 4	1,086	20.3	19,849	29.2	20,935	28.5
5 to 9	553	22.4	11,219	31.2	11,772	30.6
10 to 19	234	24.4	6,693	33.8	6,927	33.4
20 to 49	83	25.0	3,539	33.9	3,622	33.7
50 to 99	16	30.8	968	34.0	984	33.9
100 to 199	5	35.7	493	33.3	498	33.4
200 to 249	1	33.3	117	33.0	118	33.0
250 to 499	2	40.0	200	32.1	202	32.1
over 500	1	50.0	154	35.6	155	35.7

1891 Ents <- 1881 All	F	%	M	%	All	%
firm size in 1881	matches	matched	matches	matched	matches	matched
none mentioned	21,744	N/A	159,187	N/A	181,354	N/A
0	0	N/A	23	N/A	23	N/A
1	146	N/A	5,284	N/A	5,441	N/A
2 to 4	430	N/A	13,435	N/A	13,884	N/A
5 to 9	219	N/A	7,938	N/A	8,173	N/A
10 to 19	88	N/A	4,734	N/A	4,834	N/A
20 to 49	20	N/A	2,503	N/A	2,528	N/A
50 to 99	4	N/A	644	N/A	649	N/A
100 to 199	1	N/A	349	N/A	350	N/A
200 to 249	0	N/A	83	N/A	83	N/A

250 to 499	0	N/A	128	N/A	128	N/A
over 500	0	N/A	105	N/A	105	N/A

Table 8. Record linkage success rate per matchrun by firm size (number of employees) and sex. Note: the last part of this table for the 1891 Ents<-1881 All matchrun is not comparable to the rest of the table as there is no information on employees in the 1891 Census, so that only the 1881 size classes can be shown.

Note also that we are matching from Ent to all enumerated individuals. Hence, the closure or opening of any given business enterprise should in theory have no direct bearing on whether its owner can be traced over time in other census records. Hence, switches between entrepreneur and non-entrepreneur status should not affect linkage rates. However, there may be indirect effects of business exits on the record linkage success rate. At the extreme, an entrepreneur who ceases trading, especially one who falls into penury, may be more liable to disappear from records, perhaps through migration to new opportunities overseas, or as a result of becoming irregularly housed in an institution or boarding house which have often less accurately recorded age information in the census. Business closure as a result of ill health or as a cause of ill health may also result in the death of the owner, leading to a lower rate of potential matches. Nevertheless, whilst this may occur for some individuals, these possibilities do not seem to affect the matched rates achieved at a statistical level.

5.5 Occupational mobility between censuses

It is also possible to draw some preliminary conclusions on changes of occupation from one census to the next or previous for the Ent matching. The main purpose here is to assess the great disparity between male and female apparent changes of occupation. In forwards matchruns, whereas three quarters of men appear in the same occupational category from one census to the next, and this appears stable over time, a smaller proportion of women keep the same occupational category. In backwards matchruns, which tend to include younger cohorts, the proportion that change occupations among men and women is slightly higher. It is questionable whether the level of occupational change among Ent women is really so great or so different from the men, and in this respect it is informative to take a closer look at some actual matches to understand what may be really happening.

In the 1851 Ents to 1861 all matchrun there are 3,820 matches of female entrepreneurs with a stated occupation in 1851. Nearly half (42.8%) of these have no industry-specific occupational designator by 1861. The majority of these (constituting 871 women, or just under a quarter of all female entrepreneurs matched) have no occupational descriptor in 1861 at all. That this is a true reflection of their activities ten years later is unlikely in every case, especially for women who are widowed heads of households, where there must have been some source of income. A further 764 women are placed in a catch-all residual land and property owning occupational class whose members include those who live off investments, as 'house owner', 'fund holder' and so forth: which was a category of increasing size over the period. However, for some individuals the classification is misleading. Women with more detailed occupational descriptors may sometimes end up in this class too, especially where the descriptor 'wife'/'widow' of an occupation type is used (e.g. 'shoemaker's wife'). Some unlikely discrepancies arise when we consider what some of these women were doing ten years previously. In 1851 there were, for example, Ent females described as a blacksmith, market gardener, dress maker and dress master. However, in 1861 the same people had become retired blacksmith's wife, market gardener's widow, tailor and draper's wife, and gentlewoman, respectively, according to the 1861 census.

The first example may create an inequality with Ent males if retired men are typically classified to their former occupation. But the elapse of time between censuses in forwards match runs does mean that women are more likely to become retired and even paupers, and more so than men since women are generally an older group to begin with (and also quite probably a less wealthy group).

The other examples suggest the individual could well have continued in the same business activity as at the earlier census dates but that this was differently expressed, or even suppressed in the case of the high status gentlewoman. While difference of expression of an occupational descriptor might reflect some genuine change of perceived status originating from the women themselves, it could also reflect census enumerator prejudices, or the opinion of another household member who filled in the schedule. Some of these complexities can be handled by choice of how the descriptor 'xxx wife' etc. is dealt with in analysis.

5.6 Residential mobility between Censuses

The entrepreneurs linked in this pilot provide some valuable insights into mobility. Most Ents who can be traced between Censuses remained resident in the same county, registration sub-district and parish. The methodology takes account of are boundary changes by using continuous parishes (I-CeM variable ConParID). The level of mobility is shown for farmers and non-farmers separately in Tables 9 and 10.

Match run	**Same parish**	**Same RSD**	**Same county**
1851ent->1861all	74.0	86.0	93.8
1861ent<-1851all	71.8	84.7	92.9
1871ent->1881all	Not available	88.2	93.2
1871ent<-1861all	Not available	84.3	92.6
1881ent->1891all	70.0	81.8	88.8
1891ent<-1881all	68.3	81.7	89.5

Table 9. Percentage of Ent farmer/landowners staying in the same spatial unit 1851-1891.

Match run	**Same parish**	**Same RSD**	**Same county**
1851ent->1861all	74.4	82.6	91.5
1861ent<-1851all	69.0	78.9	90.1
1871ent->1881all	Not available	76.0	86.5
1871ent<-1861all	Not available	77.7	88.6
1881ent->1891all	69.8	76.7	79.3
1891ent<-1881all	64.7	73.1	84.4

Table 10. Percentage of Ent non-farmers staying in the same spatial unit 1851-1891.

Nearly all former Ents remained in the same county, about 80% remained in the same RSD, and two-thirds to three-quarters remained in the same parish between censuses. Non-farm Ents had slightly lower continuity of location, but this was still over 80% in the same county, over 75% in the same RSD, and over 65% in the same parish. The links are lower for 1881-

91, as expected by the changed and wider format of the census question in 1891. These are residential location linkages; we cannot know from this whether the business premises relocated or if the business address. Nonetheless these tables indicate a high degree of Ent community residential stability, with farmers most immobile, as one might expect for a group whose activities are directly tied to the land with tenancy agreement, and perhaps with more limited alternative opportunities. However, their stability is also surprising since there was a high level of agricultural change for the period, not least following the effect of the agricultural depression in the 1870s.

For all those tracked in the Ent population there was some evidence of increased mobility over time, especially at parish level. The extent to which this is driven by increasing urbanisation over the period needs to be interpreted with caution. It is worth remembering that farmers are, in the great majority of cases, rurally resident and display low levels of mobility as well as non-farmers. This makes it unlikely that the increases in mobility over time are driven wholly by the greater likelihood of a location having changed from rural to urban. However, greater intensity of moves into urban areas from rural ones almost certainly contributed to the observed increases in mobility over time.

Compared to the Non-ent sample, residential stability appears greater for Ents. For example, overall just over half (56%) of those in the 1881 Non-ent sample who could be linked to an 1891 census record stayed in the same parish, whereas well over two-thirds (70%) of 1881 Ents who could be linked to an 1891 census record stayed in the same parish. For Ents who remained Ents (had occupational information indicating their employer or farmer status in both censuses) the proportion that stayed in the same parish is even higher at 77.3%. The younger age of Non-ents compared to Ents does not account for this difference, since even if we restrict the Non-ents to those aged over 40 years (where the mean age of the becomes 51.5 years and thus slightly older than Ents 48.7 years), a third (34.7%) of the Non-ents move parish between the 1881 and 1891 censuses, compared to only just over a fifth (22.7%) of Ents. Among the Non-ent occupations sampled, commercial clerks were especially mobile, with nearly half (46.4%) of the 2,621 such persons who could be matched moving parish from an 1881 census record to an 1891 census record.

Commercial clerks are an interesting group for other reasons. Overall the Non-ent sample record linkage success rates from each matchrun was lower than for Ents, but there was considerable variation between occupational groups. For commercial clerks the match success rate was among the best of the Non-ent occupations, and often similar or better to the Ent match rates. For example, in 1881 to 1891 matching the success rate for commercial clerks was 26.2%, whereas for Ents it was 25.1%. General labourers had a much lower match success rate of 15.9% in 1881 to 1891 matching. Both commercial clerks and general labourers were relatively residentially mobile, and both were younger on average than Ents, especially the commercial clerks whose mean age was around 29 years, whereas it was 34 years for labourers.

5.7 Non-entrepreneur sample: matching rates

For the Non-ent sample there are some important implications to be drawn from the contrasts in match rates achieved. As already noted, the match rates are generally lower for Non-ents than entrepreneurs, but vary considerably by occupation and sector. The full record linkage success rate by county and occupation are shown in Tables 11 and Table 12, representing the earliest and latest forwards match runs, respectively.

Non-ent occupational groups that fared notably well in terms of the forward record linkage success rate in 1851 were clerks, many of the manufacturers in the sample and retail milliners, whereas sugar refiners were by far the worst, with general labourers below the mean match rate, but not as far below as expected. In the 1881 forward record linkage the highest success rate was still for commercial clerks, followed by several of the manufacturing sectors and retailers like milliners. The lowest again was sugar refiners, with general labourers again close to the mean of the match rates.

						ALL	**N**
Occupation	BEDFORD	DURHAM	LONDON	OXFORD	WARWICKS	COUNTIES	**matches**
119. Commercial clerks	25.7	21.2	13.1	28.1	19.4	14.8	1480
196. Coal Miners - underground	N/A	11.5	N/A	25.0	15.6	11.9	873
198. Coal Miners – others underground	N/A	11.9	N/A	16.0	10.4	11.9	327
246. Tinplate manufacturers	11.1	4.4	9.8	14.3	13.4	10.1	55
305. Nail manufactures	N/A	22.7	11.1	31.3	12.8	15.4	152
362. Bicycle makers & repairers	N/A	0.0	0.0	100.0	33.3	14.3	2
393. Piano & organ makers	N/A	30.0	14.4	40.0	12.0	14.5	369
405. Builders	33.3	25.0	12.0	22.7	15.2	13.4	293
412. Bricklayers	20.1	15.4	9.4	24.2	13.8	11.1	1216
426. Gasfitters	0.0	25.0	9.7	12.5	19.1	11.1	140
437. Cabinet makers	14.5	19.7	10.7	18.2	13.6	11.8	869
506. Tanners & fellmongers	12.8	25.2	9.4	17.1	18.6	11.3	196
646. Straw mat manufacturers	10.7	10.1	8.0	13.7	15.1	9.9	384
650. Milliners (not retail)	9.9	9.3	6.4	12.1	11.6	7.8	1122
652. Milliners (retail)	17.9	0.0	13.0	20.0	16.7	15.7	44
653. Tailors (not merchants)	22.2	15.6	8.2	18.9	13.2	11.2	1765
663. Shoe & boot makers & repairers	19.2	15.4	8.6	20.4	14.2	12.6	2388
691. Bakers (dealers)	19.6	14.4	7.8	19.2	14.7	9.8	1149
693. Sugar Refiners	N/A	25.0	2.8	0.0	38.5	3.3	38
709. Brewers	19.8	14.5	8.8	17.5	9.5	10.2	302
758. General shopkeepers	15.5	15.2	7.7	15.8	12.9	8.6	1075
765. General labourers	13.7	8.3	6.1	11.9	10.1	9.0	903
Total	14.4	13.2	8.8	17.5	13.8	**10.9**	15142

Table 11. Non-ent sample: forward record linkage success rate by occupation and county: 1851->1861 All

Occupation	BEDFORD	DURHAM	LONDON	OXFORD	WARWICKS	ALL COUNTIES	**N matches**
119. Commercial clerks	46.0	37.9	28.9	48.2	39.8	35.2	3520
196. Coal Miners - underground	0.0	22.7	0.0	0.0	33.0	25.1	1898
198. Coal Miners – others underground	0.0	23.1	28.6	40.7	31.1	24.8	624
246. Tinplate manufacturers	66.7	32.4	18.7	28.6	30.2	23.5	315
305. Nail manufactures	N/A	22.9	15.9	25.0	25.3	24.6	207
362. Bicycle makers & repairers	42.9	26.1	27.0	71.4	35.8	34.9	1983
393. Piano & organ makers	35.7	25.4	28.0	52.0	37.0	28.3	1532
405. Builders	42.6	35.5	25.2	45.0	35.8	27.2	1002
412. Bricklayers	41.3	27.1	19.4	33.6	27.3	24.4	2585
426. Gasfitters	39.5	24.0	22.6	25.5	33.5	24.1	1089
437. Cabinet makers	35.6	27.6	20.9	41.0	34.4	22.9	2798
506. Tanners & fellmongers	55.6	34.6	10.2	40.0	35.1	14.2	160
646. Straw mat manufacturers	31.0	20.0	17.9	0.0	27.8	30.3	2044
650. Milliners (not retail)	37.6	36.4	27.2	39.8	35.4	29.5	3574
652. Milliners (retail)	44.9	27.8	30.5	47.6	34.7	32.4	265
653. Tailors (not merchants)	50.2	26.3	16.6	39.8	28.4	23.7	3393
663. Shoe & boot makers & repairers	48.6	26.8	20.4	38.9	30.1	25.2	4290
691. Bakers (dealers)	40.5	23.4	16.8	37.1	29.4	21.7	2383
693. Sugar Refiners	30.0	39.3	17.6	50.0	11.5	18.3	140
709. Brewers	51.5	23.7	22.0	31.6	25.0	24.1	957
758. General shopkeepers	33.1	28.7	17.4	30.8	29.1	20.0	3478
765. General labourers	30.6	19.7	14.8	27.9	21.7	20.9	2088
Total	35.2	26.6	20.9	35.9	31.4	**25.3**	40325

Table 12. Non-ent sample: forward record linkage success rate by occupation and county: 188l->1891 All

5.8 Comparing urban/rural differences

Ents and the Non-ent sample of the general population have different propensities to move residence between censuses, with Ents less likely to move. When we consider the type of settlement in which a person resides, it is clear that most of the additional residential stability of Ents compared to Non-ents was concentrated among those residing in urban areas or places adjacent to towns. This may explain their higher stability: they did not need to move with the growth of increased market opportunities as urbanisation developed.

The classification of continuous I-CeM parish units into urban categories can be used to assess this further (see WP 6). The urban classification is based primarily on population density. Four categories are defined: urban, rural, urban proximity, or transition. The characteristics of the 1881 Ent and Non-ent sample matches in each of these urban categories that have been successfully matched to 1891 census records are shown in Table 13.

Sample type	All matches	Mean age/years	Matches unchanged residence parish (ConParID)	% immobile
Urban Non-Ent	22,657	41.7	12,005	53.0
Urban Ent	30,841	42.8	20,002	64.9
Rural Non-Ent	1,223	47.0	872	71.3
Rural Ent	41,050	46.9	29,333	71.5
Urban Proximity Non-Ent	1,409	41.8	844	59.9
Urban Proximity Ent	6,869	46.1	4,827	70.3
Transition Non-Ent	3,262	44.3	2,179	66.8
Transition Ent	25,235	45.8	18,608	73.7

Table 13. Mobility between 1881 and 1891 Censuses by parish of residence type, Ents and Non-ents compared.

The final column of Table 13 shows the proportion that did not move parish between censuses and are were immobile in the sense that they stayed in the same continuous parish unit, although it is certainly possible they moved house and changed street address over the ten year intercensal period. Equally, and especially in urban areas where parishes tend to encompass smaller areas, it is possible that some who moved across a parish boundary and so count as mobile did not move far geographically, and remained part of the same community after their move. However, both changes of address within parish boundaries and moves within the same settlement that crossed parish boundaries affect Ents and Non-ents alike, so that comparisons remain valid on the relative mobility of the two groups. The most striking aspect of this table is the differences between revealed in these immobility figures. The greatest difference of more than twelve percentage points is between urban-resident Ents and Non-ents. On average these groups were similar although not identical in age. Age is important because, other things being equal, younger economically active adults are likely to be more mobile than older people because of the costs and benefits associated with moving at different stages in the lifecycle: on leaving home, finding work, settling into an occupation, marrying, having children, etc. This suggests the existence of a stable core of business owners even in towns where high levels of population turnover were the norm. Although the stability was higher for Non-ents in urban proximity areas, and transition areas, these also had strong differences with the highly stable Ents.

By contrast, there is virtually no difference between rural resident Ents and Non-ents, and on average these groups were identical in age. A gradient in the Ent/Non ent residential mobility differential from rural (low differential) to urban (high differential) exists, with Ents in the other two categories of proximity to urban areas and transitional areas being more different to Non-ents. However, some caution in interpreting these differences is advisable since the average age difference is larger, particularly in the urban proximity category.

6. Conclusion

This paper demonstrates that record linkage of historical census data for England and Wales can be fairly readily achieved using I-CeM and BBCE. The method developed combines the Jaro-Winkler approach, but adds fuzzy string comparison, data blocking, and data pre-processing based on just a few variables (forename, surname, birthplace, age, sex). The method is applied to a situation where it is not necessary to seek a match for all or even a majority of individuals, and a high quality matched subsample is sufficient. Fuzzy frequency calculations for names and birthplace probabilities are key to identifying and discarding false positives - where individuals whose characteristics are not distinctive enough to achieve an unambiguous match. To achieve these results some manual calibration was necessary to set final acceptability thresholds for matches.

Record linkage of entrepreneurs and a non-entrepreneur sample of the general population taken from one census and matched to all persons enumerated in the next or preceding census have demonstrated that this method is appropriate for samples of between 100,000 and 700,000 individuals from a population of up to 29 million persons. The expected yield of matched individuals suitable for analysis obtainable ranges from 11% to 32% (recall estimated at 16-32%). Smaller samples should also work; larger samples may not be tractable without significantly increasing processing time, and will also increase rates of false positives. The aim here was to obtain a useable high quality result of at least 90% of matches being true positives. Quality checks on the profile of matches in backwards and forwards match runs, by age, marital status, and middle names suggest that this has been achieved, with indications that precision may in fact be more than 95%.

Improvements in literacy and numeracy over time may mean that similar record linkage with later censuses can attract higher recall rates. However, this is confused by varying levels of archival record survival, and regional concentrations of names. As a result geographical success of record linkage is not uniform: Wales fares much worse than England. For Scotland and potentially other countries with similar historical census data, the same method should also be possible, subject to the creation of similar birthplace standardisation coding, and depending on the diversity of the surname and forename set.

Particular care has been given to avoid possible biases arising from record linkage success differentials. Entrepreneurs proved easier to match than the Non-ent general population. In the general population there is considerable variation in record linkage success, with commercial clerks faring much better than sugar refiners, but general labourers are not as difficult to link as expected. There are also contrasts between match rates for farmers and non-farmers. Farmers with large workforces and acreages were slightly more likely to be unambiguously traceable from one Census to the next. Non-farmers are a much more diverse group, and are more difficult to match in forwards links, and especially for 1851 probably as a result of transcription deficiencies, archival data loss, and perhaps in the way the census was administered in 1851. The possible biases are small and can be readily managed in subsequent analysis by handling the data in separate sub-sets for different purposes.

A striking result from analysis of matched individuals is that entrepreneurs were very likely to remain in the same community from one census to the next, with over four-fifths continuing to reside in the same Registration Sub-District. Most entrepreneurs also remained in the same occupational sector. While some of this stability doubtless reflected the age profile of likeable entrepreneurs, which is concentrated in the middle years of adulthood, it is potentially a distinguishing feature, particularly in urban areas that were otherwise experiencing high levels of growth and population turnover. Further work remains to be done on business and occupational stability, but community stability is relatively even across all sectors, and not restricted only to entrepreneurs pursuing occupations associated with high levels of investment in plant and premises, such as manufacturers. The non-entrepreneur sample has been used mainly for comparative purposes. Development of this type of record linkage for the rest of the population should be a priority for future research. The early results reported here evidence the feasibility of linkage for the non-entrepreneur groups (even though match rates are often somewhat lower than for entrepreneurs), and the contrasts between occupational categories suggest various fruitful avenues for future investigation.

The results of the record linkage pilot reported in this paper are the basis for the construction of a database deposit of the linked individual entrepreneurs tracked between each census year. Results from these data are being developed in future publications.

Acknowledgements

Research for BBCE was mainly supported by the ESRC under project grant ES/M010953: **Drivers of Entrepreneurship and Small Businesses**. The Isaac Newton Trust research grant 17.07(d): **Business Employers in 1871** helped support the acquisition and coding of S&N data for 1871, and Isaac Newton Trust Grant 18.40(g)**: Business proprietor succession and firm size change 1851-1881** supported parts of the record-linkage development. Piloting of the research for 1881 draws from Leverhulme Trust grant RG66385: **The long-term evolution of Small and Medium-Sized Enterprises (SMEs)**.

The census database used for 1851-61 and 1881-1911 derives from Higgs, Edward and Schürer, Kevin (University of Essex) (2014) The Integrated Census Microdata (I-CeM) UKDS, SN-7481; http://dx.doi.org/10.5255/UKDA-SN-7481-1 see also E. Higgs, C. Jones,

K. Schürer and A. Wilkinson, *Integrated Census Microdata (I-CeM) Guide*, 2nd ed. (Colchester: Department of History, University of Essex, 2015).

We are also grateful to have been able to make early use of Joe Day's standardised birthplace and location name file for all years. The census database for 1871 derives from S&N to whom we are especially grateful for extraction to meet the same algorithmic rules as for the other years, led by Carry van Lieshout; it was coded by Joe Day. Other coding for 1851-1891 was done by Carry van Lieshout and Harry Smith.

The record linkage and analysis uses RSD boundary files created for the 'Atlas of Victorian Fertility Decline' project (PI: A.M. Reid) with funding from the ESRC (ES/L015463/1), based at Campop; the boundary data were created by Joe Day (2016) *Registration sub-district boundaries for England and Wales 1851-1911*; is a future open access data deposit. The Day RSD dataset has been created using Satchell, A.E.M., Kitson, P.M.K., Newton, G.H., Shaw-Taylor, L., and Wrigley E.A., *1851 England and Wales census parishes, townships and places* (2016) available at:
https://www.campop.geog.cam.ac.uk/research/occupations/datasets/catalogues/documentation/.
The Satchell et al. dataset is an enhanced version of Burton, N, Westwood J., and Carter P., *GIS of the ancient parishes of England and Wales, 1500-1850.* Colchester, Essex: UK Data Archive (May 2004), SN 4828. This is a GIS version of Kain, R.J.P., and Oliver, R.R., *Historic parishes of England and Wales: An electronic map of boundaries before 1850 with a gazetteer and metadata*. Colchester, Essex: UK Data Archive, May, 2001.

References

Anderson, M. 1988. Households, Families and Individuals: Some Preliminary Results from the National Sample from the 1851 Census of Great Britain. *Continuity and Change* 3: 421-38.

Anderson, M., B. Collins, and C. Scott. 1979. National Sample from the 1851 Census of Great Britain. [data collection], UK Data Service, http://doi.org/10.5255/UKDA-SN-1316-1.

Bennett, R.J. and Newton, G. (2015) 'Employers and the 1881 Population Census of England and Wales', *Local Population Studies*, 29-49.

Bennett, R.J., H. Smith, C. van Lieshout, P. Montebruno, and G. Newton. (2019) *The age of entrepreneurship: Business proprietors, self-employment and corporations since 185*1. Abingdon: Routledge. https://doi.org/10.4324/9781315160375

Bennett, Robert J., Smith, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2020), *The British Business Census of Entrepreneurs 1851-1911 (BBCE)* [data collection]. UK Data Service, SN: pending.

Bennett, Robert J., Smith, van Lieshout, Carry, Montebruno, Piero and Newton, Gill (2020), *The British Business Census of Entrepreneurs 1851-1911 (BBCE)*: *User Guide*, https://doi.org/10.17863/CAM.47126

Christen, P., Churches, T. and Hegland , M.(2005) *Febrl - Freely extensible biomedical record linkage*, originally from Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Springer Lecture Notes in Artificial Intelligence, Volume 3056; Release 0.3.1, http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html

Crayen, D. and Baten, J. (2010) ‘Global Trends in Numeracy 1820-1949 and its implications for long-term growth’, *Explorations in Economic History*, 47, 82-89

Day, J (2018) ‘Enriching I-CeM: Matching Individuals Birthplaces to a GIS’, Unpublished Working Paper, University of Cambridge.

Higgs, E., Jones, C., Schürer, K. and Wilkinson, A. (2015) *Integrated Census Microdata, 1851-1911, User Guide version v. 2 (I-CeM.2)*, Second edition, Colchester: Department of History, University of Essex. https://www1.essex.ac.uk/history/research/icem/documentation.html

Lee. R. and Lam, D. (2011) Age distribution adjustments for English censuses, 1821 to 1931, *Population Studies*, 37,3, 445-464.

Montebruno, P, Bennett, R.J., Smith, H. and van Lieshout C. (2020) Machine learning classification of entrepreneurs in British historical census data, *Journal of Information Processing and Management*, 57, 3,https://doi.org/10.1016/j.ipm.2020.102210

Porter E.H. and Winkler, W.E. (1997) *Approximate String Comparison and its Effect on an Advanced Record Linkage System*, Research Report RR97/02, US Bureau of the Census.

Ruggles, S. Fitch, C.A. and Roberts, E (2020) 'Historical Census Record Linkage', *Annual Review of Sociology*, 44, forthcoming 2018. https://doi.org/10.1146/annurev-soc-073117-041447

Sandberg, L. (1979) 'The Case of the Impoverished Sophisticate: Human Capital and Swedish Economic Growth before World War I', *Journal of Economic History*, 39.1, 225-241

Schofield, R. (1973) ‘Dimensions of illiteracy, 1750-1850’, *Explorations in Economic History*, 10.4, 437-455.

Schürer, K., Higgs, E. (2014). Integrated Census Microdata (I-CeM): 1851-1911. [data collection]. UK Data Service. SN: 7481, http://doi.org/10.5255/UKDA-SN-7481-1

van Lieshout, C., Bennett, R., Smith, H.J. and Newton, G. (2017.) 'Identifying businesses and entrepreneurs in the Censuses 1851-1881'. WP 3, https://doi:10.17863/CAM.9640

van Lieshout, C., Bennett, R.J., and Smith, H. (2020) The British Business Census of Employers and firm-size, 1851-1881: new data for economic and business historians, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* https://doi.org/10.1080/01615440.2019.1707140

Winkler, W.E. (1995) ‘Matching and Record Linkage’, in B. G. Cox et al. (eds.) *Business Survey Methods*, New York: Wiley (1995), 355-384.

Winkler. W.E. (2014) Matching and Record Linkage, *WIREs Computational Statistics*, 6, 313–325.

Winkler, W.E. and Thibaudeau, Y. (1991) *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census*, Research Report RR91/09, US Bureau of the Census.

Zhichun Fu, H M Boot, Peter Christen and Jun Zhou: 'Automatic Record Linkage of Individuals and Households in Historical Census Data, *International Journal of Humanities and Arts Computing* 8.2 (2014), 204-225.

Other Working Papers:

Working paper series: ESRC project ES/M010953: *'Drivers of Entrepreneurship and Small Business',* University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure. For updates see: www.bbce.uk

WP 1: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Drivers of Entrepreneurship and Small Businesses: Project overview and database design.* https://doi.org/10.17863/CAM.9508

WP 2: Bennett, Robert J., Smith Harry J. and van Lieshout, Carry (2017) *Employers and the self-employed in the censuses 1851-1911: The census as a source for identifying entrepreneurs, business numbers and size distribution.* https://doi.org/10.17863/CAM.9640

WP 3: van Lieshout, Carry, Bennett, Robert J., Smith, Harry J. and Newton, Gill (2017) *Identifying businesses and entrepreneurs in the Censuses 1851-1881*. https://doi.org/10.17863/CAM.9639

WP 4: Smith, Harry J., Bennett, Robert J., and van Lieshout, Carry (2017) *Extracting entrepreneurs from the Censuses, 1891-1911.* https://doi.org/10.17863/CAM.9638

WP 5: Bennett, Robert J., Smith Harry J., van Lieshout, Carry, and Newton, Gill (2017) *Business sectors, occupations and aggregations of census data 1851-1911.* https://doi.org/10.17863/CAM.9874

WP 6: Smith, Harry J. and Bennett, Robert J. (2017) *Urban-Rural Classification using Census data, 1851-1911.* https://doi.org/10.17863/CAM.15763

WP 7: Smith, Harry, Bennett, Robert J., and Radicic, Dragana (2017) *Classification of towns in 1891 using factor analysis.* https://doi.org/10.17863/CAM.15767

WP 8: Bennett, Robert J., Smith, Harry, and Radicic, Dragana (2017) *Classification of occupations for economically active: Factor analysis of Registration Sub-Districts (RSDs) in 1891.* https://doi.org/10.17863/CAM.15764

WP 9: Bennett, Robert, J., Montebruno, Piero, Smith, Harry, and van Lieshout, Carry (2018) *Reconstructing entrepreneurship and business numbers for censuses 1851-81.* https://doi.org/10.17863/CAM.37738

WP 9.2: Bennett, Robert, J., Montebruno, Piero, Smith, Harry, and van Lieshout, Carry (2019) *Reconstructing business proprietor responses for censuses 1851-81: a tailored logit cut-off method.* https://doi.org/10.17863/CAM.37738

WP 10: Bennett, Robert, J., Smith, Harry and Radicic, Dragana (2018) *Classification of environments of entrepreneurship: Factor analysis of Registration Sub-Districts (RSDs) in 1891.* https://doi.org/10.17863/CAM.26386

WP 11: Montebruno, Piero (2018) *Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911.* https://doi.org/10.17863/CAM.26378

WP 12: van Lieshout, Carry, Day, Joseph, Montebruno, Piero and Bennett Robert J. (2018) *Extraction of data on Entrepreneurs from the 1871 Census to supplement I-CeM.* https://doi.org/10.17863/CAM.27488

WP 13: van Lieshout, Carry, Bennett, Robert J. and Smith Harry (2019) *Extracted data on employers and farmers compared with published tables in the Census General Reports, 1851-1881.* https://doi.org/10.17863/CAM.37165

WP 14: van Lieshout, Carry, Bennett Robert J. and Montebruno, Piero (2019) *Company Directors: Directory and Census record linkage.* https://doi.org/10.17863/CAM.37166

WP 15: Bennett, Robert, J., Montebruno, Piero, Smith, Harry and van Lieshout, Carry (2019) *Entrepreneurial discrete choice: Modelling decisions between self-employment, employer and worker status.* https://doi.org/10.17863/CAM.37312

WP 16: Satchell, M., Bennett, Robert J., Bogart, D. and Shaw-Taylor, L. (2019) *Constructing Parish-level Data and RSD-level Data on Transport Infrastructure in England and Wales 1851-1911.* https://doi.org/10.17863/CAM.37313

WP 17: Satchell, M. and Bennett, Robert J. (2019) *Building a 1911 Historical Land Capacity GIS.* https://doi.org/10.17863/CAM.42285

WP 18: Bennett, Robert, J., Smith, Harry, van Lieshout, Carry and Montebruno, Piero (2019) *Identification of business partnerships in the British population censuses 1851-1911 for BBCE.* https://doi.org/10.17863/CAM.43890

WP 18: Bennett, Robert, J., Smith, Harry, van Lieshout, Carry and Montebruno, Piero (2019) *Identification of business partnerships in the British population censuses 1851-1911 for BBCE.* https://doi.org/10.17863/CAM.43890

WP 19: Montebruno, Piero (2019) *Datasets and guide: downloads for reconstructing British census responses 1851-1881 for the BBCE.* https://doi.org/10.17863/CAM.42285

WP 20: Smith, Harry, van Lieshout, Carry, Montebruno, Piero and Bennett, Robert, J. (2019) *Preparing Scottish census data in I-CeM for the British Business Census of Entrepreneurs (BBCE).* https://doi.org/10.17863/CAM.44963

WP 21: van Lieshout, Carry, Bennett, Robert, J., and Smith, Harry (2019) *Additional codes and people in the British Business Census of Entrepreneurs (BBCE) not available through I-CeM*. https://doi.org/10.17863/CAM.45322

WP 22: Bennett, Robert, J. (2020) *Employers and self-employed in the census 1921-2011 and alignment with BBCE: Entrepreneurs, business numbers and size distribution.* https://www.repository.cam.ac.uk/handle/1810/300054

WP 23: Bennett, Robert, J., van Lieshout, Carry and Schürer, Kevin (2020) *Missing in the Census 1851-1911: The 'lost', 'missing', and 'gaps' in I-CeM and BBCE, with weights to adjust RSD populations*.

WP 24: Newton, Gill and Bennett, Robert J. (2020) *Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM.*

WP 25: Montebruno, Piero and Bennett, Robert J. (2020) *Inter-census record-linked entrepreneurs and non-entrepreneurs 1851-91 using BBCE and I-CeM: database structure, assessment, downloads and User Guide.*

Full list of downloads with all Working Papers available at:
http://www.geog.cam.ac.uk/research/projects/driversofentrepreneurship

and

www.bbce.uk