Inter-census record-linked entrepreneurs and non-entrepreneurs 1851-91 using BBCE and I-CeM: database structure, assessment, downloads and User Guide.

Piero Montebruno and Robert J. Bennett

pfm27@cam.ac.uk   rjb7@cam.ac.uk

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

Inter-census record-linked entrepreneurs and non-entrepreneurs 1851-91 using BBCE and I-CeM: database structure, assessment, downloads and User Guide

Piero Montebruno and Robert J. Bennett

Working Paper 25: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge

1. Introduction

This paper describes a database of record-linked individuals tracing the development of business proprietors using information in the England and Wales population censuses 1851-1891. In addition a sample of non-entrepreneurs is also developed. The database will be available as a download accompanying this working paper. This is an additional database to that available in the *British Business Census of Entrepreneurs* (BBCE) data deposit at UK Data Archive/Service (UKDS), but deriving from that source data.[1] BBCE uses transcripts of the censuses derived from the UKDS data deposit of *The Integrated Census Microdata* (I-CeM),[2] with additional content not available in I-CeM: for additional individuals, coding, and data for 1871 derived from S&N.[3] The BBCE, I-CeM and S&N can be linked through the individual identifiers for each proprietor identified in the censuses. This provides in BBCE cross sections of entrepreneur data for each census year 1851-1911. Record linkage provides a means to track individuals in BBCE from one census to another; for this paper this is developed for 1851-91.

---

[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] 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; see also Higgs, E., Jones, C., Schürer, K. and Wilkinson, A. (2015), *Integrated Census Microdata (I-CeM) Guide*, Colchester: Department of History, University of Essex. Edward Higgs, Christine Jones, Kevin Schürer and Amanda Wilkinson, *Integrated Census Microdata, 1851-1911, User Guide version v. 2 (I-CeM.2),* Second edition, Colchester: Department of History, University of Essex.
The I-CeM version used here and in BBCE derives from an enhanced version of I-CeM v.1 at UKDS. This version was developed by Kevin Schürer, an I-CeM group at Campop, and further enhanced for identifying and correctly coding entrepreneurs by the BBCE team: see *BBCE User Guide* https://doi.org/10.17863/CAM.47126.

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

The analysis reported here, the BBCE data deposit, and the record-linked database, are outcomes from the ESRC-supported project ES/M010953 '*Drivers of Entrepreneurship and Small Businesses'*. This project uses I-CeM as its main source. However, as I-CeM does not identify and code any entrepreneurial elements; BBCE is the source that must be used to identify business proprietors.

The database reported here is derived from a pilot reported in WP 24. It uses the *entrepreneurs* from that linkage analysis as a sample of *non-entrepreneurs*. The paper is restricted to England and Wales and to 1851-91, but the methods can be readily extended to the equivalent data for Scotland and other years available in BBCE and I-CeM. The paper documents the database of record links. Linkage is based on extending one of the mainstream algorithms available for census analysis where there is no 'truth' data available for calibration: the Jaro-Winkler string comparison method, extended to use fuzzy name frequencies, data blocking, and standardisation of variables.

The next section of the paper summarises how entrepreneur record links were made and tagged, and the level of success achieved. Section 3 demonstrates the representativeness (in general) of the entrepreneur record linkage data. Section 4 defines the database structure. Section 5 summarises the non-entrepreneur sample of record linkages. Section 6 gives the variable names. The format of the database and variables follows that for I-CeM and BBCE to facilitate users referring to the original census records. Individuals are identified with the same RecID as used in BBCE and I-CeM, which are identical between these two sources (except for 1871, and a few extra people in BBCE in 1851: see WPs 12 and 21). Users should familiarise themselves with the I-CeM *User Guide*[4] and the BBCE *User Guide*[5] to understand how to use and access these two data sources.

2. Record linkage database construction: entrepreneurs

The record linkage was constructed from two sets of entrepreneur records in the census (see WPs 2, 3 and 4). First, for 1851-81 the data was extracted from a census question on employers and masters:[6]

---

[4] 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

[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

[6] e.g. 'General Instruction', Census of England and Wales, *Householder's Schedule*, 1851; and *Householder's Schedule*, 1851; all bracketed terms as given in original as examples.

'In TRADES the Master is to be distinguished from the Journeyman and Apprentice, thus – "*(Carpenter – Master employing [6] men)*;" inserting always the number of persons of the trade in his employ on March 31st.'

This was extended in 1861 to:

'In TRADES, MANUFACTURES, or other Business, the Employer must, in all cases, be distinguished. – *Example: 'Carpenter – Master, employing 6 men and 2 boys;'* inserting always the number of persons of the trade in their employ, if any, on April 8th [the time of the Census]. In the case of Firms, the number of persons employed should be returned by *one partner* only'.[7]

The data that are the centre of the extraction process are tagged from elements of the phrasing above (see WPs 2 and 3). The method for extraction used in WP 3 is a development of the algorithm from Bennett and Newton (2015). The characteristics of the main data extracted for employers and their workforce is assessed in a recent paper by van Lieshout et al. (2020) which shows the BBCE records match well with the published census tables and, although they under-represent some of the smallest employers compared to the published census tables, they give more larger employers than the clerical process was able to identify in census publications. This confirms the value of the BBCE and I-CeM databases as important resources for record-linkage of entrepreneurs identified from the occupational responses to the census questions summarised above.[8]

The second data source is for 1891 derived from the direct census question of the 'employment status' of individuals and employers, own account or workers.[9] For 1891 entrepreneurs the record linkage data were derived directly from the status question as recorded in I-CeM, but after extensive cleaning. The linkage uses only employers (not those operating on own account) (see WPs 2 and 4).

For 1851-81, entrepreneurs were identified by searching all the census records and 'tagging' the individual employers and masters who responded to the census questions above, and also those

---

[7] 'General Instruction', Census of England and Wales, *Householder's Schedule*, 1851; and *Householder's Schedule*, 1851; all bracketed terms as given in original as examples.
[8] van Lieshout, C., Bennett, R.J. and Smith H. (2020) The British Business Census of Entrepreneurs 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
[9] 'General Instruction', Census of England and Wales, *Householder's Schedule*, 1891

who responded in various other ways that indicated that they were employers or own account by using synonyms of business entrepreneurial status, such as 'proprietor', 'director', 'partner', 'owner', 'landlord/landlady' of an inn, etc. This was achieved by algorithms and subsequent clerical checks and hand cleaning (Bennett and Newton, 2015; see also WP 3). These data are referred to as the 'Extracted' entrepreneurs, in BBCE. They are tagged into groups (variable: GROUP in BBCE) using the highlighted terms in the census question above, as well as other information given in individuals' strings of occupational information (Bennett et al., 2019; see also WPs 9 and 9.2):

> **Group 1:** anyone who stated having employees (tagged <u>employer</u>).
>
> **Group 2**: anyone stating 'employer', 'emp', or similar, but with no employees stated (tagged <u>own account</u> if farmer, treated as non-respondent if non-farmer),
>
> **Group 3:** anyone described as 'master' (cleaned for spurious masters, such as 'station masters' who were employees of railways) (tagged <u>own account</u>),
>
> **Groups 4 and 5:** any farmers not included in the previous categories, Group 4 with no acres stated (tagged <u>worker</u>), Group 5 with acres (tagged <u>employer</u> if acres exceeded a threshold level by Registration Sub-District, otherwise tagged <u>own account</u>: see Montebruno et al., 2020),
>
> **Group 6:** other owners or proprietors of business assets not stating employees: e.g. mine and quarry owners, ship-owners (but not land/housing owners, which were too ambiguous to be used as a business identifier (tagged <u>employer</u>).

The Groups were used as the starting point to identify the target individuals (entrepreneurs) for record linkage in each year. The tagging of status allows status as employer, own account, or worker to be found for those extracted. GROUP is a variable contained in the BBCE database.[10]

The 1851-81 data have non-responses from individuals who were employers and masters but either did not give sufficient information to be identifiable among the extracted Groups (such as statements of 'emp' with no employees), or the transcription of their records was deficient in truncating the relevant information of their employees etc. These non-responding individuals are identified in BBCE by a reconstruction method that supplements the census responses for 1851-81 to infer employment status from individuals' responses to other parts of the census enquiry. This chiefly uses the occupations that people stated, demographic characteristics, and their relationship to the 'head' of the household – who was the respondent who provided the census information (see WP 9 and 9.2; and Montebruno et al. 2020). Two reconstruction methods are given in BBCE:

---

[10] Bennett et al. (2020) *User Guide,* p.17.

EMPSTATUS_NUM and EMPSTATUS_IND.[11] They are compared with machine learning approaches in Montebruno et al. (2020). They are not discussed further here but are key parts of BBCE for assessing and using the record-linked data in subsequent analysis and publications which are crucial for research users.

It is important to recognise that the linkage analysis suffers from a limitation that data for 1871 in England and Wales do not exist in I-CeM. This deficiency was substantially overcome by acquiring census transcripts for BBCE from an alternative supplier: S&N (see WP 12). This provides comparable input information for the 'extracted' population of employers in 1871 to that for the other years, giving the six same Groups of extracted entrepreneurs as listed above; and it includes the variable GROUP in BBCE. However, there is a key exception that most of Group 4 who were not included in BBCE; they exist in S&N but resources did not permit their full extraction. They are mostly workers and so of little consequence for analysis of entrepreneurs. However, more important, the S&N data acquired for BBCE covered only extracted entrepreneur groups, again for resource reasons. This means that searches for linked records from adjacent years cannot search the whole of the 1871 population. Hence, any searches *into* 1871 have an unacceptably high probability of false positives, false negatives, or complete failure of linkage despite an individual being recorded in the census. Thus for the five census years of record linkage, both forward and backward linkage can be applied between 1851 and 1861, and between 1881 and 1891, but only forward links can be fully explored for 1871 into 1881, and only backwards links fully explored for 1871 into 1861. The result is that six linkage exercises were undertaken between the five years 1851-1891, as shown in Table 1.

1851	1861	1871	1881	1891	Linkages N	Match rate
FORWARDS	1851-61 →				72,844	18.9
1861-51 ←	BACKWARDS				96,248	25.8
	1871-61 ←	BACKWARDS			65,205	21.9
		FORWARDS	1871-81 →		65,624	22.0
			FORWARDS	1881-91 →	104,132	25.1
			1891-81 ←	BACKWARDS	217,552	32.4

Table 1. The six linkage processes and match rates (as per cent of the entrepreneur GROUP 1-6 for 1851-81, and all employers for 1891); see also WP 24 (Table 4).

---

[11] Op. cit. ,p.15.

The record linkage pilot was undertaken by searching between years for matched individuals against the whole population available, using forward and backward record linkage, but with only one directional linkage for 1871. The linkages found, as shown in Table 1, were determined using a FEBRL Jaro-Winkler string comparison method modified with fuzzy name matching scores, pre-processing to standardise the data, and with extensive data blocking and checks. Linkage was undertaken between five variables: first and second names as transcribed, modified with fuzzy name matching, and standardised data with exact matching of birthplace, age (by year), and sex.[12] A strong priority was given to minimising false positives, resulting in using a high threshold to accept Jaro-Winkler scores of 0.88 and above, usually 0.95. For 1871, the linkage benefited from the help of Joe Day who applied an equivalent consistent method to code the S&N birthplace and locational data in BBCE. This ensures there is consistent standardisation across the data for the five census years used.

The number of matched records by year summarised in Table 1 displays the constraints on the database resulting from the more restricted information available for 1871, with only forward links possible for 1871-81, and only backwards links possible for 1871-61. However, overall linkage rates are generally comparable and achieved at least a 19 per cent matching for 1851-81 and somewhat higher for 1881-91. The rate achieved is typical for census linkage studies that prioritise minimising false positives, where no accurate training set of 'truth' data is available, and where the strings contain transcription error and variations in name recording (Ruggles et al., 2018).[13] It is lower than could have been potentially achieved because of the high emphasis placed on minimising false positives (Type I errors). The match rates between years are similar, but do vary because of several factors: 1851 was generally less well transcribed which makes it more difficult to obtain high similarity scores, and there is also missing data of about 2% of the 1851 records and about 3.7% of the 1861 records, which have been lost from the archives sources which makes those people unavailable as potential matches (see WP 23). These factors explain why the 1851-61 forwards links achieved a lower match rate than for the later census. In addition, all forwards links have reduced success because a proportion of the population had died in the ten years covered, whereas in backwards links all individuals are alive in both periods. For 1871 there is the limitation noted earlier, that the full population is not available to provide links, which produces

---

[12] The Jaro-Winkler algorithm was derived from the probabilistic record linkage package FEBRL (Freely Extensible Biomedical Record Linkage). 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)
[13] 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)

lower potential match opportunities. However, only a slightly lower match rate was nevertheless achieved for those individuals that had 1871 records available in BBCE, and that rate is almost identical for forwards (1871-81) and backwards (1871-61) linkage.

3. Representativeness of the linked entrepreneur records

A critical test of the validity of record-linked data is whether the linked data is representative of the underlying target population from which it was drawn: is the linked data an unbiased sample of the population used as a starting point. In the record linkage here we are using a target set of individuals we wish to track: the entrepreneur Groups identified in 1851-81, and employers self-nominating in 1891. Hence, the appropriate test is: Were the entrepreneur extraction Groups represented in the linked data with the same frequency as in the original extractions? Record linkage is sometimes criticised because it has a declining coverage of linked individuals as they become older: this is not surprising because there is an attrition rate through death. It is also criticised for over-representing one category over another (e.g. for different sexes). But in terms of the representativeness of the method, the key objective is that those found through linkage are representative of the groups from which they were extracted. Other patterns that reproduce the characteristics of the underlying population can be considered as the subject of the analysis. For example, changes in the level of entrepreneurial status with age is an interesting research focus, but a decline with age has to be separated for artefacts of attrition that less older people are found in forward links because some will have died. Hence, differences in match rates for variables such as age, marital status, location, as identified in WP 24 are not in themselves an issue for bias or validity of the matched data when used for analysis; they are artefacts of entrepreneur behaviour to be investigated in subsequent research.

In general, tests of matches achieved for this record-linkage exercise show that the match rates were representative of the original population of the entrepreneur Groups. Two-way significance tests of 'goodness of fit' of the combined linked samples compared with the original population by extraction Group show no significant differences between the two for almost all sets of linkages, for both farm and non-farm, for the whole sample, and for female entrepreneurs only (see Tables 2 and 3). The null hypothesis of no significant difference of response rates of the original census population by Group was accepted in almost all cases.

Non-farm	1851-61 forwards	1861-51 backwards	1871-61 backwards	1871-81 forwards	d.f.
Chi square	11.08*	1.66	8.12	1.97	5 (4 in 1871)
Farm					
Chi square	1.49	1.11	3.34	1.26	4 (3 in 1871)

Table 2. Test of representativeness that sample of linked from Groups 1-6 are the same as whole population of groups 1-6. All genders (male only have closely similar results).   Farms = Occode 173. Critical values of Chi square at * p>0.05 and ** p>0.1 are, respectively, 11.07 and 15.09 with 5 d.f.; 9.49 and 13.28 with 4 d.f.; 7.82 and 11.34 with 3 d.f.;

Non-farm	1851-61 forwards	1861-51 backwards	1871-61 backwards	1871-81 forwards	d.f.
Chi square	8.29	1.11	14.3**	3.11	5 (4 in 1871)
Farm					
Chi square	1.38	0.73	3.54	0.76	4 (3 in 1871)

Table 3. Test of representativeness that sample of linked from Groups 1-6 are the same as whole population of groups 1-6. Female only, d.f. and p as in Table 2.

However, there are two exceptions: in Table 2 the non-farm 1851-61 forwards links for both sexes, just meet the critical value at p=0.05, and hence can be accepted as on the boundary of being non-biased; and in Table 3 the female backward links 1871-61 are significantly different at p=0.01. The first mainly arises because of the lower rate of matching for Group 3 (masters without employees tagged as own account), which has 28 percent of the sample linked but 36 per cent of the original census Groups, whereas the employer linkage (Group 1) achieves a closer match of 64 percent compared to 59 per cent, respectively. This contrast of higher link rates for Groups 1 than Group 3 is the same in all the linkages, but is only marginally significant at p=0.05 for the 1851-61 non-farm forward links. The second departure from representativeness for non-farm females is more major. This is for the 1871 links that are expected to be problematic. The cause is similar to

the first, there is a higher rate of employer matches (Group 1), and lower rate of own account matches (Group 3). Female proprietresses are more difficult to link than male: only 10% are linked among Group 3 compared to 26% in the Group 1 population. These possible sources of bias can be managed by treating the employer and own account groups separately rather than jointly, and exercising care with any female-only comparisons, or can be re-weighted. The effects also need to be assessed when making interpretations.

A similar test of 'goodness of fit' of the linked samples compared with the original population by firm size shows no significant differences between the two for all combined linkage data (where the forwards and backwards kinks are combined), shown in Table 4. Comparisons of combining the results of forward and backward linkage, rather than looking at them individually, are preferable since record-linkage gives the links as a whole, not the specific steps by which they were constructed, with backward linkages always being more complete because there is no attrition through deaths.

	1851	**1861**	**1871**	**1881**	**d.f.**
All categories	0.31	0.30	0.24	0.62	7
Excluding < 5	0.22	0.18	0.16	0.38	5

Table 4. Test of representativeness that sample of linked are the same as whole population for different firm size categories. Critical values for Chi square at p>0.05 are 14.07 and 11.07 for 7 and 5 d.f., respectively.

For testing, the size categories are aggregated into eight to avoid categories with small numbers, the effect of extremes in the largest size category, and also to leave to analysis the question of how far the very largest firms shifted in size over time. The categories are: 1 employee, 2-4, 5-9, 10-19, 20-49, 50-99, 100-199, and >200 employees. Because of uncertainties about the census coverage of the smallest categories, and consistency between years, the test is repeated for six categories by excluding the smallest two size classes to focus on those with 5 and above, which unlike those under 5 in 1851 are believed to be consistently surveyed each year (see WP 13; Bennett et al., 2019, Chapter 5, and van Lieshout et al. 2020). The null hypothesis of no significant difference of

response rates of the original census population by firm size was accepted in all cases, for both the 8-catgeories and 6-categories. Indeed the differences are very small across all size categories. This is an important finding since it occurs in the face of the potential inconsistency of the census process which leads to differences in the frequency of firm sizes detected in the 1851 census compared to subsequent years. The 1851 census administrators appear to have followed a slightly different process of collecting data for the smallest firm-sizes classes which resulted in the smallest from being surveyed more thoroughly in the 1851. This is particularly true for those with only one employee, and probably also affects those with up to 3 or 4 employees. This is believed to be the result of the census including in the employee counts more in-household assistants and probably more family members in 1851, although this does not fully explain the 1851 differences observed (see Bennett et al., 2019, Chapter 5; van Lieshout et al., 2020). The result is increased size of the Ent population in 1851 used for the forward search, by including a higher proportion of individuals from the smallest enterprises (see Table 8 in WP 24). The tests in Table 4 demonstrate that despite fluctuations in the magnitude of the business population between years, the linkage process appears robust and unbiased by firm size.

Overall, the main linkage data are robust for most cases as representative of the underlying population. The population of farmers was expected to be slightly easier to link as they were often more stable with less mobility out of entrepreneurial status, but actually linkage rates are slightly lower for farm employers and masters than non-farm. Similarly, female entrepreneurs were expected to be difficult to link; but although they do have slightly lower link rates than male farmers, for non-farm females the link rates are higher and reach 70 per cent in some years. It is also important to note that despite the difficulties of linkage for 1871, all samples except female 1871-61 backwards links are statistically representative of the original population across the Groups. Possible sources of bias that arise can be managed by treating the employer and own account groups separately rather than jointly, and exercising care with any female-only comparisons. Similarly, the higher frequency of the smaller firm categories in 1851 do not affect linkage rates, but must be handled appropriately when comparing years in subsequent data analysis.

4. Database structure

The original data links shown in Table 1 can be linked pairwise between all years to give a continuous linkage for anyone found in each successive census, as shown in Table 5. Because only

some individuals are linked between the different pairwise exercises, the total number of the same linked individuals is reduced by the number that appear in two or more years. For example, between 1851 and 1861 there are 72,844 forward linkages (1851-61), and 96,248 backwards links (1861-51): a total of 169,092. But of these, only 131,216 are different people. The remaining 37,876, or 28.9% are duplicates who were the same people identified both forwards and backwards. The only other two-year direct comparisons of forwards and backwards links, between 1881 and 1891, had a 19.3% duplicate rate.

Pairwise link	N Obs
1851-61	72,844
1861-51	96,248
1871-61	65,205
1871-81	65,624
1881-91	104,132
1891-81	217,552

Table 5. Pairwise links between years for each individual linkage.

The links shown in Table 5 are coded into the database structure by organising them into the non-duplicate pairs between each year, with the continuing links of the same individuals for the preceding and ensuing years (as relevant).

The N for each individual link that can be tracked between years is shown in Table 6. Thus for the first section for linked individuals between 1851 and 1861, there are 131,216 links between the same individuals in those two years. Of these, 17,106 of the same individuals have been also found as linked in 1871, 7,945 in 1881, and 2,832 in 1891. The links are pairwise between the years, so that looking back from 1881-91 links at the foot of the table where there are 269,566 individuals linked 1881-91, and the 2,832 linked back as far as 1851 (which are the same people as linked forwards form 1851 to 1891). The N of individuals linked between the other years is given in the different sections of the Table 6. The full panel of individuals that can be linked between all five census years is 2,832, but for 2- or 3- or 4-year census links the N is larger. In Table 6 'Rxx' refers to the number of links for year xx in question; NewID71 is used to remind that this year is constrained by the data for potential links that can be identified.

Links 1851-61 and continuing	*Variable*	*N Obs*
	R51	131,216
	R61	131,216
	NewID71	17,106
	R81	7,945
	R91	2,832
Links 1861-71 and continuing	*Variable*	*N Obs*
	R51	17,106
	R61	65,205
	NewID71	65,205
	R81	27,984
	R91	9,703
Links 1871-81 and continuing	*Variable*	*N Obs*
	R51	7,945
	R61	27,984
	NewID71	65,624
	R81	65,624
	R91	19,005
Links 1881-91 and continuing	*Variable*	*N Obs*
	R51	2,832
	R61	9,703
	NewID71	19,005
	R81	269,566
	R91	269,566

Table 5. Number of linked individuals between years for each linked year.

The N available in any combined linked year in the database has the total N for the pairwise links for their reference years, but also has other links for the same people in the other years. The total N for each set of links is shown in Table 7.

Linked year	*Variable name*	*N Obs*
1851-61-51	R51	131,216
1871-61	R61	65,205
1871-81	NewID71	65,624
1881-91-81	R91	269,566

Table 7. Record linked database: N for each year

5. Record linkage database construction: non-entrepreneur sample

The main purpose of record linkage developed from BBCE and the ESRC '*Drivers of Entrepreneurship and Small Businesses*' project was to track entrepreneurs over time. However, it is valuable to have a comparative sample of non-entrepreneurs. For this purpose a sample of non-entrepreneurs was developed that provided contrasts of sector of operations and type of occupation based on a stratified random quota design, with quotas within two strata: for sectors and geography. Twenty sectors with contrasted were identified that were likely to have different levels of switching and mobility between occupations, level of skill specialisation, gender participation, histories of development. Selection was based on I-CeM Occodes, after cleaning. Five counties were chosen as the geographical sampling frame to cover a range of different levels of urbanisation and relation to the selected sectors, as well as having sufficient completeness of records in the scanned e-censuses (see WP 23): Bedfordshire, Durham, London, Oxfordshire, and Warwickshire. The final sample 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**

Table 8. Stratified quota random sample for non-entrepreneurs in 21 sectors in 5 counties.

For most occupations the population available in the five counties was sufficiently small to give the constrained sample needed. But for two occupations (commercial clerks, and general labourers) the numbers were large so that a quota was drawn by random selection limited to 10,000. Hence, the final sample is a mix of a stratified quota full population and random selection. Because of the limitations of working with the 1871 e-records available, no attempt was made to develop record linkage related to this year. As a result, only two non-entrepreneur linkage databases were constructed: 1851-61, and 1881-91. As shown in Table 9.

1851	1861	1871	1881	1891	Linkages N	Match rate
FORWARDS	1851-61 →				15,142	10.9
1861-51 ←	BACKWARDS				21,586	16.3
			FORWARDS	1881-91 →	28,551	17.5
			1891-81 ←	BACKWARDS	40,325	25.3

Table 9. Linkages and match rates (as per cent of the non-entrepreneur sample selected).

The different linkage exercises were then combined to give one set of matches for the two groups of years for the non-entrepreneurs in the same way as for the linked entrepreneurs; see Table 10. It is this set of combined links that available as a download.

1851	1861	1871	1881	1891	Linkages N	Match rate
Combined 1851-61 / 1861-51					32,174	14.1
			Combined 1881-91 / 1891-81		61,167	19.3

Table 10. Combined match numbers for non-entrepreneur sample.

6. Record-linked database: variable names

The variable list derives from the RecID in the I-CeM data with additions from BBCE (see I-CeM User Guide; and BBCE User Guide).[14] The variables are listed below with a brief definition of each. The record-linked data are for England and Wales only. A RecID entry for a given year means that individual has been matched to what is believed to be the same person in another year, with the entry in that other year giving their RecID then. The numbers of years with entries is the number of years for which a link has been found. A non-entry does not necessarily mean a failed match; the individual could also be dead in that year or absent from the country. Age is given as a check so link files to I-CeM and BBCE.

6.1 Entrepreneurs record-linked database download

The download is in CSV format as a txt file. The files can be opened by any text editor, database management system (Access) or statistical package (Stata, SPSS)

Variable list:

Age51
Age61
Age71
Age81
Age91
RecID51
RecID61
NewID71
RecID81
RecID91

Variable: Age xx [year date]

Variable label: Age

Type: Numeric

Max length: 4

---

[14] 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
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

Description: Age as given from I-CeM or BBCE: provided solely as a means for secondary checking of data matching with I-CeM and BBCE. For 1871, like other variables in BBCE this is derived from S&N transcriptions. The label Agexx is used to distinguish from the label xxAGE in I-CeM and BBCE.

Coverage: 1851-1881 England and Wales

Variable: RecID xx [year date]

Variable label: RecID   Unique person identifier

Note: same as in I-CeM except for additional people identified in 1851, and all 1871

Type: Numeric

Max length: 8

Description: Unique identifier for each person's record, for each year. Key variable to link with I-CeM 1851-61 and 1881, and to BBCE for 1871 and additional people for 1851. RecID is identical to I-CeM and BBCE, with additional people not in I-CeM given high numbers not used by I-CeM; and the 1871 RecID unique to BBCE which does not relate to any future I-CeM deposits for 1871.

Coverage: 1851-1881 England and Wales

The RecID allows users to follow a given person who has been linked from one year to the next. Each data row gives a linked person, with the columns giving the RecID for the year that the same person has been found as linked.

6.2 Non-entrepreneur record-linked database download

The download of the linked non-entrepreneur sample is in the same format as the entrepreneurs: CSV format as a txt file, with the same variable definitions. Because 1871 was not available, a RecID entry for a given year is linked only between two year pairings: 1851/61, and 1881/91. A single entry means that they were linked only in one direction; if in 1851, from that to 1861; in 1861, from that to 1851. An entry in both years means they were linked both backwards and forwards. Age is again given as a check so link files to I-CeM and BBCE. The variables given are:

Variable List:

      Age_51

      Age_61

      RecID51

      RecID61

      Age_81

      Age_91

      RecID81

      RecID91

7. Conclusion

This paper provides users with the background to the database of record-linked individual business proprietors tracked between England and Wales population censuses for 1851-1891, and also a sample of non-entrepreneurs in 21 sectors for five counties that can be used for comparison purposes. The paper gives the definitions of variables as in the record-linked database. The method used to construct the record linkages cover only for England and Wales. However, the methods can be readily extended to the equivalent data for the whole population and to Scotland for all years available in BBCE/I-CeM.

The paper assesses the representativeness of the entrepreneur record linkage data against the target groups, demonstrating that in general it is representative of the underlying population from which it is drawn of the extraction Groups used to identify entrepreneurs. Tests of the representativeness of the links for each extraction Group in the linked data show that they occur with the statistically the same frequency as in the original extractions. Similarly, tests of representativeness by firm size confirm that the linked data are generally fully presentative. Hence, in general, there is no greater chance of being found in the linked data than the original data by extraction Group or firm size. However, there are *two exceptions* where the linked data have an inbuilt bias:

(i) For non-farm proprietor 1851-61 forwards links, where the null hypothesis of no difference between the record-linked and original data for Groups is not quite accepted at the 95% level.

(ii) For female backward links 1871-61 where there is a significant difference of the record-linked from the underlying Group population at the 99% level.

The first is overcome by combining the data for forward and backward linkage – as given in the record-linked database; and the effects of the second is greatly reduced by combining linkages. However, in any case, these possible sources of bias can be managed by treating the employer and own account groups separately rather than jointly, exercising care with any female-only comparisons, and by combining forward and backwards links where available – as given in the record-linked database.

In general a significant finding of this paper is the importance *combining* matching data for forwards and backwards linkages. For example, a difference in match rates that occurs for forward matching 1851-61 for non-farm firm size (as noted in WP 24) is not significant when the combined linkage results with backwards linkage 1861-51 are used. This demonstrates the importance of using the combined database of forward and backwards links provided here.

The entrepreneur linkage data are robust by firm size despite the much higher numbers of the smaller firm sizes recorded by census administrators in 1851; thus linkage rates appear to be independent of the peculiarities of the census process. In future linkage research it will be valuable to know how far, in general, linkage is independent of different census processes, but the results for firm size, as well as that for extraction Group, is suggestive that linkage methods are generally independent of the census process. This suggests that, despite other possible administrative changes, the census processes were consistent in recording the variables that are critical to linkage (surnames, forenames, birthplaces, ages and sex – the last three after standardisation). This is indeed to be expected given the overwhelming emphasis placed by the census on these variables in order to accurately count the population. Hence, the record linked data have placed greatest emphasis on the most consistent parts of the data.

The database and variables for the record linked data are given in databases linked to this WP. Their format follows that for I-CeM and BBCE to facilitate users referring to the original census records. Individuals are identified with the same RecID as used in BBCE and I-CeM, which are identical between these two sources. Users should familiarise themselves with the I-CeM *User*

Guide[15] and the BBCE *User Guide*[16] to understand how to use the record-linked database and to enlarge the range of variables available by using two data sources. Updates will be provided at www.bbce.uk.

The database of linked individual entrepreneurs tracked between each census year will provide an important resource for other researchers to begin to make large scale investigations of links between people's life development and career history in business during the nineteenth century.

Future research should focus on expanding the record-linkage material available from the censuses. This can be usefully expanded to the whole non-entrepreneur population, to a range of other linkage methods, including machine learning (as an extension of Montebruno et al. , 2020), and to comparisons with more detailed clerical identification of records from genealogical and other sources where checks of algorithmic material would be enhanced by use of a wide range of external data. Whilst full 'truth' data will never be available for the individuals in these historical censuses, there is a wide range of other external data that can be used to test record linkage results for some of the individuals. For entrepreneurs this range of possibilities is being explored in future papers by the current authors and collaborators.

The algorithmic approach to record linkage only really works for large data where it is possible to accept an error rate that is below the significance levels that are being used for the rest of analysis; then researchers can concentrate on the overall statistics and interpretations that can be made. However, if individual cases are being pursued for individual business histories of specific firms the extensive external sources that are often available for firms from archival sources should be used. There are horses for courses, with different methods useful for different purposes. The record-linkage method used here, and BBCE as a whole, is most valuable not for the individual cases but as a statistical source for large-scale analysis of many businesses in the whole population. It is also an important finding aid for more individual study, as well as giving a broader basis for comparisons and hence helps to assess the generality and representativeness of individual case studies.

---

[15] 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

[16] 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

Acknowledgements

References

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., Smith, H., van Lieshout, C., Montebruno, P. and Newton, G. (2019) *The Age of Entrepreneurship: Business Proprietors, Self-employment and Corporations Since 1851*, Abingdon: Routledge https://doi.org/10.4324/9781315160375

Bennett, R. J., Smith, H., van Lieshout, C., Montebruno, P. and Newton, G. (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 (accessed 2018)

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

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

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

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.J. and Smith H. (2020) The British Business Census of Entrepreneurs 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

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