Identifying businesses and entrepreneurs in the Censuses 1851-1881.

Carry van Lieshout, Robert J. Bennett, Harry Smith, and Gill Newton

cv313@cam.ac.uk rjb7@cam.ac.uk hjs57@cam.ac.uk ghn22@cam.ac.uk

Working Paper 3:
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.

May 2017

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

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

JEL Codes: L26, L25, D13, D22

Identifying businesses and entrepreneurs in the Censuses 1851-1881

Carry van Lieshout, Robert J. Bennett, Harry Smith and Gill Newton

Working Paper 3: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge.

1. Introduction

The household returns from the censuses between 1851 and 1881 provide a major opportunity to identify entrepreneurs in the nineteenth century, as they contain information on the number of employees of each business and the acres of each farm. These censuses each followed a broadly similar structure and are the earliest large scale and most complete source of information on business size, because of the information that employers provided on the number of workers in their employ. While the early censuses are less definitive for identifying own account business proprietors than the later censuses, many individuals can be confidently identified, whilst the population of all occupied can be used to examine workers and others who were likely to have been self-employed. This working paper describes the early censuses and the method through which different groups of entrepreneurs were extracted and parsed. It distinguishes between three major groups: employers with employees; own account proprietors who can be identified with certainty; and people owning certain business assets that indicate they were very likely to be either employers or were self-employed working on own account.

This paper explains how the target individuals can be identified and extracted from the original manuscript records of household returns to the population census for 1851-1881. The database for Entrepreneurs 1851-1911 referred to in this and other project Working Papers for ESRC project ES/M010953 *Drivers of Entrepreneurship and Small Businesses* is an amalgamation of several sources. The data referred to in this working paper for 1851-1881 is

in part derived from the Integrated Census Microdata (I-CeM) deposited at UK Data Archive (UKDA), which has been used in a revised and updated form to improve its accuracy of coding.[1] The I-CeM records are derived from the transcriptions made by the commercial genealogy provider Find My Past (FMP) (part of BrightSolid) in conjunction with The National Archive (TNA). However, because of major gaps in the required records for employers and other entrepreneurs in FMP and hence in I-CeM, additional data have been extracted from other sources to obtain the full records required to satisfy the target of a complete and consistent database. The additional material is derived from three sources: first, for 1851 approximately 55,000 additional records missing or truncated in FMP and I-CeM have been supplied by S&N Genealogy Supplies (TheGenealogist.co.uk); second, for 1861 approximately 28,000 records of entrepreneurs truncated in FMP and I-CeM have been obtained by direct inspection of the original census manuscript pages; third, for 1881 the earlier version of records transcribed by the Genealogy Society of Utah and deposited at UKDA, with corrections made by Campop, has been preferred to the version of these records available in I-CeM.[2]

This paper describes how individual entrepreneurs can be identified and extracted from the original Census Enumerators Books (CEBs). It must be borne in mind that the population census was not a business census but designed by the General Register Office (GRO) to count the population. As a result, the way in which the census gathered material constrains the employer information that can be obtained. The material collected in each census and its value for identification of entrepreneurs as a raw data base and published tables are described in Working Paper 2 from the ESRC project: *Employers and the self-employed in the censuses 1851-1911: The census as a source for identifying entrepreneurs, business numbers and size distribution*. An overview of the project is given in Working Paper 1: *Drivers of Entrepreneurship and Small Businesses: Project overview and database documentation*.

[1] Higgs, Edward and Schürer, Kevin (University of Essex) (2014) *The Integrated Census Microdata (I-CeM)* UKDA, SN-7481, derived by FindMyPast using a variety of original FMP transcriptions. Version 2 of I-CeM includes a range of valuable additional inputs from colleagues at Campop; see Schürer, K., Higgs, E., Reid, A.M., Garrett, E.M. (2016) *Integrated Census Microdata V.2 (I-CeM.2).*

[2] Schürer, Kevin and Woollard, Matthew (University of Essex) (2000) *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)* [computer file] UKDA, SN-4177, transcribed by Genealogical Society of Utah and Federation of Family History Societies. Although the two versions for 1881 are nominally the same, the version used includes many corrections and updates to occupational and other codes made at Campop.

1. The census form and reports

In the censuses 1851-1881, enumerators were asked to record employee numbers for occupational groups. As discussed in Working Paper 2 from, the instructions for 1851 requested that for "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". In addition, farmers were asked to provide "the number of acres, and of in and out-door labourers".[3] In 1861 the instruction for trades remained the same, but the example provided read "*Carpenter – Master, employing 6 men and 2 boys*" while the farmers were explicitly asked for "men and boys employed on the farm".[4] By 1871 the instruction for the farmers included women, now reading "number of men, women, and boys employed on the farm", although the example still only mentioned men and boys. The instructions for employers in trades referred to "workpeople".[5] Both instructions remained the same for 1881.[6] While the instructions only asked farmers and people in trades to return employees, in practice many others, such as merchants, landowners, and mine owners did the same. In addition, although the instructions only asked for limited categories (men and boys, later women, and only labourers for the 1851 farmers), in practice many other descriptors were used for employees as well, including apprentices, girls, porters, or masons; furthermore, the workforce was often broken down by gender.

The 1851 census was the earliest that explicitly sought to differentiate employers from others, and masters from men. It was also the only census to actually publish the results of the census enquiry of employers over the 1851-81 period. However, the GRO was very critical of the quality of the returns of masters, which were considered imperfect, as not all masters had identified themselves as such, and it was suspected that some of the masters had not returned their employees.[7] As a result, the report concluded that the tables created based on the return of workpeople by masters and farmers was "tentative, and a mere auxiliary to our inquiry."[8] The results of the farmers' returns were likewise viewed with some suspicion; however, as the results were considered "of so much interest on a matter so imperfectly understood", separate division and county tables were published showing the size of farms and the number of

[3] General Instruction, Census of the Population, 1851.
[4] General Instruction, Census of England and Wales, 1861.
[5] General Instruction, Census of England and Wales, 1871.
[6] General Instruction, Census of England and Wales, 1881.
[7] Census of Great Britain, 1851. Population Tables II. Ages, civil condition, occupations, and birth-place of the people Vol 1; lxxviii.
[8] Census of Great Britain, 1851. Population Tables Vol 1; lxxviii.

labourers employed on them, in addition to the general tabulation of employees.[9]

The 1851 report is valuable as the only occasion over 1851-81 for which the GRO tabulated employers and the number of men in their employ. This provides two valuable elements: first, it provides insight into how GRO interpreted the responses to this census question; and second, it provides a comparator and check on the methods of extraction which have been used here drawing from the original CEBs. The GRO tables were based only on those returns which conformed to the exact formulation of the census question. First, it only included people explicitly identified as masters, excluding people who just stated their number of employees. The reason provided was that they might be journeymen – even though it was recognised that in certain trades this disqualified almost all employers.[10] Secondly, it only counted men – even though the question in the census had mentioned persons – while employed women and children were sometimes quoted separately at division level, this was only for a few of the branches of trade of firms where women and children were employed in significant numbers.[11] Women and children employed on farms were excluded from the county level tables of farmers' labourers as well.

In 1861 the census gathered information not only on workpeople but broke this down further into men, women, and children, and the instructions to householders and enumerators were amended to reflect this.[12] However, no tables were prepared on the workforce of tradespeople, and only limited analysis was performed on the farmers' returns, as it was anticipated that an agricultural census would be introduced in the near future.[13] The farm tabulation of returned acres and workforce again excluded women and children, and was only aggregated for 10 counties.[14] By 1871 any attempt to tabulate by GRO was abandoned, and there was no mention of the returned employees of tradespeople in the GRO published report. However, labourers returned by farmers were tabulated for 17 counties and the results shown against the 1851 figures.[15] The report was more optimistic about the quality of the employee numbers returned, stating that "the return of the farmers is quite in accordance with the return of the labourers themselves".[16]

[9] Census of Great Britain, 1851. Population Tables Vol 1; lxxviii.
[10] Census of Great Britain, 1851. Population Tables Vol 1; cclxxvi.
[11] Census of Great Britain, 1851. Population Tables Vol 1; 119.
[12] Census of England and Wales, 1861. Vol. III General Report, 27.
[13] Census of England and Wales, 1861. Vol. III General Report, 29.
[14] Census of England and Wales, 1861. Vol. III General Report, 139-143.
[15] Census of England and Wales, 1871. Vol. IV General Report, xlvi-xlvii.
[16] Census of England and Wales, 1871. Vol. IV General Report, xlviii.

By 1881 the GRO explicitly abandoned any attempt at published tabulations, describing the analysis of the occupations as "the most laborious, the most costly, and, after all, perhaps the least satisfactory part of the Census".[17] No tables of either the employees of tradespeople or farmers' labourers were made, and the recording of employees and labourers was abandoned in the design of the 1891 census (see Working Paper 2).

2. Overcoming deficiencies in FMP/I-CeM

Various checks have been undertaken with all the data extraction at different stages to assess how far they provide complete coverage and represent as far as possible: (i) a full extraction of everything that is in the archival census records, and (ii) how far they include all anticipated business so that a reliable and consistent count of the total British business population can be achieved. These checks are reported in various parts of the analysis of the records as they arise. In this working paper the key issue is completeness of the archival census records that are captured by I-CeM; subsequent working papers address the problem of the completeness and representativeness of the coverage of the whole business population There are three major deficiencies in the data supplied by FMP used in I-CeM that affect the database that can be constructed for employers: first, for 1851 approximately 55,000 employer records are missing or truncated in I-CeM; second, for 1861 approximately 28,000 employer records are truncated in I-CeM; third, for 1871 there are no records in I-CeM for England and Wales and, though a form of them is being constructed to enhance I-CeM, this does not contain any occupational information and hence prevents employers being identified. It has been possible to overcome each of these deficiencies, as discussed below.

2.1. The 1851 Census

For 1851 one of the obvious checks to undertake was to compare the data extracted from I-CeM with the published GRO report, which for 1851 alone published an analysis of the records of employers. After the initial stages of data extraction, following the various steps outlined in section 3 of this paper, it was discovered that there was a serious deficiency in the number of employers that can be identified using I-CeM. A comparison of the non-farmers at

[17] Census of England and Wales, 1881. Vol. IV General Report, 25

Division level, which was the lowest geographical level published by the GRO, is shown in Tables 1a and 1b. This makes clear that there is a total deficit of about 23,000 employers in I-CeM compared to the GRO report, a number that is far too large, and clearly not randomly distributed, to be accounted for by transcription or keying errors. Even more worrying the deficits are heavily concentrated in three divisions (North West, London and South East) that are the most populated and contain many, if not most, of the major employers in the whole UK. To satisfy the aims of the project to achieve a consistent identification of all the main businesses in the country, some method had to be found to overcome the deficiency. The explanation for these gaps appears to be that the varied sources used by FMP for the 1851 census were working at different levels of transcription. The 1851 transcription for FMP derived from two sources: (i) family history society existing transcriptions, which often stopped transcribing after the main occupation had been captured, and did not bother with details of employee numbers etc.; and (ii) FMP's own transcriptions. For FMP's own transcription it appears that there was a combination of deficiencies: some areas were missed entirely, some individuals had truncated lines because they were never fully keyed (in the same way as family history societies), and some data were truncated and lost at some stage in the transfer of databases between IT systems, with the occupation text-string cut off at 100 characters.

Fortunately, as well as FMP, there are a number of other genealogy suppliers that provide data on the census. One frequently used alternative, Ancestry, however, did not have full occupation strings either. But the deficit can be made good by using S&N Genealogy Supplies. S&N have undertaken an entirely independent transcription of all the censuses derived from the original TNA microfilms and this covers almost all of the deficiencies in FMP/I-CeM. To narrow the problem down a further analysis of the I-CeM data was undertaken to identify which parts of the country had wholly or partially deficient data. This was done at Registration Sub-District (RSD) level, analysing which RSDs contained either no employer returns at all, which is implausible at RSD level, or only a very low percentage of the population. The resulting analysis is shown in Figure 1. Clearly for missing data there is uncertainty about whether the data is actually missing, or never existed in the first place. By using the RSD level minor differences of enumerator capacity at ED or parish level will be smoothed out, leaving low-employer RSDs as likely locations where all or part of the employer information has been missed in I-CeM. Note that at this stage Scotland has not been tested for completeness in I-CeM.

	I. LONDON		II. SOUTH-EASTERN		III. SOUTH-MIDLAND		IV. EASTERN		V. SOUTH-WESTERN		VI. SOUTHWEST-MIDLAND	
NrEmployees	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**
1	743	3182	2093	2664	1637	1746	1480	2331	2371	2115	2444	2536
2	644	3092	1725	2190	1172	1256	1100	1638	1982	1837	1961	2026
3	336	1922	969	1219	643	698	563	829	1116	1124	1138	1235
4	237	1338	684	774	443	447	402	513	768	710	834	844
5	111	710	352	387	239	233	194	272	406	363	484	508
6	115	729	328	392	245	260	198	245	387	349	447	479
7	71	329	189	203	133	132	105	136	179	161	213	243
8	59	322	162	159	103	107	94	98	197	160	241	271
9	32	183	115	99	72	56	53	67	127	92	148	145
10-	174	985	549	484	381	299	281	288	555	462	756	935
20-	60	416	148	122	126	79	86	63	139	117	225	283
30-	36	183	65	58	56	48	31	20	49	51	119	62
40-	17	121	39	23	28	28	13	10	29	23	73	42
50-	26	100	48	35	31	30	19	27	37	26	122	113
75-	8	37	24	11	14	10	9	7	21	15	64	53
100-	7	39	19	12	12	9	6	8	14	15	59	69
150-	4	14	4	4	9	11	2	3	8	13	41	26
200-	1	10	1	3	1	2	2	0	10	3	22	17
250-	1	5	5	1	2	0	1	1	3	2	13	13
300-	1	5	4	2	1	3	2	0	2	1	16	8
350-	4	7	8	2	9	4	4	5	21	9	37	28
Grand Total	**2687**	**13729**	**7531**	**8844**	**5357**	**5458**	**4645**	**6561**	**8421**	**7648**	**9457**	**9936**

Table 1a. Comparison of preliminary extraction of employers identified in I-CeM and the GRO report of 1851 Divs. I-VI. Non-farmers only.

	VII. NORTH-MIDLAND		VIII. NORTH-WESTERN		IX. YORKSHIRE		X. NORTHERN COUNTIES		XI. WALES	
NrEmployees	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**	**I-CeM**	**Published**
1	1927	1899	453	3216	2582	2079	1189	1377	990	1200
2	1332	1246	342	2682	1993	1249	977	1125	658	902
3	719	660	156	1626	1226	725	631	693	362	446
4	524	513	123	1264	891	513	384	409	235	299
5	289	299	74	676	490	265	207	233	116	162
6	306	260	70	769	544	325	220	213	125	157
7	161	151	46	425	276	132	101	123	49	58
8	118	140	48	455	260	135	96	118	57	79
9	75	81	25	222	166	86	61	71	41	38
10-	440	390	113	973	972	451	371	354	175	205
20-	137	122	43	388	344	183	96	101	45	55
30-	79	57	29	212	210	104	51	60	18	23
40-	41	35	13	145	119	54	30	22	12	11
50-	52	48	21	177	179	72	48	37	12	16
75-	33	28	11	143	96	26	16	13	5	5
100-	32	28	11	162	124	33	21	9	5	6
150-	18	14	12	106	51	24	15	20	4	1
200-	12	8	3	65	38	12	9	12	2	3
250-	5	3	4	50	25	10	3	3	1	0
300-	6	5	1	36	22	5	0	0	0	0
350-	15	11	8	141	96	18	5	3	0	0
Grand Total	**6321**	**5998**	**1606**	**13933**	**10704**	**6501**	**4531**	**4996**	**2912**	**3666**

Table 1b. Comparison of preliminary extraction of employers identified in I-CeM and the GRO report of 1851 Divs. VII-XI. Non-farmers only.

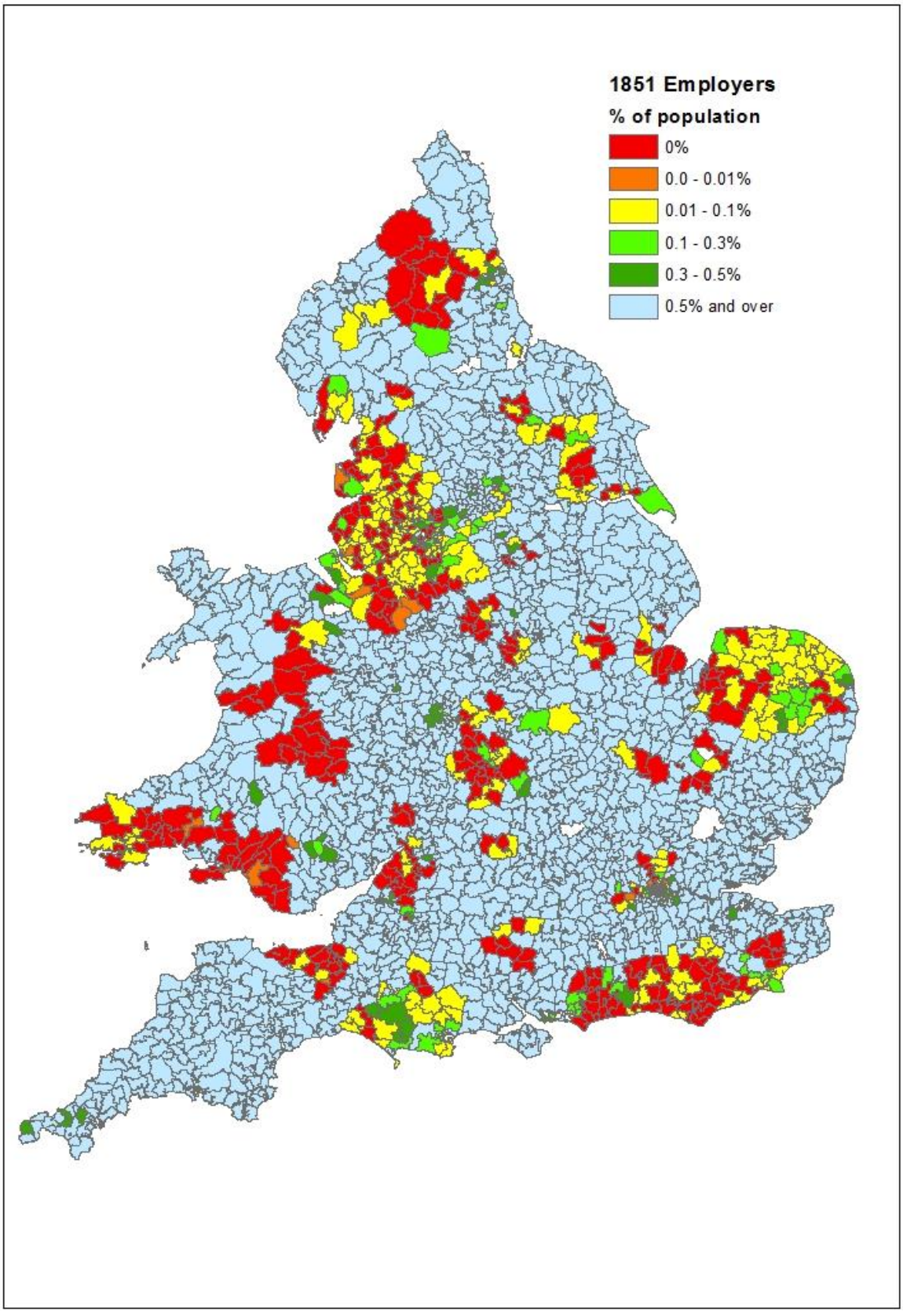

Figure 1. The ratio of employers to the total population at RSD level 1851 from I-CeM records.

Figure 1 was used as a starting point for supplementation of 1851 employer data. All areas with no employers were judged as almost certainly truncated, areas with less than 0.1% employers were judged as very likely to have been truncated, and those with 0.1-0.3% employers as possibly truncated. Areas with 0.3-0.5% employers were also candidates for truncations and were checked. S&N supplied records for all the areas on this map with possible truncation: the whole of the 14 counties where records are wholly absent or very partial in I-CeM (Carmarthenshire, Cheshire, Derbyshire, Dorset, Glamorgan, Gloucestershire, Lancashire, London, Middlesex, Montgomeryshire, Norfolk, Radnorshire, Sussex, and Warwickshire), and the 87 Registration Districts within 27 other counties that otherwise appeared fairly complete, but where records were wholly absent or very partial at RSD level in I-CeM. These counties were Berkshire, Brecknockshire, Cambridgeshire, Cornwall, Cumberland, Denbighshire, Devon, Durham, Essex, Flintshire, Hampshire, Huntingdonshire, Kent, Leicestershire, Lincolnshire, Merionethshire, Monmouthshire, Northumberland, Oxfordshire, Pembrokeshire, Shropshire, Somerset, Staffordshire, Surrey, Wiltshire, Worcestershire, and Yorkshire. In total, S&N supplied 75,000 records.

The S&N and I-CeM records were then compared against each other at the level of individuals, and where gaps were found the S&N records were input. This process is far from simple because the parish and other identifiers in S&N differ from those in I-CeM so a considerable resource had to be deployed to ensure accurate matching. While up to 80% of records could be linked using automated matching, the remainder had to be checked at an individual level to check that the correct match had been found.

There are several other complexities. Most records of individuals identified in S&N have counterpart records in I-CeM. However, about 1570 records in S&N do not have records in I-CeM at all. This is either because of transcriber or other keying errors that prevent the matched individuals being found (about 450 cases; 29%), or because of totally omitted parishes, usually in batches (about 1,120 cases; 71%). These S&N records with no I-CeM counterpart were added to the entrepreneur database manually, coded to the same format as I-CeM, but given additional IDs. Conversely, there are some records in I-CeM for areas which otherwise appear to have deficiencies in employer entries, which do not have a counterpart in S&N. These I-CeM records were retained in the entrepreneur database. In cases where I-CeM already contained the full string, and the S&N match was a duplicate with slight variance in transcription, the I-CeM transcription has been retained to maintain greater consistency.

The result of this infill was an additional 55,000 individual employer records identified that would otherwise be absent from the entrepreneurs database. This is approximately 25% of the total employers identifiable for this year, and covers some of the most significant locations and industries in Britain.

2.2. The 1861 Census

The data supplied by FMP and obtainable from I-CeM truncate the occupational string field at 50 characters for 1861. This information was not contained in the original I-CeM documentation and required significant resource to unravel. Checks directly with FMP confirmed that these data were either never fully keyed, or more likely were lost at some stage in the transfer of databases between IT systems. Many strings for employers, especially the largest and most complex businesses, have considerably longer than 50 characters, and these are some of the most significant employers that are required for analysis in this project. I-CeM was checked to identify all potentially truncated strings based on length of string, string ending in '###' (which was a key indicator of truncation during data transfer) or the presence of … in the string (which was the main indicator that parts were not transcribed). There were about 35,000 truncated strings identified by these methods for which manual corrections were added, of which 28,000 were employers. This is about 20% of the employer records for this year. This method should ensure that the target for as complete coverage of the census records as possible is achieved, despite the truncation problem in I-CeM.

2.3. The 1871 Census

For 1871 there are no usable occupation strings in I-CeM or that can be derived from FMP. They were never transcribed by FMP or Ancestry. Other items missing from the FMP transcriptions and hence from I-CeM are marital status and birthplace. For this year the extraction algorithm described below was applied directly to the S&N data and then cleaned and parsed in the same way as the other data. This provides complete coverage of employers for 1871 in the same way as for the other censuses 1851-81. It also provides birthplace, but not marital status as this was not transcribed by S&N (the same as FMP). However, there are various differences in the coding of the data with I-CeM because of differences in the structure of the S&N data. Aligning coding required a significant exercise for which additional support was provided by the Isaac Newton Trust. The full documentation of the 1871 extraction is discussed in another working paper. For the extraction of

employers, however, the methods of extraction, cleaning and parsing were identical to that for the other census years; only possible differences in transcription may influence comparisons between censuses.

3. Extracting entrepreneurs from the CEBs: employers with employees

Employers with employees were extracted using an algorithm developed for a pilot study on 1881 financed by the Leverhulme Trust.[18] This was refined for 1851, 1861, and 1871 as described here. The primary source is the text strings that described occupations. The algorithm picks out from all occupation strings those containing digits or written numbers and an employee-word, such as men, women, boys, girls, sons, daughters, males, females, labourers, servants, apprentices, assistants, people, person, hands, journeymen/woman, and employees. In addition, all strings containing the word 'employs' or 'employing' were extracted. This identifies as potential employers around 15% of all unique strings, but as many of these were unique to one individual, this accounts for < 1% of the population as a whole.

3.1. Parsing the employees

The employees contained in the extracted occupational strings were parsed through a second algorithm, adapted for 1851-71.[19] This algorithm separated the occupational descriptor from the declared employees, and subdivides the employees into the following categories:

- men, including descriptors that were obviously men, such as tradesmen
- women, including descriptors easily identifiable as women such as shopwoman
- boys, including lads
- girls, including young ladies
- labourers, with indoor and outdoor labourers added up
- apprentices
- journeymen/women
- male, when gender is stated but no indication of age, such as in the case of ‘son’, or for contractions such as ‘20 men and boys’
- female, when gender is stated but no indication of age, such as in the case of ‘daughter’, or

[18] See Bennett, R. J. and Newton G. (2015) Employers and the 1881 population census of England and Wales, *Local Population Studies*, 94, 29-49.
[19] Bennett, R. J. and Newton G. (2015) *Employers and the 1881 population census*.

for contractions such as ‘14 women and girls’

- children, including young persons and contractions such as ‘boys and girls’
- other, including general non-gendered titles and occupations such as servants, clerks, assistants, baker, hands or where contractions include both males and females and/or unknowns, e.g. ’65 men, women, and boys’ or ‘9 males and apprentices’.

In cases where descriptions would allow an employee to fit into more than one category, they were assigned in the order of the list above. In practice, this only applied to a small number: fewer than 100 female labourers in 1851.This arose as a result of farmers being asked to return ‘labourers’ in that year. Since the example give in the household schedule enumeration form for trades mentioned ‘men’, non-farmers tended to make the distinction between men and women, whereas most farmers just returned labourers, only broken down into indoor and outdoor labourers.

The list of employees and their spelling variants was compiled in a data-driven manner: the corpus of distinct extracted strings was processed and the non-parsed residual examined iteratively, allowing for the most common missed variant each time, until the residual fell below 20% of all distinct strings (meaning that more than 80% had been processed by algorithm). The residual strings that did not parse were then passed to research assistants for manual processing. Their accuracy rate was good: a random sample of 1,000 strings returned 35 mistakes (3.5%) including missed dual occupations they had been asked to flag. An additional check was made to ensure that there were no mistakes of addition of total employees for all the largest employers, as the small numbers in this category have potential to distort analysis of this group in later stages of the project. Also larger employers often had the most complex strings that were most difficult to parse, and were more likely to list their workers broken down by men/women/boys/girls, and then give the total, leading to potential double-counting errors.

3.2. Cleaning the employers with employees

The employers with employee strings required some additional cleaning. For 1851 the nature of the I-CeM data meant that the occupational string field was truncated at 100 characters. For 1881 the same field was similarly truncated at 80 characters but subsequently augmented with additional elements up to 100 characters provided by Kevin Schürer. Some of the occupational information appeared to exceed these limits, and long strings were likely to be employers with employees as these tended to require more characters. All long strings that were not picked up by the extraction

algorithm were checked, and if relevant, were added manually to the database, and parsed by hand. In addition, strings that had either 100 or 99 characters that were picked up and parsed normally were checked manually against the CEBs, as due to their cut-off point some part of the string including employees could be missing. The manual repairs necessary to fix the 1861 data meant that this step could be skipped for that year, as the full string had been entered as part of this process. For 1871, and for 1851 data supplied by S&N, string length had been checked by S&N internally before data extraction, which eliminated this step of cleaning as well.

One feature of the data is that in many families the occupational description of the Head of Household (HoH) is repeated for the rest of the family with an indication of their relationship. For example, a HoH would be described as a 'shoemaker employing 3 men', and his wife's description was 'shoemaker employing 3 men wife', with his children 'shoemaker employing 3 men son' and 'shoemaker employing 3 men daur'. In some cases, however, the relation was not mentioned and the HoH's occupation string was repeated verbatim in the occupation fields of the rest of the household. All these strings were picked up by the algorithm. In many cases, it was unclear whether these were genuine descriptors, as in the case of two people working in a partnership and returning identical occupational descriptors, or whether they were familial descriptors.

The first check on employers was based on age. All people under 14 were removed, while all people of 16 and over were considered genuine potential partners in the first instance. Those in between were considered on a case by case basis with many checks on the original CEBs, but in practice, all those checked were found to belong to a different person than in the database, or to have had their age mis-keyed.

Women enumerated with their husband's occupations as well as 'wife', whether they were employers or not, could not be relied upon as an indicator of their actual activities as the extent to which the occupations of wives was recorded varied significantly between enumerators. The phrase was probably often used as a signifier of social identity rather than an economic function. Certain enumerators described all married women in their district as their husband's occupation's wife, so in these places the descriptor was not a meaningful indicator of occupation or entrepreneurship.[20] As a result of these complications, it was decided to flag all those who had entrepreneur occupational descriptors and had the additional descriptor (such as wife, son, and daughter, etc.) so that they can be included in analysis of entrepreneurs or excluded as desired. There is a relatively small number

[20] You, X., 'Women's Employment in England and Wales, 1851-1911'. PhD Thesis, Cambridge, 2014, 196-223

of people in this category so that there will be generally little effect in most analyses.

In some cases, the occupation string was truncated in I-CeM through splitting across the records of different individuals. This resulted from the original manuscript CEBs displaying a long occupation string over two or more consecutive lines. In these cases the occupational descriptor was separated from the declaration of employees, and the algorithm only picked up part of it. This was identified as a significant problem, especially for females, in the pilot study for 1881.[21] There were two variations of this problem. In the first type, the split occurred after the occupational descriptor, meaning that this was not picked up by the extraction algorithm, but the next line started with 'employs' and contained the employees, which meant the second half was picked up. In the second type, the split occurred after the word 'employs', meaning that the first half of the string was picked up by the extraction algorithm because it contained (a variation of) 'employs', and the second half because it contained the employees, but they were split between two people. In the first case, the employees were attached to the wrong person, usually the employer's wife; in the second case information was divided between two people, again usually husband and wife. In the first case order the households of all strings without an occupational descriptor and all strings starting with 'employ' were examined, and in case of a split, the actual employer was brought into the database, the strings united and the erroneous person deleted. In the second case, the strings were linked up and the erroneous person removed as an entrepreneur.

A final problem was that many of the employers with employees had been allocated the wrong occode in I-CeM. This was a consequence of the employee descriptors in the strings, which caused people who employed e.g. porters, to have been allocated the occode for porter. This was significant enough to distort the analysis in coding many worker-type occupations to the employer. All strings were recoded based on the employer's occupation descriptor extracted and separated from the employee description as part of the parsing algorithm. At this stage, employers with multiple occupations were identified as well, and allocated second or multiple occodes.

4. Extracting entrepreneurs from the CEB records: Own Account

Unlike the later censuses, the 1851-1881 censuses have no reliable indicator for self-employed sole proprietors; the individuals that later censuses described as 'own account': people who neither

[21] Bennett, R. J. and Newton G. (2015) *Employers and the 1881 population census*. pp. 42-3.

employed others nor worked for an employer. However, there are certain assumptions we can make about the employment status of certain groups that allows some of these to be identified form their occupational descriptors. These groups consist mainly of masters/mistresses, farmers, and partners/company owners.

4.1 Masters and mistresses

The 1851 GRO report made a point about distinguishing masters from men in order to assemble data that could be used to interpret industrial status rather than just occupations.[22] Although the report stated that this return was considered imperfect as not all masters returned themselves thus, the descriptors of masters without employees can be used to identify potential self-employed own account proprietors.

The instructions in the 1851 census indicated that "In TRADES the Master is to be distinguished from the Journeyman and Apprentice, thus – (*Carpenter – Master employing* [6] *men*".[23] This instruction was dropped in 1861, where the instruction simply asked for the employer, and in 1871 the General Instruction to the householders stated that "Masters must, in all cases be distinguished", and again in 1881, "Masters must, in all cases, be so designated".[24] Masters who did not employ any men – or who did not record this – still distinguished themselves from journeymen and apprentices by stating they were masters, either as e.g. MASTER BAKER or as CARPENTER, MASTER. In addition, although the language used was explicitly male, many women who qualified as mistress of their trade were recorded as such.

To extract all masters and mistresses from these records, a wildcard search was used for m*st*r and m*str*s on the residual of the algorithm used to extract the employers with employees. Next, occupations which included the term master/mistress but did not refer to own account tradespeople – the spurious masters – were removed from the database.

The list of spurious masters believed to be of employee rather than own account status excluded from the database were as follows:

POSTMASTER
STATION MASTER

[22] Census of Great Britain, 1851. Population Tables Vol 1; lxxvii-lxxviii
[23] General Instruction, Census of the Population, 1851.
[24] General Instruction, Census of England and Wales, 1861, 1871 & 1881; also see Working Paper 2.

MASTER OF WORKHOUSE
MASTER OF UNION
HEADMASTER
WARD MASTER
TASK MASTER
DOCK MASTERS
HARBOUR MASTERS
LOCK MASTERS
TIDE MASTER
LIGHTERMAN MASTER
PIER MASTER
CHOIR MASTER
MASTER OF ARTS
BARMASTER – (refers to the Barmoot court - a Derbyshire leadmining term); also other court-related masters.
MASTER RN
QUARTER MASTER
BAND MASTER
BARRACK MASTER
PAYMASTER
DRILL MASTER
RIDING MASTER (unless private teacher),
FENCING MASTER (unless private teacher).
BARGE MASTER (unless explicitly stated that they were (part-) owner)
FLAT MASTER (unless explicitly stated that they were (part-) owner)
BOAT MASTER (unless explicitly stated that they were (part-) owner)
SHIP MASTER (unless explicitly stated that they were (part-) owner)
MASTER MARINER (unless explicitly stated that they were (part-) owner)
SAILOR MASTER (unless explicitly stated that they were (part-) owner)
[MASTER OF NAMED VESSELS] (unless explicitly stated that they were (part-) owner)
MASTER (just 'masters' was always shipping related)
SCHOOLMASTER and SCHOOLMISTRESS, with a few exceptions (unless proprietors of private academies etc.)

Comparisons with the later censuses for 1891-1911, where the employee and own account status of these categories is explicitly identified, were used also as a guide to these choices. One of the most difficult categories was schoolmasters/mistresses. The nineteenth century schooling system was complicated. Broadly speaking, before the Education Acts of the 1870s there were three main kinds of schooling for the majority of the population: Sunday Schools, charitable day schools (National School, Lancastrian/British), and Dame and private schools. The last category was usually small scale (10-30 pupils) and met at the proprietor's house, meaning that schoolmasters and mistresses of these schools were essentially own account business people. However, as most occupational descriptors did not specify the type of school, schoolmasters and schoolmistresses were removed unless it was explicit that they taught in a private school or were proprietors. In order to make sure that this whole group was captured, an additional extraction was performed using 'teacher' and 'tutor' as search terms, with private school teachers/tutors included in the entrepreneur-database. Dancing masters and music masters were also included as own account, as well as teachers described as working in these fields, as these were normally all private peripatetic teachers. Language and mathematical masters and teachers were excluded from own account. While some of these may have been private peripatetic, the majority were employees of a school and so they were not included as a group.

A similar strategy applied to the shipping-related masters. While some of these would have been working on their own account, the shipping list suggests that only 16% of barge masters also owned the boat, while the rest worked for a company or master.[25] This was corroborated by using employment data from the 1891 census, which showed that only about 12% of barge masters responded as own account. As the majority were employees, they have been excluded from the entrepreneur's database unless explicitly stated that they were (part-) owner of a boat.

All masters and mistresses that were retained were checked for multiple occupations and, if necessary, recoded. Additional checks were run on age, with most masters under 20 years old removed.

4.2. Farmers

A second group of people assumed to be working on own account were farmers. Farmers were considered masters by the GRO in the 1851 report.[26] In addition, as landowners they were in

[25] See e.g. *Gloucester Registration Authority Canal Boat Registrations 1879-1891*, and *Stratford-upon-Avon Canal boat registrations 1879-1890*. Transcribed copies available through The Eureka partnership, Aylesbury 2010-2014
[26] Census of Great Britain, 1851. Population Tables Vol 1; lxxviii

definite possession of business assets.

Farmers were extracted based on I-CeM occupational codes in the residual of the employers with employees extraction. All people with the occode 173 (Farmer, grazier) as well as 177 (Farm bailiffs stewards and foremen), 184 (Market gardeners) and 187 (Others in agriculture) were extracted, as well as all people who returned an acreage.

The general instructions to the householders from 1851 onwards made clear that "the term FARMER to be applied only to the occupier of land", who were asked to return their number of acres.[27] In 1881 the instructions stated that farmers should return the number of acres occupied, without clarifying what farmer meant.[28] In practice, bailiffs, market gardeners, cottagers and other people occupying land regularly also returned an acreage as well. An algorithm originally used for the 1881 pilot and adapted here extracted people with acreages, and parsed the number of acres returned in a separate field.[29] All farmer-related professions were parsed with this algorithm, with the remainder given to research assistants for manual acreage-parsing.

However, it proved ambiguous whether people who returned an acreage were actually farming it. For instance, there were bakers also occupying 2 acres who were unlikely to be farmers as a main occupation. A difficulty that affects this issue is that 'farmer' took precedence in the census instruction so that normally farmer was given first before any other occupation, even if the other occupation was clearly the 'main' occupation.[30] The way in which I-CeM coded farmers also gave them a similar precedence. This was corrected as far as possible through occode-cleaning, though portfolio entrepreneurs and some others will have been corrected further though inspection of the CEBs. In addition, for very small farmers or others with acres it is difficult to judge whether this was farming for the market, subsistence farming, a piece of land where they kept some animals, or just wasteland. The geographical variation in the area of land required for non-subsistence farming varies, so that it is difficult to be sure how far these were true farmers. In addition, there was a sizable subgroup of 'cottagers' who returned an acreage that was on average much smaller than that returned by farmers. Cottagers were usually tied to a landowners' estate in a feudal system but had a higher level of independence than agricultural labourers; but these would not normally be classed as entrepreneurs since they were usually only of subsistence on the land, and/or were de facto agricultural labourers with some land of their own for household support. For the purposes of

[27] General Instruction, Census of the Population, 1851.
[28] General Instruction, Census of England and Wales, 1881.
[29] Bennett, R. J. and Newton G. (2015) *Employers and the 1881 population census*.
[30] See e. g. Higgs, Edward (2005) *Making Sense of the Census Revisited*.

defining entrepreneurs these had to be excluded. It was decided to define as farmers all who either:

a) explicitly mentioned farmer, farming, farm, or grazier,
b) occupied land and employed people to work it (these would have been extracted as part of the employers with employees),
c) but to classify those who were described as cottagers as their main occupation as a separate group (tagged in order to be able to include or remove them depending on the form of analysis adopted).

Family members and bailiffs who returned an acreage kept their original occode, but have been included as own account workers as they probably returned the farm's information on behalf of a farmer, whose information would otherwise be missed (where they also returned employees they are counted with other farm employers as surrogates for the actual farmer). However, other occupations who returned land were excluded from the own account category, though retained with a separate tag with land owners (see below).

All farmers who were retained were checked for multiple occupations and, if necessary, recoded to their 'main' occupation if clear. They were also checked for age, with all farmers under 14 removed.

4.3. Partners and company owners

A final category of people who can be assumed to be working on own account were partners or owners of companies or firms. These were extracted from the residual of the employers with employees algorithm by search term on the terms p*rtn*r, firm, with, joint, and company. They were checked for spurious partners.

People who were enumerated as working with their parent, e.g. BAKER WITH HIS FATHER, SEAMSTRESS WITH MOTHER, were removed if they were under 24, as it was assumed they were workers or in a *de facto* apprenticeship. Those who were 24 or over were assumed to be junior partners.

Partners and firm owners were checked for multiple occupations and recoded if necessary. Partners were subsequently also given additional codes so that those identified in the census as operating in

partnerships could be subjected to specific analysis, as has been done for the 1881 pilot study.[31] As noted in that pilot, the census records cannot be taken as a record of all partners or all partnerships, but they are a large sample of the population who were partners.

5. Identifying asset holders, directors, and landowners

There were several other groups of people who were likely to be working on own account or employ others, but who would not have returned this to the census as they were not in either of the categories asked to list workers. People described as owning certain business assets which would make it likely they either employed people or used that asset as their main means of income have been extracted separately from the residual of the extraction of employers with employees. This includes people such as threshing machine owners, mine owners, and ship owners. The main search terms used were own*r and propr*tor, which were supplemented by additional checks for mine, colliery, and ship. These strings were cleaned and, if necessary, re-coded as a separate group in the entrepreneurship database. In some parts of the analysis these people are added to the employers with employees, as it is assumed they at least indirectly employed people. Some additional enrichment of the owner category has been undertaken using other sources, especially for mine and quarry owners, as discussed in another working paper.

People described as owners of land and/or of houses are more ambiguous, as these can include some large scale landowners employing many people as well as any owner of a piece of land. In addition, an owner of a house could refer to the house the respondent lived in as part of their occupation if their main income was from lodgings, or they rented out the whole house. This group should be added to the own account people in certain analyses, as they are likely to be self-supporting.

A last group extracted are the directors of Limited Companies. Company directors officially did not employ anyone, nor owned the business assets, as the workforce was employed by the company and the assets were owned by the company (and ultimately by the shareholders and/or debt holders). However, they provided the human link as entrepreneurs who were the responsible decision-makers of that business. They can in some cases be identified in the census and have been extracted. However, as the majority of directors did not record themselves explicitly primarily as directors this

[31] Bennett, R.J. (2016) Interpreting business partnerships in late Victorian Britain, *Economic History Review*, 69,4,1199-1227.

category has been supplemented through data enrichment in a separate analysis, as discussed in another working paper.

6. Conclusion

This paper describes the several extraction and infill methods that together produce a new entrepreneurship database for the years 1851 to 1881, which can be linked to I-CeM and other census records to obtain socio-economic, household, and other information to allow statistical analysis of entrepreneurs. Other Working Papers examine further details of aspects of the extraction and coding of individuals and occupations to yield the final database and variables that can be used for statistical analysis, notably the reconstruction process by which the estimation of the own account population in 1851-81 can be improved. Other papers also describe how the extraction that is possible from the census is supplemented by data enrichment from a variety of sources, and tests are made of the completeness of coverage of the whole business population.

As noted at the outset, the census is an imperfect source of information about employers. However, this paper, and others, demonstrates the considerable potential for identifying employers and information about their workforces, as well as the potential for identifying self-employed sole proprietors.

Acknowledgments:

This research has been supported by the ESRC under project grant ES/M010953: **Drivers of Entrepreneurship and Small Businesses.** Piloting of the research for 1881 draws from Leverhulme Trust grant RG66385: **The long-term evolution of Small and Medium-Sized Enterprises (SMEs).**

The database referred to for 1851-61 and 1881-1911 derives from K. Schürer, E. Higgs, A.M. Reid, E.M Garrett, *Integrated Census Microdata, 1851-1911, version V. 2 (I-CeM.2)*, (2016) [data collection]. UK Data Service, SN: 7481, http://dx.doi.org/10.5255/UKDA-SN-7481-1; enhanced; 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).

The I-CeM data for 1851 are infilled and enhanced for about 75,000 from records supplied by records S&N Genealogy Supplies (TheGenealogist.co.uk), with assistance for checks of employers from the 1851 manuscript CEBs by Gavin Robinson.

For 1861 the I-CeM data for approximately 35,000 incomplete employer records have been checked and infilled from the manuscript CEBs by Mark Latham, Gavin Robinson, and Tiffany Shumaker.

The data for 1871 are entirely derived, from processing special tabulations from S&N Genealogy Supplies (TheGenealogist.co.uk), with additional financial support from the Isaac Newton Trust.

The data used for the 1881 pilot, supported by the Leverhulme Trust, derives from Schürer, Kevin and Woollard, Matthew (University of Essex) (2000) *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)* [computer file] UKDA, SN-4177, transcribed by Genealogical Society of Utah and Federation of Family History Societies.

Additional research assistance for parsing data for 1851-1871 was supplied by AEL Data, and for 1881 by Thymus.

The GIS boundary files for RSDs were constructed by Joe Day for the ESRC fertility project directed by Alice Reid:

http://www.geog.cam.ac.uk/research/projects/victorianfertilitydecline/publications.html

These used as a starting point the GIS parish files of Satchell, A.E.M., Kitson, P.M.K., Newton, G.H., Shaw-Taylor, L., Wrigley E.A. (2006) *1851 England and Wales census parishes, townships and places*, 2006, ESRC RES-000-23-1579, supported by Leverhulme Trust and the British Academy; Satchell, A.E.M. (2015) *England and Wales census parishes, townships and places*; which is an enhanced and corrected version of Burton, N, Westwood J., and Carter P. (2014) *GIS of the ancient parishes of England and Wales, 1500-1850*, UKDA, SN 4828; which is a GIS version of Kain, R.J.P., and Oliver, R.R. (2001) *Historic parishes of England and Wales: An electronic map of boundaries before 1850 with a gazetteer and metadata*, UKDA, SN 4348.

A special acknowledgement of thanks is made to Kevin Schürer for advice and all his help in developing improved versions of I-CeM, and to Alice Reid, Eilidh Garrett, Joe Day, Hanna Jaadla, Xuesheng You, Leigh Shaw-Taylor and other members of the Campop I-CeM group who, with the current authors, have collectively worked on the new versions of I-CeM.

References.

Bennett, Robert J. (2016) Interpreting business partnerships in late Victorian Britain, *Economic History Review*, 69, 4, 1199–1227.

Bennett, Robert J. and Newton Gill (2015) Employers and the 1881 population census of England and Wales, *Local Population Studies*, 94, 29-49.

Higgs, Edward (2005) *Making Sense of the Census Revisited: Census Records for England and Wales 1801-1901*, Institute of Historical Research, National Archives, London.

Higgs, Edward and Schürer, Kevin (2014) *The Integrated Census Microdata (I-CeM)* UKDA, University of Essex, SN-7481.

Schürer, Kevin and Woollard, Matthew (2000) *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)*, UKDA, University of Essex, SN-4177.

Schürer, Kevin, Higgs, Edward, Reid, Alice M., Garrett, Eilidh M. (2016) *Integrated Census Microdata V.2 (I-CeM.2)* [data collection].

You, Xuesheng (2014) *Women's Employment in England and Wales, 1851-1911*, PhD Thesis, Cambridge University.

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.

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.*

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.*

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

WP 4: Smith, Harry J., Bennett, Robert J., and van Lieshout, Carry (2017) *Extracting entrepreneurs from the Censuses, 1891-1911.*

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.*

Full list of all current Working Papers available at:
http://www.geog.cam.ac.uk/research/projects/historyofentrepreneurship/