Extraction of data on Entrepreneurs from the 1871 Census to supplement I-CeM

Carry van Lieshout, Joseph Day, Piero Montebruno, and Robert J. Bennett

cv313@cam.ac.uk jd466@cam.ac.uk pfm27@cam.ac.uk rjb7@cam.ac.uk

Working Paper 12:
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.

September 2018

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

Extraction of data on Entrepreneurs from the 1871 Census to supplement I-CeM

Carry van Lieshout, Joseph Day, Piero Montebruno, and Robert J. Bennett

Working Paper 12: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge.

1. Introduction

This paper sets out the process of obtaining, processing, and evaluating the entrepreneurship database for the 1871 census as created and deposited by ESRC-supported project ES/M010953 '*Drivers of Entrepreneurship and Small Businesses'*. This project uses *The Integrated Census Microdata* (I-CeM) as its main source.[1] However, I-CeM does not cover England and Wales for the 1871 census, and if it is deposited in future it will lack crucial data that were never transcribed from the census manuscripts (occupation, marital status, and birthplace). While there is a skeleton version of 1871 available from Kevin Schürer, this lacks the occupation string field with all the variables (such as occode) that are based on this field, as well as marital status and birthplace. The occupation string is essential for the analysis of entrepreneurs in the early censuses as it provides the only indication of employment status, and is also used to parse sector, acreage, and number of employees.

After successful collaboration with S&N [theGenealogist.co.uk] on missing employer data in I-CeM for 1851, as discussed in Working Paper 3, data on 1871 entrepreneurs was obtained from their database. This was acquired and processed as part of an additional project, funded by Newton Trust Grant 17.07(d): *Business Employers in 1871*.

This paper sets out the process of acquiring the S&N data, explains the differences between the S&N and I-CeM derived data, and evaluates the data. The general process of extracting, parsing, and cleaning the early census data is the same for both data sources, as described in Working Paper 3. Working Paper 1 gives an overview of the entrepreneurship project as a whole, and Working Paper 2 defines in detail the different censuses and the challenges they

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

present for identifying entrepreneurs. A full list of Working Papers is included at the end of this paper.

2. Extracting the data

2.1. The method

After piloting with S&N to test feasibility and local operator implementation, they were provided with the 3 algorithms that were developed by the entrepreneurship project, which they ran on their 'profession' text strings of their full 1871 census database. We are very grateful to S&N for implementing these searches directly from their full database. The algorithms were:

i) divide.pl A perl algorithm that extracts employers with employees;

ii) acres.pl A perl algorithm that extracts and parses individuals with acres; this was run on the residual of the divide algorithm, meaning it only identified people with acres but without employees.

iii) An extraction query in SQL This identified masters, owners, directors and partners. This was run on the residual as well.

S&N returned the raw results of these algorithms, i.e. the text strings, as well as a count of individuals associated with a particular text string. As the raw strings still include many spurious hits, with the help of the string counts the final number of individuals to order could be reduced. As these had to be purchased per string, cost constraints prohibited the acquisition of the full anonymised database (comparable with I-CeM), or the full data as extracted for other census years. This constraint has some consequences for the full entrepreneur candidates that can be identified, as discussed below.

2.2. Differences between I-CeM and S&N extraction

The algorithms used were the same as used on the I-CeM data, and for most entrepreneurs the extraction method resulted in corresponding results between the early censuses. However, there remain several key differences between the 1871 S&N data and the data extracted from I-CeM for the 1851, 1861, and 1881 censuses.

For the purpose of analysis, several distinctive Groups of entrepreneurs were created based on the method by which they could be identified and extracted. Table 1 summarises the extraction process for each Group and the resulting differences between I-CeM and S&N data.

Group	**Description**	**Extraction process**	**Difference S&N – I-CeM**
1	All employers and any others (such as masters, proprietors or owners) with stated employees; farmers with stated employees; partners with stated employees	divide.pl on all occupation strings (see WP3 section 3)	No difference
2	All 'employers' with no employees; 'masters' and anyone else who includes 'emp' in their occupation descriptor; and partners without stated employees	divide algorithm and the extraction query (see WP3 sections 3 and 5)	No difference
3	Master etc. anyone including 'master' or 'mistress' in their occupational descriptor but with no employees.	extraction query, with further cleaning (see WP3 section 4.1)	No difference
4	'Farmer' not stating 'emp' or acres.	In I-CeM, farmers were identified based on occode, and then parsed on acreage (see WP3 section 4.2). In addition, everyone who returned any acreage was extracted (although many of these were later cleaned out)	In S&N, filtering on occode was not possible, so only those who returned acreage were extracted. This means that farmers who just described themselves as FARMER without mentioning employees or acres (but perhaps mentioning a second occupation) have not been extracted. The wider search was too large to be financially viable, and would provide only limited extra information

5	Farmer giving acres but no stated employees, and two or more acres of land (those with less than two acres with no employees were excluded; it was assumed that they work on other farms)	See Group 4. However, since these all have acres, they should be picked up through the acres algorithm	Should correspond to the I-CeM extractions, although I-CeM also includes farmers with less than 2 acres
6	Owners or proprietors of business assets: mine/quarry owner, shipowner, barge owner and others with any business assets (other than land/housing).	extraction query, with further cleaning (See WP3 section 4.3)	No difference
7	'Owners' with no other information (not in 6), including landowners with no employees or only with acres, and house proprietors with no employee information.	extraction query, with further cleaning (See WP3 section 5)	No difference. However, due to financial constraints these were eventually dropped
8	Directors	extraction query, with further cleaning (See WP3 section 5)	No difference

Table 1: S&N vs. I-CeM extraction by entrepreneur Group.

2.3. Pilot study: Derbyshire

A pilot was commissioned from S&N to run all algorithms on Derbyshire only. At an early stage it was found that S&N initially supplied truncated strings, but fortunately they had the full strings as well and were able to supply these (long strings were in lower case and had to be converted before running any algorithms).

S&N supplied from the algorithms:

1) 1,891 unique strings relating to 2,024 individual employers with employees
2) 1,279 unique strings relating to 3,057 individuals with additional acreage
3) 894 unique strings relating to 2,241 extra OA individuals.

As a check, the same algorithms were run on Derbyshire 1861 I-CeM and 1881 pre-I-CeM (GSU), all using raw data in order to compare like with like, although it should be noted that 1881 was particularly well transcribed.

In comparison with the same extractions for Derbyshire in the previous and following census years it appears that both employers and acreages are under-estimated in the 1871 S&N extractions:

Year	Employers	Acreage	OA extraction
1861 (I-CeM)	2292 (2488)	1275 (3235)	818 (2667)
1871 (S&N)	1891 (2024)	1279 (3057)	894 (2241)
1881 (GSU)	2415 (2491)	1635 (4082)	892 (2007)

Table 2: Raw numbers extracted for Derbyshire pilot study. Number of unique strings and individuals (in brackets).

From this it appears that there might be people/strings missing in the employers and acreage categories, i.e. Group 1 entrepreneurs and Group 5 farmers. For the employers, a number in the range of 2,300-2,400 was expected rather than under 1,900, so it appears that around 20% of employers could be missing if a constant employer ratio is assumed. In addition, for the 1861 census certain areas are known to be lost, which includes over half an RSD in Derbyshire (see subsequent Working Paper on this), meaning that the 1861 numbers are already an under-estimate.

In order to check whether there was a transcription issue the complete data from one Registration District, Derby, was provided by S&N. This included 290 employers, 11 additional acreages, and 307 further OA extractions. Note that, as this was an urban RD, acreage was expected to be low. The number of employers in the Derby RD seemed more in line with expectation, based on the other years: 1861: 261 employers; 1871: 290 employers; 1881: 395 employers.

The data received from S&N covered the Derby pieces numbered 3560-3576. Each piece consists of 140 to 200 Census Enumerators Book (CEB) pages containing 20-25 people each, with 6 empty pages at each Enumeration District break. CEB pages were checked for employers and individuals with acres, and these were checked against the data supplied by S&N. Piece 3560 was checked as a whole, then pieces 3561-3565 were checked for the first

100 pages. Pieces 3566-3569 were checked for 50 pages taken from the middle (counting towards the end so as a range slightly skewed towards the end), 3570 to 3574 were checked 50-70 pages from the end. 3575 was checked in its entirety, and 3576 was checked for ¾ taken from the end. In total over 1400 pages were checked, containing roughly 28,000 people, which is almost half of Derby.

The transcriptions were mostly correct, with a total of 7 employers and 1 person whose acreage was missed in transcription.[2] As about half the pages were checked, the assumption is that another 7 may be missing (although they seem to be clustered in certain pieces and towards the end of a piece, so this may not be the case). 14 employers missing on 290 found employers is just under 5%, so this does not account for the 20% of additional employers expected to be found.

This left the following possibilities:

1) There are other parts of Derbyshire which have been transcribed really poorly, worse than piece 3570,
2) There are other parts of Derbyshire missing from the S&N database, or
3) 1871 just does not contain/record as many employers/acres as the other years.

Possibility 1 could not be checked except by doing the above for all RDs, which would be very time-consuming. Possibility 2 was ruled out by requesting a population count in the S&N database by RD for Derbyshire as well as the rest of England and Wales. This was checked against published population totals. This showed that S&N's 1871 is complete (apart from some missing people in Anglesey which they have since corrected). Possibility 3 is unlikely. Later analysis suggests that possibility 1 was indeed the case (see Figure 1 below).

[2] These were: in piece 3562: a cowkeeper with missing acres; In piece 3570: one flour miller employing 4 men, plus a farmer of 40 acres employing 4 men and cotton spinner employing 30 (all taken from the end of the piece with only their profession transcribed - further checks towards the front of this piece proved everything there was correct). In this piece S&N found 17 employers, and 3 were missed, so 3/20 which is close to the 1/5 that seems to be missing; in piece 3571, a silk merchant employing 11 men 55 women, and a master silk manufacturer picked up but missing employees (towards end but some later ones had been done properly - all early employers were fine); in piece 3573, an iron manufacturer at the very end, and the employees of a boot and shoe maker master towards the end of the piece.

2.4. Full England & Wales Extraction

Based on the pilot it was decided that S&N provided sufficient quality of coverage to commission the full England and Wales data. S&N supplied the England and Wales output from the same 3 algorithms, providing unique strings and string counts. The raw data included:

1) 106,819 unique strings relating to 145,239 individual employers with employees
2) 23,140 unique strings relating to 108,408 additional acreage
3) 35,005 unique strings relating to 177,100 extra OAs.

In comparison with the same extractions for England and Wales in the previous and following census years it appeared that the employers and OA extractions were lower in 1871:

Year	**Employers**	**Acreage**	**OA extraction**
1861	131,604 (181,310)	21,263 (97,435)	33,768 (231,052)
1871	106,819 (145,239)	23,140 (108,408)	35,005 (177,100)
1881	146,637 (186,189)	24,647 (112,822)	51,265 (258,295)

Table 3: Raw numbers from the E&W national extraction. Unique string numbers and (individuals in brackets).

After concerns that there might be a sectoral bias, the proportion of farmers over all employers strings was checked, which was (unique/individuals) 53/54% for 1861, 49/51% for 1871, and 46/46% for 1881, so it seems the missing employers were not biased towards farmers.

From this it appears that there might be people and/or unique strings missing in the employers and OA extraction categories. Acres look fine overall, even though there had been some concerns in the pilot. For the employers, again around 20% was missing, which corresponds to what was expected based on the pilot. The lower number of OA individuals (rather than unique strings) was not expected based on the pilot. It does not seem that whole

categories are missing – so this might be either an attribute of the actual data or systematic transcription errors.

2.5. Commissioning the data from S&N

Based on budget constraints, initially the aim was to request around 200,000 individuals, meaning that obtaining the full 430,000 individuals resulting from the algorithms was infeasible. To resolve this, the results of the acreage algorithm and the extraction query contained some overlaps, which were removed. S&N provided age and county which allowed us to remove those from British Islands and those under 14 years old. Agricultural labourers were removed from those with acreage, but it was decided to retain bailiffs, market gardeners, and cottagers. These steps removed around 7,000 individuals. Since the extraction query extracted spurious masters as well, the most numerous strings of these were pre-cleaned (see WP3 section 4.1 for spurious master cleaning), which removed 93,000. The extraction query results were broken down into categories to help prioritise strings.

The 330,000 remaining individuals consisted of:

• Employers with employees:	145,000	These were considered a priority
• Acreage:	106,000	Could be broken down:
○ over 5 acres only:	96,000	
○ Including 5 acres:	100,000	
○ Including 4 acres:	102,000	
○ Including 3 acres:	104,000	
○ Including 2 acres:	105,000	
• Cleaned masters/mistresses:	34,000	
• Directors:	200	
• Partners:	1,200	
• ‘Owners’:	9,100	[of ships, mines, newspapers, carts etc.]
• houseowners:	4,100	
• landowners:	22,600	
• 'tail' of strings to be cleaned:	6,000	still contains some masters and owners

As it was deemed impossible to cut down to 200,000 people, new negotiations focused on cutting down to below 300,000, with the recognition that some additional strings would need to be ordered at a later stage to fix the split string problem for the employers (this was around 2,000 strings).

It was decided to cut those with an acreage under 2 acres; to cut landowners (an ambiguous category and not a priority in analysis), and to cut house owners (an ambiguous category and mostly spurious as entrepreneurs). Note that the 'tail' of strings, containing strings that only occurred a few times, still contains masters and owners, as well as spurious masters and owners. These were cleaned as part of the occoding process.

Final numbers ordered were:

Employers:	142,978
'Farmers' (acreage over 2):	105,330
Cleaned master/mistress:	33,736
'Owners':	8,987
Partner/director:	1,337
'Tail':	5,840

Total: 298,208

2.6. S&N vs I-CeM data

Comparison of the final S&N content achieved, compared to I-CeM, is shown in Table 4.

Category		Group	ICEM	S&N	Notes
Employers with Employees		1 & 2	√	√	Should be the same based on method, however there are an estimated 20% 'missing'
Masters/Mistresses		3	√	√	Should be the same based on method
Farmers as defined in WP3 p. 21					
	With employees (may also state acres)	1 & 2	√	√	Identified through employees; should be present but subject to an estimated 20% 'missing'
	With acres, without employees	5	√	>2 acres only	Identified through acreage, can be filtered
	Without stated acres or employees	4	√	n/a	Identified through occode in I-CE-eM, not available S&N
Non-farmers with acreage (market gardeners, cottagers, some others with acres)		0	√	>2 acres only	Identified through acreage, can be filtered
Partners/company owners		1,2 & 6	√	√	Should be the same based on method
Asset holders					
	Owners of carts, ships, mines, newspapers, shops, mills, hotel etc.	6	√	√	Should be the same based on method
	Houseowners	0	√	n/a	Excluded
	Landowners	0	√	n/a	Excluded
Directors		8	√	√	Should be the same based on method

Table 4: Commissioned data received form S&N by category/Group compared to I-CeM.

3. Processing the data

The S&N extracted individuals had to be parsed in the same way as I-CeM extractions. In addition, they had to be made compatible to the I-CeM database structure. The main tasks were a) aligning the spatial data as provided by S&N with the I-CeM parish dictionaries, to allow spatial analysis and mapping, and b) creating occodes for the occupation string data. In the I-CeM derived extractions only the 'occupation' part of the employer strings had to be

coded from scratch, while the other occodes had to be checked for accuracy and portfolios. In the case of 1871, the full dataset had to be provided with (multiple) occodes.

3.1. The S&N data

S&N provided its individuals with a pipe | separated csv file:

combined_wheat_1871_RAW_DATA containing:

uid	the S&N unique identifier
natural_order	gives an idea of the order that people appear on a page but is not unique
household_id	
family_id	iterative and resets per household
piece	
page_number	
street	
parish	
area	RD
county	
forename	
surname	
gender	a generated field, not transcribed, so may not always be 100% accurate
relationship	
age	
Profession	
birth_parish	
birth_county	

S&N, like I-CeM, does not have Marital Status nor birthplace, which remain major gaps for analysis.

3.2. Spatial alignment

The S&N sample contained the following information pertinent to matching the data to the 1871 parish dictionaries:

County	These were standardised to correspond to the counties in the dictionary.
Area	These largely corresponded to the registration district in the dictionary.
Parish	Although ostensibly a 'parish' identifier, these also included vessels, townships, hamlets and tythings which were not included in the dictionary or described in census reports.
Piece	Piece numbers refer to collections of individual enumerators' books for a district.
Address	Address information varied in quality from an exact street address or misspelt road names through to hamlet/township names or vague descriptions, e.g. 'cottage'.

This information was used to match the S&N places to the 1871 Cambridge Group Parish Dictionary which included:

Registration County	Census of England and Wales was divided into 52 registration counties (RC)
Registration District	Sub-divided into 627 registration districts (RD)
Registration Sub-District	Sub-divided into 2,195 registration districts (RSD)
Civil Parish/Part	Sub-divided into 16,028 civil parishes/parts of parishes

Matches were first made between: County and Registration County, Area and Registration District and Parish and Civil Parish/Part. However, as complete strings in both lists seldom matched exactly and were described in slightly different terms (e.g. "Saffron Hill, Hatton Gardens and Ely Rents Liberty" in the S&N database vs. "SAFFRON HILL, HATTON GARDEN, ELY RENTS, AND ELY PLACE" in the parish dictionary) matches were made by the number of words matched. In the S&N database string "Saffron Hill, Hatton Gardens and Ely Rents Liberty", six words match the string "SAFFRON HILL, HATTON GARDEN, ELY RENTS, AND ELY PLACE" in the parish dictionary. This is more than any other string. The S&N string is therefore matched to the parish dictionary. Where an S&N string is

matched to more than one place in the parish dictionary, these are manually checked. Approximately 13,000 strings could be automatically matched using this method.

This automated matching was checked using the S&N 'piece' number. As a rule, a piece number ought to refer to only one RSD. Where this automated matching resulted in places in the same piece number being allocated to parIDs in the parish dictionary in two or more RSDs, these were manually checked and removed. Of those places that could not be matched to a single parID in the parish dictionaries – either because of a misspelling or because there was more than one potential match – the piece numbers were used to narrow down the potential places to which a place in the S&N database could be matched. Where no matches were made in the initial parse (e.g. Castelldwyran in piece 5508) given that all other places in that piece were in RSD 594.4, Castelldwyran could be manually matched to "Castle Dyrran" in RSD 594.4. Limiting the number of possible matches in this way significantly reduced the matching workload.

This process reduced the number of unmatched places to approx. 3,000. The next step required searching for townships and places that were not given in the parish dictionary, which were matched using Joe Day's look-up dictionary of place names and on-line gazetteers.[3]

The next step was to match those S&N places which could be matched to multiple RSDs in the parish dictionaries (e.g. Paddington and numerous other parishes in London/other major cities were split across multiple RSDs). For this the address information was utilised. RSD boundaries were mapped with underlying street maps and addresses given in the piece number were used to determine which RSD should be matched to which piece number.

This left a residual of places to be matched which could not be matched using a piece number in conjunction with other descriptors – e.g. those S&N places with a piece number that had no pre-existing match to the parish dictionary. Generally, piece numbers follow the same order as RSDs and therefore, if piece number 2126 was matched to RSD 139.4 and piece number 2136 was matched to RSD 140.2, the piece numbers in between were matched to all

[3] Joe Day (2017*) Leaving Home and Migrating in Nineteenth-Century England and Wales: Evidence from the 1881 Census Enumerators' Books (CEBs)*, PhD, University of Cambridge, Faculty of History.

RSDs between RSD 139.4 and RSD 140.2 inclusive. These were then manually searched to select the correct matches.

Once all S&N places were assigned at least one match in the parish dictionary, several checks were undertaken to ensure matches were correct and as accurate as possible. Since piece numbers are generally in the same order as RSDs, RSD numbers should be in numerical order when sorted on piece numbers. Where this was not the case, the match was checked and if necessary corrected. S&N places that were matched to more than one parID in the parish dictionary were checked to see if this could be refined and be matched to fewer places, however, in some cases this was inconclusive. While all individuals could be allocated to an RSD, the remaining individuals who had multiple ParID matches have been allocated the first ParID in their RSD as 'assignedParID'.

3.3. Occodes

The employers with employees were parsed according to the usual method (see WP3 section 3.1). There were 105,609 unique employer occupations. 86,599 were auto-parsed using parse.pl (82%). The remaining 19,010 were sent to AELData for manual processing, and checked. The parsing process split the employer occupation from the employee information. The employer occupation strings were added to the unique occupations from the other algorithms, resulting in 65,590 unique occupation strings that required up to 6 occodes for each of the separate entrepreneurial occupations mentioned. While a proportion of these could be matched against the existing occode dictionaries resulting from previous census work, consistency checks and the residual had to be performed manually.

Subsequently, individuals were cleaned to remove spurious masters, non-entrepreneurs in the 'tail', and descriptors of relations to an entrepreneur ("BAKER MASTER'S ASSISTANT").

3.4. Further alignment to I-CeM

In addition to the spatial data and occodes, some further issues had to be aligned as well.

- Birth county: S&N's birth county has been matched to I-CeM county and country codes. Yorkshire was kept together as one county.

- Relationship: the S&N relationship descriptor was not coded and had to coded to theRELA_10 codes used in other parts of the analysis used in the project.

3.5. Database checks and cleaning

S&N assisted the split line cleaning by providing newly transcribed occupations of people who had been identified as potentially split over several lines. This added 957 new entrepreneurs to the database.

In addition to checks on all large employers with over 100 employees (70 for farmers), all female employers with more than 20 employees were checked. As in the case of I-CeM, many of these turned out to be cases of the husband's occupation wrongly allocated to the wife. As the full S&N database was not available to pull these real employers in (and delete the spurious ones) the original entries have been altered to M in sex with the appropriate age and HEAD in relationship, and have been allocated a NewID that likely corresponds to the correct person's uid in S&N's database. The old uids have been preserved to maintain the link back to S&N. In order to save time, and since names are not subject to analysis, the forename has been changed to 'husband of ORIGINAL NAME' except in cases when there was no family relation (e.g. if the spurious entry was a servant).

4. Evaluation

The total number of entrepreneurs extracted by S&N is summarised in Table 5.

Group	**N**
1	138,751
2	4,244
3	31,484
4	230
5	93,903
6	9,343
8	153

Table 5: Total N in 1871 'cleaned' entrepreneurs database derived from S&N

The number of Group 4s (farmers without acres and/or employees) is incomplete and only includes farmers that were picked up accidentally as part of one of the other search terms.

4.1. Employers

The number of Group 1 entrepreneurs (employers with employees), as predicted by the pilot and preliminary checks, is lower than expected based on the 1851-1881 trends. In addition, the national employer-ratio (% of employers in the whole population) is 0.61%, while for the other census years this was in the 0.8-0.9% area. However, this is not evenly distributed throughout the country. In 112 RSDs there were no employers at all, and another 116 RSDs have an employer-ratio of less than 0.05%.

Figure 1 shows the areas where employer ratios where lowest. Analysis of other early census years has shown what while there are a few areas that genuinely have 0.1 to 0.5 % employers, such as South Wales, parts of East London, and parts of the North East, overall a coverage of at least 0.5% employers would be expected. It is therefore likely that these areas have transcription error and omissions. Using Derbyshire, the subject of the pilot study, as an example, 1 RSD (Tideswell in Bakewell) had no employers at all, 1 RSD in Chesterfield had a ratio below 0.1%, another part of Bakewell was at 0.1%, 2 other RSDs in Chesterfield were at 0.2% and a further 9 RSDs had a ratio between 0.2 and 0.5%.

There are some counties which consist fully or nearly fully of over-0.5% (light blue) RSDs. An analysis of these counties shows they are mainly in line with the expected numbers of employers based on 1861 and 1881. These include Brecknockshire, Cambridgeshire, Huntingdonshire, Lincolnshire, Rutland, Suffolk, and Yorkshire East Riding. In some additional counties, the missing areas are small and/or concentrated in urban areas, so there is no problem with these for farmer-only analysis. These counties include: Bedfordshire, Essex, Kent, Oxfordshire, Warwickshire, Wiltshire, Glamorgan, Flintshire, and Merionethshire.

Unfortunately, the GRO in 1871 did not create tables of employers with their employees on county, division, and national level as they did in 1851, so no detailed comparisons with published data were possible. However, they did tabulate some limited analysis for farmers only.

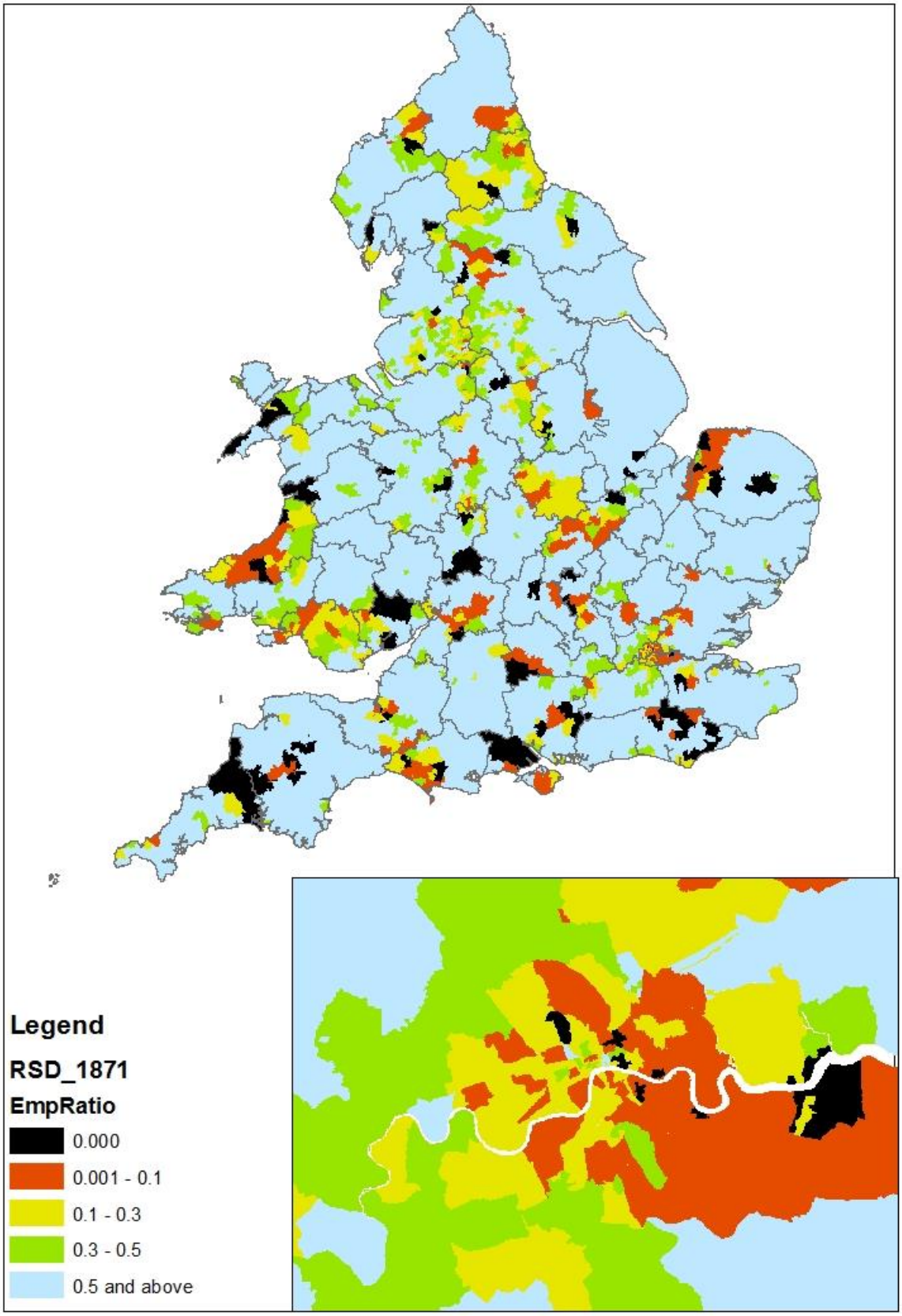

Figure 1: Employer-ratio in 1871 in England and Wales.

4.2. Farmers

The GRO performed a limited analysis of farmer employers and their employee numbers. This was conducted on a group of 17 'representative' counties in England only, namely Surrey (Extra-Metropolitan), Kent (Extra-Metropolitan), Sussex, Hampshire, Berkshire, Essex, Suffolk, Norfolk, Leicestershire, Rutland, Lincolnshire, Nottinghamshire, Derbyshire, Durham, Northumberland, Cumberland, and Westmorland.

The published figures can be compared with the S&N extraction for the same 17 counties. It should be noted that while some of these are amongst the better-transcribed counties when it comes to employers (e.g. Lincolnshire, Suffolk), there are also some of the poorer ones, such as Hampshire.

The published report for the Census 1871 shows that of the farmers who employed 1 or more labourers, the average number of employees was less than 6.[4] The average number of employees in the extracted data is 7.8, an indication that the extraction picks up more large rather than small farms. This is also evident in the breakdown by size between the published and the extracted farms, as shown in Tables 6 and 7. Since the extraction method picked up the farmers without employees through a different algorithm, the results for farmers with zero employees have not been included in the total in Table 7.

[4] 1871 Census England & Wales, General Report, 1873 [872-I] Vol LXXI Part II, p. xlviii

employees	Surrey		Kent		Sussex		Hampshire		Berkshire		Essex		Suffolk		Norfolk		Leicestershire	
	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N
0	299	247	934	871	962	987	721	687	254	228	400	310	753	985	1692	2197	1220	698
1	130	86	428	278	423	150	295	76	131	62	281	141	579	318	879	422	502	127
2	142	73	454	312	422	196	285	111	164	58	389	191	548	333	625	391	386	119
3	126	81	347	234	300	189	230	107	128	61	307	185	468	313	462	326	272	94
4	88	80	294	211	235	153	170	83	103	57	279	151	434	290	333	240	216	80
5	73	75	219	198	206	107	153	87	98	49	264	168	300	245	236	190	118	76
6	64	46	234	191	193	107	134	60	103	44	214	153	281	226	226	160	102	40
7	60	50	129	133	130	98	104	61	71	43	162	150	219	179	171	137	59	29
8	55	42	157	139	125	98	137	61	112	53	199	118	220	167	183	134	41	21
9	35	32	114	102	84	72	75	51	52	35	145	114	132	140	124	108	25	23
10-	105	129	329	379	259	256	263	188	200	146	469	500	417	475	415	360	74	45
15-	37	51	132	180	87	128	84	101	89	98	209	255	163	220	193	175	12	16
20-	16	30	89	96	52	76	52	71	57	69	131	153	69	114	85	122	3	7
25-	6	12	44	76	21	30	30	24	20	33	42	87	37	59	53	54	2	4
30-	3	9	26	46	17	30	10	20	10	17	31	50	17	46	33	36		1
35-	1	6	10	23	4	12	2	11	2	26	13	34	11	23	15	20		3
40-	2	3	12	16	4	4	3	4	6	11	16	18	9	12	6	11	1	1
45-		1	5	9		7		3	3		4	14	2	11	7	7		
50-	1	1	6	5	3			1		2	5	8	3	5	1	6		
55-			3	3	2	3			1	3	4	8	2	4	4	5		
60-		2	11	20	3	4	1			3	10	20	6	12	5	12		
total	1243	**1056**	3977	**3522**	3532	**2708**	2749	**1807**	1604	**1098**	3574	**2828**	4670	**4177**	5748	**5113**	3033	**1384**

employees	Rutland		Lincolnshire		Nottinghamshire		Derbyshire		Durham		Northumberland		Cumberland		Westmorland	
	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N	Pub	S&N
0	196	157	5144	4940	1923	1411	4068	2661	2412	1679	1268	1017	1610	2644	1762	1677
1	91	36	1251	747	503	213	650	383	303	133	206	118	385	242	161	107
2	65	33	969	681	400	222	354	241	302	167	273	175	322	219	107	97
3	52	45	614	500	220	163	204	190	178	107	231	120	166	123	45	49
4	39	35	444	363	184	120	84	109	104	53	182	121	95	81	33	30
5	24	32	326	299	86	87	31	67	47	44	92	86	54	45	11	12
6	13	17	255	263	77	79	11	30	35	24	102	64	22	25	11	14
7	9	21	181	224	54	66	7	23	22	15	52	45	14	15	6	2
8	11	19	189	164	40	51	7	21	28	16	65	54	17	11	3	6
9	7	7	115	138	23	32	2	5	7	16	32	34	4	4	1	4
10-	22	21	370	491	74	116	11	18	15	23	89	109	15	25	2	6
15-	7	16	112	222	14	31	2	8	3	8	34	40	1	5	2	3
20-	2	2	67	123	2	16	1	3		4	13	26	2	3		
25-		2	27	50		4	1	1	1	4	3	16	1			
30-			10	24		2		3		3	4	15				
35-			4	17				1			4	7				1
40-		1	9	19	1		1				1	4		1		
45-		1	3	5	1	1	1					1		1		
50-				5	1	1					1	1				
55-			1	2								1				
60-			2	6		1				1	2	4				
total	538	**445**	10093	**9283**	3603	**2616**	5435	**3764**	3457	**2297**	2654	**2058**	2708	**3444**	2144	**2008**

Table 6: Published vs extracted S&N farm data by number of employees by county.

Table 7 shows a total under-estimate of 25% of farmers with employees. However, these are not evenly distributed among the farm sizes. Small farms, with fewer than 10 employees, are up to 49% down, while the larger farms are over-counted. In order to check whether these are transcription errors or reflect under-counts by the GRO's clerks, additional checks were performed on the largest categories of farms, those with 50 or more employees, for the 17 counties only.

	Total 17			
employees	Published	S&N	**Difference**	**%Difference**
0	25618	23396	-2222	-8.7
1	7198	3639	-3559	-49.4
2	6207	3619	-2588	-41.7
3	4350	2887	-1463	-33.6
4	3317	2257	-1060	-32.0
5	2338	1867	-471	-20.1
6	2077	1543	-534	-25.7
7	1450	1291	-159	-11.0
8	1589	1176	-413	-26.0
9	977	917	-60	-6.1
10-	3129	3287	158	5.0
15-	1181	1557	376	31.8
20-	641	915	274	42.7
25-	288	456	168	58.3
30-	161	302	141	87.6
35-	66	184	118	178.8
40-	71	105	34	47.9
45-	26	61	35	134.6
50-	21	35	14	66.7
55-	17	29	12	70.6
60-	40	85	45	112.5
Total (excl 0)	35144	26212	-8932	-25.4

Table 7: Published vs S&N extracted farm data by number of employees for the 17 counties.

While the largest farms, those with 70 employees or more, had already been checked as part of the general cleaning process, the additional cleaning of the 50-70 employees category for these 17 counties revealed an additional 26 farmers who had been wrongly transcribed. This is mainly due to the 'ing' part of 'employing' in the text string being mistaken for either a 4, 7 or 9 (the main reason for cleaning down to 70 employees), or the '&' in a descriptor of 'employing X men & X boys' being mistaken for a 3, 4, or 5. While a proportion of the large

firm over-counts is therefore an inherent error in the data source, a greater proportion is due to oversight of the GRO's clerks. After careful checking of all farms of over 50 employees, 149 were found in the extraction, where the GRO only listed 78 – roughly half of the farmer employers present in the data (see Table 8).[5] Taking into account that the extracted sample includes counties from which parts are known to be missing, the 'real' number of large farmers is likely to be even higher, implying that the GRO missed over half the large farms in the census returns. For instance, looking at Essex and Suffolk, some of the better transcribed counties, half of the farms in the larger categories are missing in the published totals. Similarly, many of Norfolk's known large farms are missing in the published records, and since this county includes some badly transcribed areas, there are likely to be more.

Number of Employees	**Published**	**Original Extraction**	**Cleaned Extraction**	**New difference**
50-	21	50	35	14
55-	17	34	29	12
60 and over	40	91	85	45
Total	**78**	**175**	**149**	**71**

Table 8: Published vs S&N extracted farm data after further checks.

Similar checks further down the farm size scale were infeasible due to numbers, but the implication of the larger farm checks is that the extraction and GRO possibly miss out a large number of small farms as well.

4.3. Acreage

In addition to their employee data, the GRO also reported the acreages for the farmers in the 17 counties. This included both Group 1 farmers, since those who stated their employees often also stated their acreage, as well as Group 5s (farmers with acreage but no listed

[5] The entrepreneurship project considered landowners with labourers as farmers, which the GRO probably did not. However, of the 149 only 7 did not explicitly use 'farm' in their descriptor, so this cannot account for the difference. In addition, another 16 were farmer as their second occupation (with the first being either landowner or magistrate). It is not clear whether the GRO included them, although their 1851 methodology suggests they did. Even so, the remaining 48 'missing' farmers were all of the standard 'Farmer occupying X acres employing Y employees' formula, and should have been picked up under any definition of farmer.

employees). They calculated that the average farm size in these counties was 152 acres, with more than a fifth of farms occupying less than 20 acres.[6] This corresponds exactly with the extracted data, which has an average of 152.4 acres.

Acres	Published	Extracted	Difference	%Difference
Under 5	1984	1198	786	39.6
5-	4017	2845	1172	29.2
10-	6074	4731	1343	22.1
20-	4193	3416	777	18.5
30-	3363	2795	568	16.9
40-	3048	2597	451	14.8
50-	6370	5401	969	15.2
75-	4113	3486	627	15.2
100-	7341	6351	990	13.5
150-	4706	3834	872	18.5
200-	3927	3205	722	18.4
250-	2324	1902	422	18.2
300-	2226	1819	407	18.3
350-	1166	947	219	18.8
400-	1824	1493	331	18.1
500-	1098	875	223	20.3
600-	666	565	101	15.2
700-	390	304	86	22.1
800-	270	218	52	19.3
900-	188	159	29	15.4
1000-	249	186	63	25.3
1200-	159	118	41	25.8
1500-	84	61	23	27.4
2000upwards	90	85	5	5.6
Total	**59870**	**48591**	**11279**	**18.8**

Table 9. Published vs extracted farm data by acreage.

[6] 1871 Census England & Wales, General Report, 1873 [872-I] Vol LXXI Part II, p. xlvii

Table 9 shows the comparison between the published and extracted acreages broken down by farm size. The 19% missing acreage overall corresponds with the estimated missing people due to transcription error, but it is interesting that these are distributed more evenly across the farm sizes than the missing employers, and that there are no over-counts. This indicates that the GRO clerks were better at extracting acres than employer numbers. Here as well though, the largest percentage of missing farms are the smallest ones, which is proportionally twice as much as the overall missing acreage. This may indicate that enumerators or the GRO processes excluded some very small operators and small holders, but the actual process GRO used is not fully documented.

In total therefore, the extracted data is missing small farms, both in acreage as well as employee size, as measured against the GRO data. But the GRO also missed about half of large farms.

5. Alignment

With the available resources, the missing data in the 1871 S&N records cannot be added since this would require the original CEBs to be thoroughly searched for all counties. However, there are some possible fixes to weight the available data to reconstruct the database.

5.1. Employers and farmers

The GRO data introduced in section 4 provides an additional source of information: it gives the number of farmers without acres or employees for each of the 17 counties. Although this will not be fully accurate, it makes possible construction of a proportional breakdown between E, OA, and W farmers for these counties. This is shown in Table 10.

published	**%**
E	49.3
OA	35.9
W	14.8

Table 10: Proportional breakdown farmer type in the published 17 counties.

The total number of farmers for the whole England and Wales was published by GRO: 249,907. Applying the breakdown of the representative 17 counties, this gives estimates that there were 123,209 employer farmers, 89,813 OA farmers, and 36,885 worker farmers.[7] Comparing this to the extracted data gives us 75,691 E farmers, and 95,633 OAs. This shows two things: first, many of the OA farmers were wrongly transcribed, giving only their acreage and not their employees. However, this can be adjusted by using the farm reconstruction model (outlined in WP 9). Second, the extracted total E+OA is 19.6% down on the estimated totals, confirming that the data should be weighted at national level in the range of 19-20%. Alternatively, as Table 7 shows, while the farmers with employees are down 25.4%, the farmers without employees are only down by 8.7%. In employer-only analysis, the E farmers can be weighted more heavily, while the OA farmers should be given a smaller weight.

5.2. The farming model

The farm reconstruction model used more generally over 1851-81 (see Working Paper 9) seeks to compensate for non-responses to the census questions to farmers to give employee numbers. For 1871 to be aligned to the general reconstruction estimates the S&N data need to be weighted to compensate for the areas with missing data. To develop these weights, five sub-samples of increasing quality were identified to assess the parish-level data using the employment ratio of RSDs (0.1%, 0.3%, 0.5%), and the counties where transcription was judged to be good. These can be divided into locations that appear to have good agricultural responses (parishes from the 9 counties identified in section 4.1 where the missing transcription areas were small or confined to urban areas – called the 'good' counties), and those that look complete overall (parishes from the 7 counties identified in section 4.1 that seemed to have been transcribed completely – called the 'very good' counties). The better the extracted dataset, the less is the need to weight.

The way of calculating the weights uses the inverse of the number of observations of the size of the sample under consideration. The samples were ordered in decreasing quality so as to

[7] For 1851, both the 17 counties and the whole E&W are available. A comparison of this breakdown shows that the counties are not completely representative with the E and OA slightly up and the workers underestimated. This is probably because the 17 counties skew towards southern counties, and exclude Wales. Extrapolating the 17 counties up to E&W therefore might slightly overestimate the E and OAs, but these were likely underreporting themselves.

pick higher weights as one goes down the list ordered by the decreasingly quality of the data. The weights correct for the reduced number of farmers caused by poor transcription. Consequently, the full sample has higher weights for the smaller and highly accurate samples down the following list:

Sample	No. Parishes	Regression Sign
Whole	12,600	-
0.1%	10,900	-
0.3%	9,300	-
0.5%	7,800	-
Good	3,700	+
Very good	1,900	+

The list shows the consecutive samples from the whole population consisting of 12,600 parishes. A test is used to indicate the effects of weighting. This estimates the Acreage coefficient as an independent variable in a cross-section regression with the log of the ratio of Employers and Own account as the dependent variable. The coefficients have a worrying negative sign for the Acreage coefficient for most of the sample, and for the entire sample without weighting. The coefficient changes to the expected positive sign only for the parishes with good and very good extractions. This demonstrates that even parishes with a fairly good extraction ratio (of 0.5%) weighting is essential to obtain valid results.

As a summary, we have created a weighted sample where the more accurately extracted locations are weighted more, which permits to use the best part of the sample, but preserves the value of the rest of the sample but using smaller weights and using as much as data as possible from the 1871 extraction.

6. Conclusion

This working paper has laid out the procedures of collecting and processing the 1871 data as deposited by the ESRC Entrepreneurship project. The data align mostly with the extracted I-CeM data from the other early census years, however, there are some important differences due to data constraints. These mostly relate to geographical pockets where the employers

with employees have been badly transcribed, leading to a national deficit of an estimated 20% of employers. There is also a the lack of farmers without employees or acres due limitations on finance that restricted extraction of all S&N data. In addition, the acquired 1871 data only covers entrepreneurs who could be identified and extracted through their occupational string. Without access to the full economically active population of 1871, a reconstruction of all Employers and Own-Account as discussed in WP 9 is not possible.

The paper has also evaluated the extracted data against the figures published by the GRO. Since the actual process the GRO used is not fully documented, it is difficult to assess our data against their tabulations. Comparisons confirm that around 20% of the farmer employers are missing. However, it has also become clear that the GRO's numbers of farmers are deficient. In the case of the largest farms in their 17 selected counties, at least half of the farms recorded in the CEBs were missing from the GRO's report.

Acknowledgements

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).**

Additional support for the coding of the 1871 census data derived from part of the ESRC project, with additional support for data coding and cleaning by Joe Day from Isaac Newton Trust research grant 17.07(d): **Business Employers in 1871**.

We are especially grateful to S&N for making the data available for this research, and helping so fully in the data extraction process.

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 authors, have collectively worked on the new versions of I-CeM.

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*.

WP 6: Smith, Harry J. and Bennett, Robert J. (2017) *Urban-Rural Classification using Census data, 1851-1911*.

WP 7: Smith, Harry, Bennett, Robert J., and Radicic, Dragana (2017) *Classification of towns in 1891 using factor analysis.*

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*.

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

WP10: Bennett, Robert, J., Smith, Harry and Radicic, Dragana (2018) *Classification of environments of entrepreneurship: Factor analysis of Registration Sub-Districts (RSDs) in 1891*.

WP11: Montebruno, Piero (2018) *Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911*.

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