# Reconstructing business proprietor responses for censuses 1851-81: a tailored logit cut-off method

**Robert J. Bennett, Piero Montebruno, Harry Smith, and Carry van Lieshout**

rjb7@cam.ac.uk pfm27@cam.ac.uk hjs57@cam.ac.uk cv313@cam.ac.uk

Working Paper 9.2:
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 2019

[Extension of analysis of WP 9 of February 2019]

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**

# Reconstructing business proprietor responses for censuses 1851-81: a tailored logit cut-off method

**Robert J. Bennett, Piero Montebruno, Harry Smith, and Carry van Lieshout**

Working Paper 9.2: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge.

## 1. Introduction

This paper extends the reconstruction method developed in WP 9 to identify entrepreneurs 1851-81. Its aim is to identify the *individual* employers and own-account business people for 1851-1881, where employment status was not explicitly identified in the population censuses. This paper develops a method of variable logit cut-offs tailored to each occupation code. This allows the original census responses can be supplemented to give approximately all employers and own account. The aim is to provide a further resource for subsequent researchers, which is available in the database deposited at UK Data Archive (UKDA) as the *British Census of Entrepreneurs 1851-1911* (BBCE), derived from the ESRC-supported project ES/M010953 *Drivers of Entrepreneurship and Small Businesses*. An overview of the project is provided in WP 1, which shows how 1851-81 differed from the subsequent censuses. WP 2 defines in detail the different censuses and the challenges they present for identifying entrepreneurs. WP 3 describes how the data for 1851-81 were extracted from the censuses from different sources. A full list of Working Papers is given at the end of this paper. The method described in this paper is extended to England and Wales and can thus be compared with the results from WP 9; the method here is sole method applied to Scotland (see WP 20).

As in WP 9 the source used is transcripts of the census, mainly as they are entered into the I-CeM electronic database for the censuses 1851-1911 produced by a team at the University of

Essex, deposited at the UKDA: *The Integrated Census Microdata (I-CeM).*[1] Also used are infills of gaps of truncations in I-CeM. For 1861 this has come from clerical inspection of the Census Enumerators Books (CEBs); this infilled about 28,000 employers. Other infills have been added from a separate source (S&N: The Genealogist) for 1851 (about 53,000 individuals, and 26% of all employers), and for 1871 (100% of selected records). For 1871 the census records available are restricted in the project to those extracted from Groups 1-6 (as defined below); employers, masters and farmers as described in WP 1. The 1871 data provide invaluable sources for the intelligence-led approach but are not available for the whole population and hence cannot be used directly in the reconstruction process developed here (but do inform the analysis).

The sources provide transcriptions of the original CEBs as well as enhancing the data with various codes for household structure, and occupations. Within the occupational data is information on employment status: whether individuals are *employers* (those who employed others), *sole proprietor own account self-employed* (who employed no-one else and described themselves as 'masters' or similar), employees, or the unoccupied.

As in WP 9, the aim is to supplement employer and own account information for 1851-81 where many were not explicitly recorded. Although the 1851-81 censuses identified 'employers' and 'masters', this was only partial for many occupations since there was no explicit identification of 'own account', or explicit distinction from other statuses such as worker, unoccupied, etc. The aim is to reconstruct these missing responses as accurately as possible, within specified probability bounds with definitions aligned as far as possible, between the data for 1851-81 and the later census series for 1891-1911. This paper seeks to extend WP 9 by improving on accuracy of identification for *individuals* (rather than aggregate numbers). However, even in this version the results at the *individual level should be used only with care*, since validations can usually only be made for aggregates not individuals.

The background to the cleaning and screening of the census data is given in WPs 1, 3, 4, and 9. This gives the methods of extraction of those identified as 'employers' 1851-81 from the

[1] 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).

occupational strings, and the identification, cleaning, screening and correction of all employment statuses 1891-1911. The background to what the different censuses covered is provided in previous working papers (primarily WP 2). This paper provides an alternative reconstruction method. The paper first summarises what WP 9 contributed (Section 2). Section 3 outlines the alternative method developed which is based on variable logit cut-offs and discusses its outcomes; Section 4 gives an overview of the main trends and implications of the alternative compared to the original WP 9 estimates. The estimates for each individual using the new version developed here and from WP 9 are available in the *British Census of Entrepreneurs 1851-1911* (BBCE) database at UKDA, planned for deposit later in 2019. The detailed decision made for each sector/occupational category is recorded in a separate data download mounted with this paper and WP 9. Full information on how the method was implemented is given in WP 19, which includes the estimation methods used in both WPs 9 and 9.2, with downloads of the estimation process, for both England and Wales, and Scotland.

The proprietors identified by both the intelligence-led and tailored cut-offs are available by I-CeM RecID in the UKDA deposit of BBCE as:

- EMPSTATUS_IND (for the outputs derived from this WP 9.2), and
- EMPSTATUS_NUM (for the outputs derived from WP 9).

## 2. Reconstruction: overview of intelligence-led methodology in WP 9

### *2.1 Census questions and data extraction*

The core of the reconstruction method for England and Wales starts with those that can be reliably identified and extracted from the census instructions that respondents followed. The method was not applied to Scotland. The 1851 census instruction was:[2]

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

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

In 1861 the wording was widened:

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

Similar instructions were used for 1871-81. For farmers a question asked for workforce and information on the 'number of acres' occupied. For 1851, which was almost identical to 1861-81, this was:

> 'The term FARMER to be applied only to the occupier of land, who is to be returned – "*Farmer of [317] Acres, employing [12] labourers*;" the number of acres, and of in or out-door labourers, on March 31st, being in all cases inserted.'

Census responses to these questions resulted in extended occupational descriptor strings which can be used to identify each employer and their workforce using algorithmic searches and parsing, supplemented by significant clerical checking.[3] See also WP 3.

The individuals identified directly from these instructions are referred to as those *extracted* from the census responses. As outlined in WPs 1 and 3, the extractions derive from searching for the census and similar terms in the descriptor strings for employers and others to identity their workforce, acres (for farmers), and 'master' or synonyms and equivalents. This results in six extraction Groups:

> **Group 1:** ***all employers and others (such as masters, proprietors or owners) stating employees***; farmers and non-farmers;
>
> **Group 2:** ***all stating 'employer' but with no stated employees;*** Group 2 is small and possibly contains transcriber and algorithmic identification errors. Farmers in Group 2 were accepted as own account; but non-farmers were identified by the alternative reconstruction methods.
>
> **Group 3:** ***master etc.*** anyone stating 'master' as a business in their occupational descriptor with no employees were identified as own account.
>
> **Group 4:** ***'farmer'*** not stating employees or acres; assumed to be workers.
>
> **Group 5:** ***farmer giving acres*** but no stated employees. Those over two acres were split between employers and own account using their declared acreage to estimate if it

[3] For piloting of the methodology see: Bennett, R. J. and Newton G. (2015) Employers and the 1881 population census of England and Wales, *Local Population Studies*, 94, 29-49.

was large enough normally to require employees.[4] Those with less than two acres were assumed to be smallholders working on other farms. The employers identified from Group 5 were added to the employers from Groups 1 and 2 to give all farmer employers; the rest of Group 5 were assigned as own account. This results in all farmers being identified directly from the census descriptions.

**Group 6:** ***owners or proprietors of business assets***: mine/quarry owner, shipowner, barge owner and others with any business assets (excluding land/house owner). All identified as proprietors even if they state no employees; attribution as employers or own account based on descriptor strings where possible (such as coal owners/masters), otherwise by the alternative reconstruction methods.

No attempt was made to take account of specific 'partner' or 'director' information in reconstruction; they were all treated as separate proprietors identified as employers or own account through the rest of the reconstruction. If the true status of directors in I-CeM coding was as a worker or not occupied they were identified subsequently by record-linkage with director directories (see WP 14). 'Partners' and 'directors' are not explicitly considered further in this paper for reconstruction.

### *2.2 Reconstruction used for intelligence-led method*

The reconstruction method used in WP 9 is based on using the characteristics of the 1891 and 1901 census responses to supplement responses to the 1851-81 censuses. This is only possible after improving the quality of the 1891-1901 data to process the census data in I-CeM to compensate for non-response and misallocation bias. This requires use of weighted data for 1891 and 1901. The limitations of the 1891 and 1901 censuses and the weighting method are given in Bennett et al. (2019a); the weights are given as a supplement to WP 11.

The reconstruction method used in WP 9 follows six stages:

(1) data preparation through extensive data cleaning, re-coding occupations in I-CeM to their entrepreneur status, and development of 83 additional Sub-Occodes to supplement the occupations codes in I-CeM (OCCODE) to better separate individuals

[4] This uses the methodology for allocation based on a logit model described fully in WP 9 and Montebruno et al., 2019a.

within complex groups with high variance between employer, own account and worker status.

(2) Development of logit regression estimators for combined employers and own account based on the 1891 and 1901 censuses where 'employment status' (as employer, own account and worker) was information explicitly collected. The details of the logit estimator are given in WP 9 Appendix.

(3) The logit model was then used to estimate reconstructions by swapping the coefficients for 1891 (mainly) and 1901 (for some Sub-Occodes) with the 1851-81 data; i.e. using the 1891 or 1901 statistical characteristics of individuals to identify the same type of individuals from the data in the earlier censuses and assign them appropriate employment statuses.

(4) Comparisons against alternative estimators for 1851-81 using ratios of entrepreneurs in 1891-1901. Final choice of reconstruction method to give reconstructed aggregate numbers.

(5) Assignment of individuals as entrepreneurs or workers; and

(6) Repeating the process to assign identified entrepreneurs between employer and own-account status.

Comparisons were made against secondary material in published literature and directory sources at each stage in order to test if the estimates were matching known trends. As a result, in WP 9 we refer to this approach as *intelligence-led*. The 1871 data from S&N provide invaluable additional resources for employers and farmers that are used in the intelligence-led approach (especially for farmers), but they are not available for the whole population (i.e. all workers and some entrepreneurs were not available), and hence 1871 data in BBCE cannot be used directly in the reconstruction process. The intelligence-led comparisons were undertaken for each Sub-Occode. The 83 additional Sub-Occodes when added to the original I-CeM OCCODE to increase the number of occupational categories from 761 to 844. These are listed in the Appendix to WP 9. The Sub-Occodes allow a fine mesh and targeted means to assess the issue at a sector micro level.

The reconstruction process should be regarded as *a means to supplement the census for non-responses* because the information we require was only partially determined from the questions asked in the census instructions. It is therefore akin to a method of post-survey non-response adjustment.

The estimate chosen after the six stages outlined above for each Sub-Occode was then compared against the number of extracted in each Sub-Occode. Once the preferred supplementation is decided, the *extracted are used as the base which is supplemented* using the estimated reconstructed total of entrepreneurs in each category of employer and own account. The reconstruction estimates are thus used in two ways: first, to assess how far the extracted are complete; and second, to provide post-survey estimates of the individuals who need to be added to the extracted as supplements to give a full response group *as if* the census had fully gathered the data.

Bounds can be estimated for the estimates using the range of the various estimators. This uses the properties of the logit estimator, which gives a probability of being a proprietor for each individual in the whole population. These probabilities have a range from zero (for those estimated to have no probability of being an entrepreneur), to one (where individuals are certainly proprietors). In practice probabilities have decimal values between the 0-1 extremes which offer two alternatives which act as bounds to the final estimator chosen. *Unrounded*: the actual decimal values; and *Rounded*: decimal values rounded up or down to 1s or 0s (i.e. as a proprietor or not). The logit model used in the intelligence-led reconstruction adopts a value of 0.5 as a cut off for determining entrepreneur status. This is the standard method used in logit estimation: above 0.5 an individual has a probability of more than half of being an entrepreneur; below 0.5 an individual has a probability of more than half of being a worker. The values between the unrounded and rounded give the respective upper and lower bounds of the logit estimates.

## 3. Alternative reconstruction method: tailored cut-offs

The intelligence-led approach used in WP 9 predominantly uses unrounded logit estimates but also employs rounded logit, and also extrapolation ratio estimates. These have the

significant drawback that unrounded logit and ratio estimates give only the proportion within the group that was entrepreneurs, not the individuals themselves. The rest are identified in WP 9 using random assignment. This is *acceptable for tracking the aggregates* most accurately - as an intelligence-led approach informed by secondary and external sources. But it is *unsatisfactory for identifying the individuals themselves*.

An alternative approach is to use a tailored logit that exploits the range between the rounded and unrounded estimates to tune the logit cut-off to the 1891 estimates. This was re-applied to England and Wales, and was the sole method used in Scotland. The same logit estimator was used as in WP 9. This was then applied in two ways. The first took the cut-off which best predicted the 1891 actual numbers in each sub-occode when the 1891 model was used to predict 1891 entrepreneur numbers. The second used the cut-off which gave a total closest to the number of entrepreneurs in that sub-occode if the 1891 ratio between entrepreneurs and workers was maintained in 1851-81. In both cases, as with the intelligence-led reconstruction, choices were guided by the 1891 data: the choice between each comparator was determined by whether the trend produced was sensible. This *tailored cut-off* approach allows individuals to be identified and hence is better suited for further analysis at the micro-level. There is no random assignment. The logit tailored cut-offs ranged from 0.1 to 0.8. This broad range illustrates that in some occupational groups a rather low level of probability is necessary to identify the most likely individual proprietors that relate to the 1891 characteristics, whilst in other cases a higher level is needed.

The different logit cut-offs are summarised in Table 1. Only 21% of Sub-Occodes have cut-offs at the 0.5 level and above, whilst 63% are at 0.25 of below. This is a strong indication of how the characteristics of proprietors identifiable from their demography and other personal attributes are quite widespread in some occupational groups. It is also a facet of the strong skew of the distribution. In all cases the probability distribution approximates a log normal, with very large proportions of very low probability values, declining rapidly as probabilities increase. In such cases a small difference in probability between 0.1 and 0.15, for example for shoemakers, innkeepers or grocers, most other maker-dealers, refreshments, and retailers, makes a very large difference in the numbers identified as proprietors. Indeed, it is these categories, and maker-dealers in particular, that are the most difficult to differentiate between proprietors and workers, with small changes in assumptions making large changes in the number of people estimated as needing inclusion in supplementation.

| Logit cut-off | Number | | | % | | |
|---|---|---|---|---|---|---|
| | 1851 | 1861 | 1881 | 1851 | 1861 | 1881 |
| 0.1 | 315 | 315 | 314 | 37.5 | 37.5 | 37.4 |
| 0.15 | 70 | 69 | 71 | 8.3 | 8.2 | 8.5 |
| 0.2 | 90 | 89 | 88 | 10.7 | 10.6 | 10.5 |
| 0.25 | 55 | 56 | 55 | 6.6 | 6.7 | 6.5 |
| 0.3 | 41 | 40 | 42 | 4.9 | 4.8 | 5.0 |
| 0.35 | 37 | 38 | 37 | 4.4 | 4.5 | 4.4 |
| 0.4 | 28 | 28 | 28 | 3.3 | 3.3 | 3.3 |
| 0.45 | 24 | 23 | 25 | 2.9 | 2.7 | 3.0 |
| 0.5 | 39 | 41 | 38 | 4.6 | 4.9 | 4.5 |
| 0.6 | 37 | 37 | 36 | 4.4 | 4.4 | 4.3 |
| 0.7 | 73 | 69 | 76 | 8.7 | 8.3 | 9.0 |
| 0.8 | 30 | 34 | 30 | 3.6 | 4.1 | 3.6 |

**Table 1.** Cut-offs of 1891 logit estimates for *all proprietors* that best fit 1851-81: number of each Sub-Occode and percentage of all 840 Sub-Occodes containing entries.

The situation is different for employers, as shown in Table 2. Whilst there is still a large proportion with cut-offs less that 0.25 (over 26% of Sub-Occodes), and there is still a skew towards low probability values, a much large proportion has cut-offs over 0.5 (ranging from 46% to 49% per year). In general, the characteristics of employers identifiable in the census from their personal attributes are much more definitive than for all proprietors. In particular employers have a strong association with being a household head, whilst own account proprietors have very varied relationships within households and are much less commonly heads.

| Logit cut-off | Number | | | % | | |
|---|---|---|---|---|---|---|
| | 1851 | 1861 | 1881 | 1851 | 1861 | 1881 |
| 0.1 | 140 | 145 | 148 | 19.5 | 19.9 | 20.6 |
| 0.15 | 15 | 14 | 15 | 2.1 | 1.9 | 2.1 |
| 0.2 | 15 | 20 | 15 | 2.1 | 2.7 | 2.1 |
| 0.25 | 20 | 30 | 23 | 2.8 | 4.1 | 3.2 |
| 0.3 | 27 | 27 | 33 | 3.8 | 3.7 | 4.6 |
| 0.35 | 37 | 37 | 33 | 5.2 | 5.1 | 4.6 |
| 0.4 | 44 | 53 | 63 | 6.1 | 7.3 | 8.8 |
| 0.45 | 63 | 45 | 49 | 8.8 | 6.6 | 6.8 |
| 0.5 | 63 | 59 | 56 | 8.8 | 8.1 | 7.8 |
| 0.6 | 130 | 107 | 121 | 18.1 | 14.7 | 16.8 |
| 0.7 | 85 | 95 | 89 | 11.8 | 13 | 12.4 |
| 0.8 | 79 | 93 | 75 | 11.0 | 12.8 | 10.4 |

**Table 2.** Cut-offs of 1891 logit estimates for *employers* that best fit 1851-81: number for each Sub-Occode and percentage of all Sub-Occodes containing entries.

## 4. Overview of the main trends and implications

This section discusses the new estimates for reconstruction for England and Wales. The cut-off method was also applied to Scotland (as the sole method used), and discussion of these estimates is given in WP 20.

### *4.1 Range of estimates*

Figure 1 compares the estimates using the intelligence-led decision rules (Supplement 1) and the alterative using the tailored logit cut-offs (Supplement 2); in all cases farmers are not supplemented but fully estimated using the extraction Groups (as outlined under Groups 1, 2, 4 and 5 above). In addition, the figure plots the bounds of the estimates: the range is shown

by plotting estimates from alternatives using the logit estimates (based on 1891 and 1901 data) for both rounded and unrounded probabilities, and also the estimates from extrapolated ratios.[5] Estimates are shown separately for employers, own account, and all proprietors (the sum of the two). Also shown below are the numbers of employers and all proprietors identified from the census extractions alone. The figure makes clear that the unrounded estimates are usually higher than rounded, because unrounded include very many individuals even though most have low probabilities of being a proprietor. The unrounded estimates count a probability of 0.1 although only a 'part likelihood of being a proprietor' is counted entirely but there are large numbers of them. The rounded probabilities count only those values above the cut-off of 0.5. The range between the two usually gives the outer bounds of our supplementation estimates: from those with a remotely small probability of proprietorship (unrounded) to those who were more certainly proprietors (rounded). However, quite a few rounded estimates are higher than unrounded, notably any sub-occode where entrepreneurs are a high percentage of that sub occode in 1891.

Five features stand out. First, the scale of supplementation required is considerable. The bottom line in Figure 1 for extracted employers is just under one half of the supplemented numbers in 1851, and the line for all extracted proprietors is only about one third of the supplemented estimates. The scale of under-response also seems to have become larger with every year up to 1881, mostly as a result of the rapid expansion of the own account proprietors that are more poorly recorded and more difficult to identify in the census; in comparison, employer numbers change relatively slowly and response rates remain similar 1851-81.

[5] The loss of about 3.7% of records for 1861 that are no longer at TNA and hence not in I-CeM is compensated by weighting for that year.

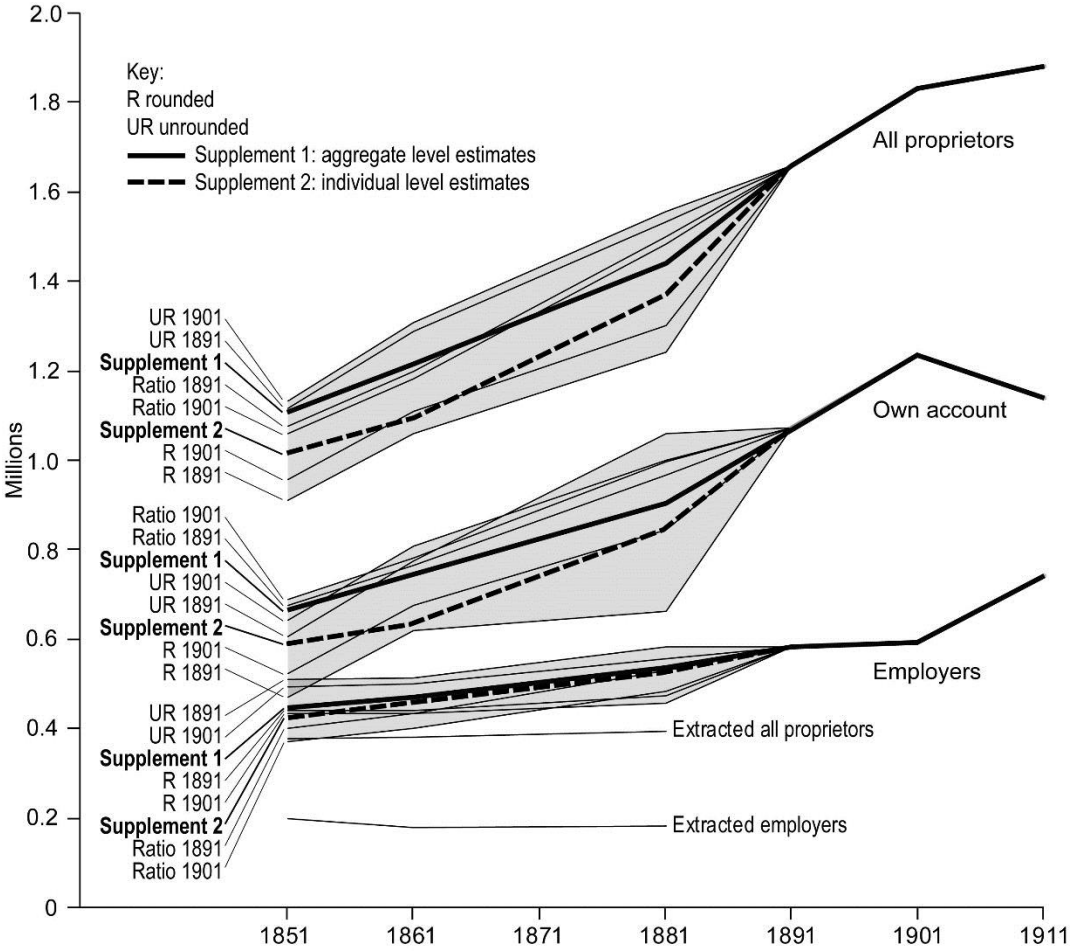


**Figure 1.** Reconstructed estimates of entrepreneur numbers for employers and own account 1851-81, with 1891-1911 from later census weighted estimates; estimates show the range of alterative estimators used and the chosen estimate, compared to the census extractions for employers and own account 1851-81; (1861 is weighted for data loss; and estimates have been straight-line extrapolated across 1871 where data are unavailable in I-CeM, and the S&N additional data cannot be used for fully reconstructing all individuals).

Second, the census extractions capture a slow rate of increase in own account numbers and slight decrease in employers 1851-61 which feeds into the various supplementation estimates. Third, the decision rules for the intelligence-led approach lead to a convergence towards similar estimates of numbers of proprietors, within ± 10 per cent. This means that whatever assumptions are made, the general trends identified from the respondents' actual descriptors when supplemented are robust against the main alternative estimates. Fourth, the range between the rounded and unrounded estimates (i.e. between certain entrepreneurs (lower

bound) improbable entrepreneurs (upper bound)), are much narrower for employers than own account, as a result of the extracted Groups in the census being better and more complete guides for employers. Fifth, the alternative estimates based on tailored cut-offs are always lower than the intelligence-led aggregated estimates, 5-10% lower in numbers overall, and mainly differ for own account. However, it is important to note that although the numbers are lower, the overall trends are very similar between the two approaches, and very close for employers.

Taken together these features indicate that, while there are many uncertainties in the supplementations, estimates based on a range of different assumptions about developments over the period are similar in general trend whatever method is used: all show a slower increase in proprietor numbers 1851-61 than subsequent years, a fairly slow increase in employer numbers over the whole period until 1901, and the alternatives are within the same broad bounds. The main differences between methods are for own account, within which the range is most strongly affected by the supplementation chosen for maker-dealers. Although, of course, the precise estimates that result must be treated with caution, the BBCE database deposit identifies the alternatives so that other researchers can opt between them.

### *4.2 Estimates on non-response*

The reconstruction allows estimation of the total of employers and own account supplemented for non-responses. This allows comparisons with the extracted proprietors from Groups 1-6 which can then be used as an indicator of the non-response levels for each Sub-Occode. These are, or course, estimates based on assumptions and hence are at best only a guide.

Those with the highest response rates were mostly from manufacturing sectors and a few commercial proprietors, as to be expected given the phraseology of the census instructions emphasising trades and manufactures. Apart from farmers, which are the single largest group of any Occode and are assumed to have a full census responses from the extractions, the top 10 of the response rates of employers compared to estimates of need for supplementation, equalling over 96% in 1851 and 1881, and over 94% in 1861, included the following nine I-CeM Occodes in each year: Occode 771 Machinists Machine Workers undefined, 555 Cotton & Cotton Goods Manufacture undefined, 589 Tent Maker, 546 Newspaper Publishers, 120

Bankers, 572 Worsted & Stuff Manufacture undefined, 560 Worsted and Stuff Manufacture Spinners Piecers, 244 Steel - Manufacture, Smelting, Founding (in 1851 and 1881), and 259 Ironfounder - Moulders, Core Makers, others & undefined (in 1861). As a result these and many other manufacturing sectors needed little or no supplementation.

| **Sub-Occode** | **Lowest response rates for employer categories having N ≥ 2,500** | **Non-respondent N** |
|---|---|---|
| 872 | **Grocers Tea Dealers E/OA - Sub-Occode from 697** | 11,493 |
| 713 | **Innkeepers, Hotel Keepers and Publicans** | 10,814 |
| 405 | **Builders** | 10,633 |
| 657 | **Dressmakers** | 8,203 |
| 682 | **Butchers & Meat Salesmen** | 8,199 |
| 628 | **Drapers Linen Drapers Mercers** | 7,589 |
| 111 | **Merchant -- Commodity undefined** | 6,259 |
| 39 | **Solicitor** | 6,062 |
| 858 | **Tailors (Not Merchants) - Sub-Occode from 653** | 3,353 |
| 116 | **Auctioneers Appraisers Valuers House Agents** | 3,350 |
| 691 | **Bakers (Dealers)** | 3,275 |
| 698 | Greengrocers Fruiterers Potato Dealers | 2,986 |
| 876 | Wine and Spirit Merchants - Sub-Occode from 722 | 2,701 |
| 235 | **Coal Merchants and Dealers** | 2,646 |
| 409 | Carpenter, Joiner | 2,646 |
| 482 | Chemists Druggists | 2,584 |
| 686 | **Corn Millers** | 2,547 |

**Table 3.** The lowest response rates for Sub-Occodes that contained more than 2,500 employers in 1881 (bold; the same Sub-Occodes also in the lowest 17 response categories in 1851).

Conversely, the lowest response rates were in sectors that were poorly described by the census instructions emphasising trades and manufactures and where respondents could be confused by the alterative census instructions. As shown in Table 3 for the 17 categories with

the largest number of non-responses of extracted compared to the supplementation in 1851 and 1881, these were mainly in retail, refreshments, builders, some professions, and some maker dealers.

More generally, non-response rates clustered around 50-80% and were similar across the census years, as shown for employers in Table 4. But fortunately the response distribution was skewed upwards: over 16% of Sub-Occodes had greater than 80% response rates in each year, over 40% had greater than 65%, and over two-thirds had greater than 50% response rates. Hence the categories shown in Table 3 were the more exceptional; and indeed only 57 Sub-Occodes needed more than 1,000 individuals to be supplemented, and only 264 of the 844 needed 100 or more to be supplemented. Nevertheless, the supplementation process preformed a major task: over 30% of Sub-Occodes had less than half of the employer members fully responding.

| **Response rate category (%)** | **1851** | **1861** | **1881** |
|---|---|---|---|
| >90 | 6.7 | 5.8 | 5.5 |
| 80-90 | 14.2 | 10.2 | 11.0 |
| 65-79.9 | 21.5 | 24.3 | 22.7 |
| 50-64.9 | 27.1 | 27.7 | 27.1 |
| 35-49.9 | 18.4 | 18.5 | 20.6 |
| 20-34.9 | 8.2 | 9.6 | 9.0 |
| <20 | 3.9 | 3.9 | 4.0 |

**Table 4.** Percentage of *Employer* Sub-Occodes at different response levels for each year.

### *4.3 Robustness check*

A robustness assessment of the reconstruction can be made using a testing dataset of labelled *entrepreneurs*. A simple test can be visualised in a diagram, see Figure 2. Suppose, for simplicity of depiction in the figure, that we are classifying *entrepreneurs* and *workers* by their age (shown on the X axis). There are two clouds of points defined by the probabilities

on the Y axis: along the top of are 'true' or labelled *entrepreneurs*, y = Ent, and at the bottom, 'true' or labelled *workers*, y = W. In this figure the logit function is estimated to assess the probability of being an *entrepreneur* according to age; the full logit model used in this paper has many other variables. An external intelligence-led input defines a cut-off which at the same time defines an age above which an individual is predicted to be an *entrepreneur*, $\hat{y}$ = Ent, while below this an individual is predicted to be a *worker*, $\hat{y}$ = W. This generates four categories True Positive (y=Ent, $\hat{y}$ = Ent), False Positive (y=W, $\hat{y}$ = Ent), True Negative (y=W, $\hat{y}$ = W), and False Negative (y=Ent, $\hat{y}$ = W). From these numbers it is easy to calculate the accuracy (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative) and other performance measures.

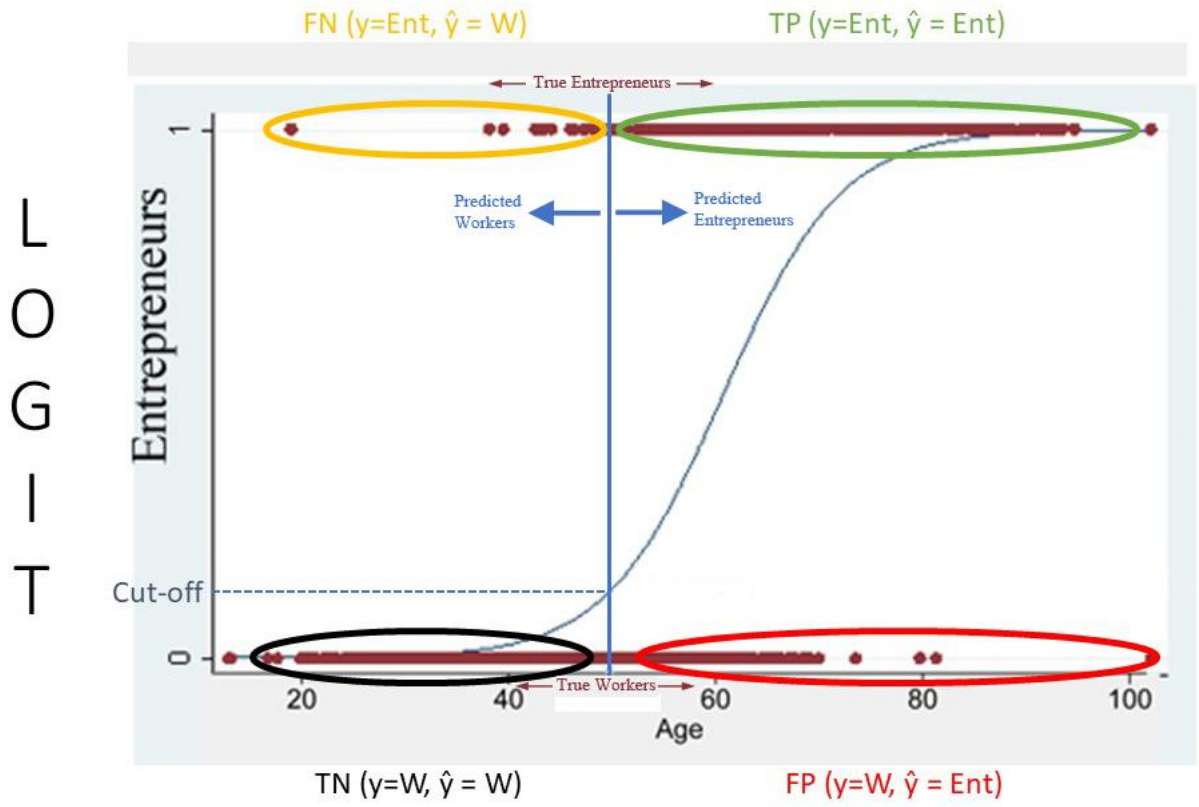


**Figure 2.** Diagram with labelled *entrepreneurs* (on top) and *workers* (at the bottom) from a logit regression to calculate the probability of being an *entrepreneur* from an individual's age. After a cut-off is set the individuals are split into four categories: True Positive, False Positive, True Negative, and False Negative. Accuracy is calculated from these categories.

As a comparison, in a methodology paper (Montebruno et al., 2019), for a subset of this reconstruction, we calculated accuracies ranging from 0.74 using a stand-alone logistic regression using 1891 as training data, rising to 0.96 using a deep learning algorithm, which increased further to an accuracy of 0.99 using the extracted same-year Groups as the training set taking the full occupation strings as a 'bag of words'. The reconstruction method developed here is thus one of a range of methods that can be used to supplement the 1851-81 census responses. However, the comparison of machine learning methods with the core results derived from the intelligence-led and alternative tailored cut-off demonstrate that each method identifies individuals in a very narrow range of alternatives.

### *4.4. Gender*

The original supplementation method as described in WP 9 had several difficulties for estimating plausible age and sex at the micro level. This led to the development of an alternative reconstruction model for the most common female entrepreneurial occupations only (see WP 9, section 5). This method was also intelligence-led and was preferable on some measures, whereas the original reconstruction was better on others. Its key feature was to estimate the women separately from the men in the most frequent categories of female entrepreneurship. The variable cut-off method developed in this Working Paper provides a potentially preferable method for supplementation.

Figure 3 compares the age and employer status distributions for female dressmakers. This compares the 1891 results for employment status (after weighting for non-response and misallocation biases) with the three methods of supplementation:

(i) The original intelligence-led reconstruction using the same logit model for both genders, with random allocation of individuals that are not identified in the extraction Groups (WP 9, sections 3 and 4).

(ii) The original intelligence-led reconstruction using a separate female-only logit model for the main female occupations, with random allocation of individuals that are not identified in the extraction Groups (WP 9, section 5).

(iii) The variable cut-off method for the logit developed here which identifies each individual directly from their logit probabilities, with no random assignment necessary (as in this WP 9.2).

Dressmakers were the largest female entrepreneurial category, constituting over 30 per cent of female entrepreneurs in all years and drove many overall trends. This is one example, but similar differences between the supplementation methods occur across all the main female occupations. For dressmakers, the intelligence-led random allocation process element of the original reconstruction based on aggregates assigned almost all women under 25 as workers, and almost all women over 35 as entrepreneurs, whereas in 1891 at least a quarter of female dressmakers over 35 were workers. The alterative female intelligence-led estimation method based on a model separately for women addressed this issue somewhat, however, when the variable cut-off allocation is used the reconstructed population of dressmakers becomes most like the 1891 distributions.

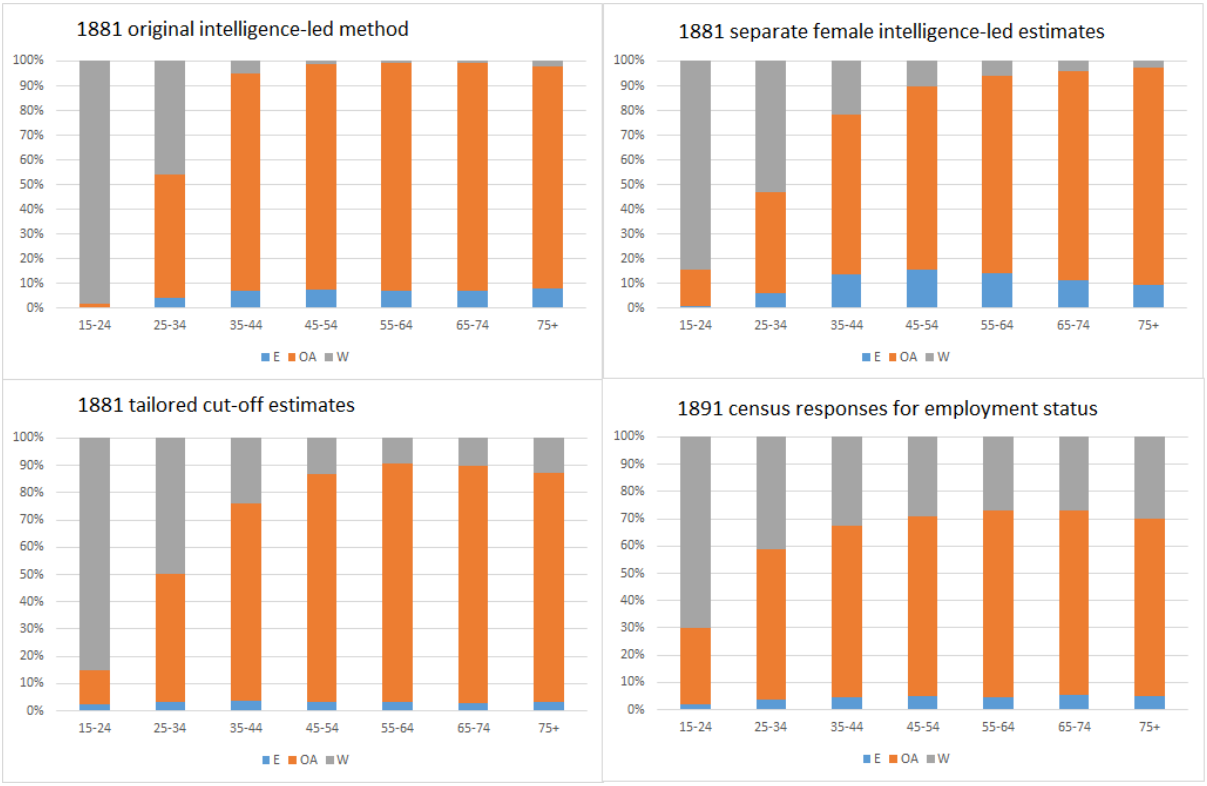


**Figure 3**. Employer status breakdown for female dressmakers by age compared between 1881 intelligence-led (top), 1881 cut-off methods (bottom left), and the 1891 census actual responses (bottom right).

The female-only reconstruction assigns more women to employers compared to the 1891 employ-code reconstruction, and the variable cut-off method improves on this feature. However, in the variable cut-off reconstruction many of the own-account entrepreneurs were married rather than single, a feature that characterised dressmaker entrepreneurs in 1891, 1901 and 1911. In addition, the variable cut-off reconstruction was optimised for individual allocations rather than for aggregates.

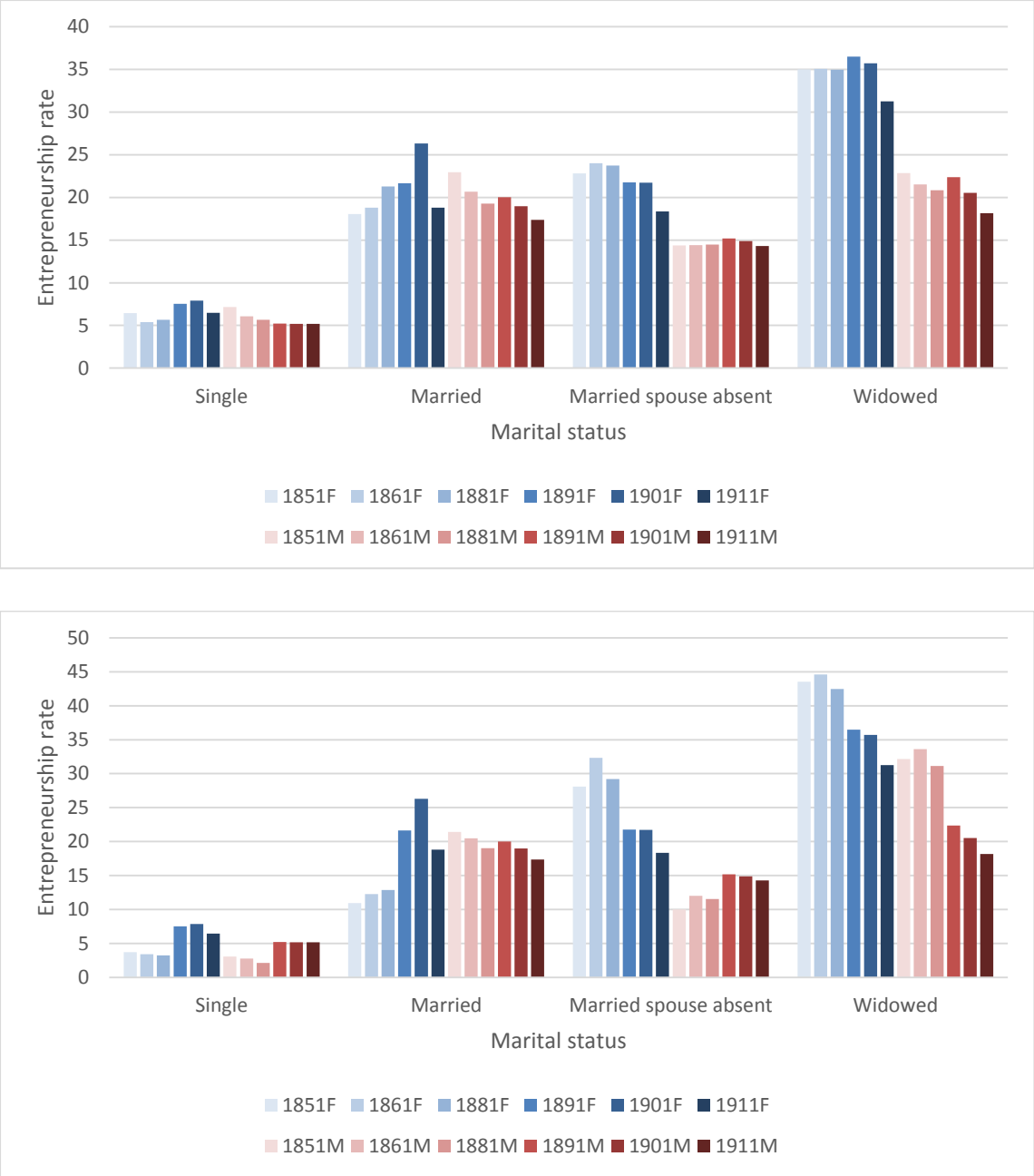


**Figure 4**. Aggregate entrepreneurship rates using the original intelligence-led supplementation method with ***no*** separate female estimates (top), and the alternative cut-off approach (bottom).

As can be seen from Figure 4, there is a possible discontinuity in entrepreneurship rates of both men and women for both methods of reconstruction (the original intelligence-led with no separate female estimation, and variable cut-offs) between the reconstructions for the early censuses and the later censuses. The variable cut-off approach (bottom chart) had perhaps more discontinuity than the original intelligence-led reconstruction which perhaps shows a smoother trend (top chart). For aggregated analysis therefore, the original reconstruction may be preferable. For the micro-level, and in particular occupations where many entrepreneurs were female, but head of households male, the variable cut-off method is better on age and employer status distributions, while the female-only reconstruction has been optimised for age and marital status, but overestimates female employers.

Testing gender as a key indicator of entrepreneurial status is a good example of the constraints on trying to supplement census data and undertake this for small categories. The reconstruction methods all work better at an aggregate level and any attempt to look at parts of the distribution for groups of people is more likely to lead to uncertainties. Hence, as shown here, even though women are a large group of entrepreneurs, when they are disaggregated into categories by age or marital status, the supplements are likely to be less reliable, especially where the sub-categories have very small numbers (such as widows, and married spouse absent; or for the older and very young age groups). Reconstruction and census supplementation is thus an imperfect method, but nevertheless at aggregate level and for large categories, provides a means to estimate the key changes in entrepreneurial developments.

## 5. Conclusion

This paper has outlined an approach using tailored cut-offs to supplement for non-responses to the census questions in 1851-81. This allows individual proprietors to be identified in a comparable the census questions for 1891-1911; hence the responses for 1851-81 are reconstructed as if responded to the 1891-1911 censuses. The reconstructed estimates can then be used to supplement data for individuals that would not otherwise be available in the censuses. These estimates are part of the database deposit in the British Census of Entrepreneurs. The alternative tailored cut-off reconstruction in WP 9.2 allows better identification of individuals than the intelligence-led method used in WP 9, and is thus more

suitable for micro-level analysis. However, the estimates are lower than in WP 9 (see Figure 1), and are they probably lower than should have been in reality. For aggregate assessment the estimates from the intelligence-led approach developed in WP 9 will normally be superior and should be used in aggregate analysis. The tailored logic developed here is preferable for micro-data level analysis of individuals. However, it also experiences major challenges when the disaggregated categories being examined have very small numbers, or the nature of the entrepreneurship being investigated has wide variance across the category being investigated – as demonstrated for the analysis of female entrepreneurs examined here. As in much data analysis, there is 'no free lunch', and there is no perfect solution to reconstructing response to questions that the censuses of the time either did not ask or asked in a rather imperfect way.

Hence, despite the improvements at the individual level, r*esearchers are cautioned that the aim of this paper and the related database deposit is to provide reconstruction estimates for aggregates.* Whilst the individual level is constructed for each economically active person, these are still identified only at a statistical level. The individual level should be used with care, taking account of how the cases were estimated. Generally, fully accurate identification of an individual's employment status is only available for the extracted individuals in the extraction Groups, as detailed in WP 3. Whilst the tailored cut-offs will be statistically valuable, the variables built into the logit estimates need to be borne in mind in interpretations.

It is also important to note that, as with WP 9, reconstructions at Sub-Occode level are not reliable for very small categories, or where changes occurred in the way information was originally defined or collected in the census or is coded in I-CeM. This particularly affects census respondents or enumerators that gave insufficient information (e.g. general categories of 'manufacturer', 'labourer', or 'cotton operative') and many large-scale female occupations such as textiles, and female occupations more generally for clothing and personal services. Generally, the constraints are greater in earlier census years because GRO tried to improve the precision of occupational descriptors used and the way households and enumerators responded.

The definitions of the reconstruction choices, and the identification of the reconstructed individuals by I-CeM RecID, with detailed decisions made for each sector/occupational category, are recorded in a separate data download mounted with this paper and WP 9 (for

the two separate methods). Full information on how the method was implemented is given in WP 19, which includes the estimation methods used in WPs 9 and 9.2, with downloads of the estimation process, for both England and Wales, and Scotland. See WP 20 for further details on the method in Scotland.

The proprietors identified by both the intelligence-led and tailored cut-offs are available by I-CeM RecID in the UKDA deposit of BBCE as:

- EMPSTATUS_IND (for the outputs derived from this WP 9.2), and
- EMPSTATUS_NUM (for the outputs derived from WP 9).

***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).** An Isaac Newton Trust Grant, 18.40(g) **'Business proprietor succession and firm size change 1851-1881'**, provided additional valuable support to help track and code Scottish employer and director data. Figure 1 was drawn by Phil Stickler at the Cambridge University Geography Department Cartographic Unit.

The census database used for 1851-61 and 1881-1911 derives from Higgs, Edward and Schürer, Kevin (University of Essex) (2014) The Integrated Census Microdata (I-CeM) UKDA, SN-7481; see also E. Higgs, C. Jones, K. Schürer and A. Wilkinson, *Integrated Census Microdata (I-CeM) Guide*, 2nd ed. (Colchester: Department of History, University of Essex, 2015). The census data for 1871 was provided by S&N.

**References.**

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

Bennett, R.J., Smith, H. and Montebruno, P. (2019a) The population of non-corporate business proprietors in England and Wales 1891–1911, *Business History*, http://doi.org/10.1080/00076791.2018.1534959

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

Montebruno, P., Bennett, R.J., van Lieshout, C. and Smith, H (2019) Shifts in agrarian entrepreneurship in mid-Victorian England and Wales, *Agricultural History Review*, 67(1).

Montebruno, P., Bennett, R.J., Smith, H and van Lieshout, C. (2019) *Machine learning for the Censuses of the mid-Victorian era*, forthcoming.

Schürer, K., Higgs, E., Reid, A.M., Garrett, E.M. (2016) *Integrated Census Microdata V.2 (I-CeM.2)* [data collection].

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.

**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.* https://doi.org/10.17863/CAM.9508

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

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

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

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

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

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

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

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

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

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

WP 11: Montebruno, Piero (2018) *Adjustment Weights 1891-1911: Weights to adjust entrepreneur numbers for non-response and misallocation bias in Censuses 1891-1911.* https://doi.org/10.17863/CAM.26378
Adjustment weights: https://doi.org/10.17863/CAM.26376

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

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

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

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

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

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

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

WP 19: Montebruno, Piero (2019) *Reconstructing British censuses 1851-1881 for the BBCE: Downloads of datasets with intermediate variables and brief tutorial.*

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

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