Estimates of GB firm-size by sector in 1881 using BBCE sector definitions (EA17 sector codes)

Robert J. Bennett, Harry Smith[1] and Leslie Hannah[2]

rjb7@cam.ac.uk harry.j.smith@kcl.ac.uk lesliehannah@hotmail.com

Working Paper 27:
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.
[1] Department of Geography, Kings College London
[2] Department of Economic History, London School of Economics

October 2021

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

Bennett, Robert, J., Smith, Harry and Hannah, Leslie (2021) *Estimates of GB firm-size by sector in 1881 using BBCE sector definitions (EA17 sector codes)*, Working Paper 27: ESRC project ES/M010953: ‘Drivers of Entrepreneurship and Small Businesses’, University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure.

Keywords: Entrepreneurship, Employers, Self-employment, Census, BBCE, I-CeM

JEL Codes: L26, L25, D13, D22

Estimates of GB firm-size by sector in 1881 using BBCE sector definitions (EA17)

Bob Bennett, Harry Smith and Les Hannah

Working Paper 27: ESRC project ES/M010953: Drivers of Entrepreneurship and Small Businesses, University of Cambridge

1. Introduction

The digital records of employers in the UK population census offer major opportunities for research. These have now been made available as an open access digital resource in the British Business Census of Entrepreneurs (BBCE) for 1851-1911 (Bennett et al., 2020a; https://www.bbce.uk/). This has the advantage that the employers can also be linked with their full digital census responses for the whole population, available in the Individual Census Microdata database (I-CeM) (Schürer and Higgs, 2014), both available at UKDS. These modern digital versions of the census are the first occasion that the complete original responses by employers to the censuses can be fully accessed and analysed. The records available are those surviving in the census enumerator books (CEBs). These records have already been used extensively to assess the scale and trends in the number of business proprietorship over the period between 1851 and 1881 (Bennett et al., 2019a, 2020b, 2021). This has allowed a number of analyses of levels of entrepreneurship.

One the more prized aspects of the employer response in the census is that, for the four years 1851-1881, each employer was asked to give the number of their employees, and additionally the breakdown of their workforce into men, women, boys and girls. This information potentially provides invaluable statistical support for analysis of how firms of different size developed in different sectors and local markets over time, how far market entry was facilitated or experienced barriers, and how individual firms related to the overall economy as a whole.

Despite the huge potential of the firm-size data, however, there were various deficiencies in the census process and its digital capture. This led to Bennett et al. (2019a, Chapter 5) limiting analysis of the firm-size data to preliminary breakouts and analysis. To take data analysis further requires a number of developments - in our understanding of the data, and in how it can be processed to provide unbiased firm-size estimates. The paper has four aims:

1. To develop a methodology (called *data re-scaling*) for re-weighting the raw response data to allow for non- response and any biases in digital data preparation;
2. To apply this data re-scaling to sector distributions based on the BBCE standard EA17 sector codes;
3. To provide a first set of estimates of firm-size by sector for 1881 as a pilot;
4. To prepare the ground for using other sectoral coding schemes that allow comparisons over time and between Britain and other countries.

Part 1 of the paper introduces the census data on workforce size. Part 2 reviews previous analysis of non-response and data truncation, and the challenges for developing unbiased firm-size estimates. Part 3 develops the re-scaling method used for estimation of firm-size by sector for 1881, commenting on alternative choices that can be made. The 1881 estimates by size and sector are presented. Part 4 of the paper compares the 1881 estimates with alternative contemporary data from Booth (1886) and the Factory Returns for 1878 and 1885. There are no strictly comparable data available for the whole GB population of individual firms, but these comparisons of aggregate data for some sectors give useful confirmation of the general accuracy of the estimates and where improvements in method are required. An additional comparison is made with 'truth' data for the actual workforces of firms of 1,000 employees and over, developed from Les Hannah's research. Part 5 evaluates the re-scaling method and concludes on how to combine with 'truth' data for the very large firms.

The paper records work in progress; it is subject to revision of method, and updating of the firm-size estimates.

Part 1: The census data on employers and their workforce size

The UK population census question in 1851-81 was a unique effort to ask employers to state 'the number of persons of the trade in his employ'.[1] The question has never been repeated in the population censuses, though other efforts to collect firm-size information began in the

[1] General Instructions, Census of England and Wales, 1851, 1861, 1871 and 1881.

census of production in 1907, and modern government statistics give aggregate information through the BPE, LFS, and other sources from the 1980s, with the IDBR offering potential for analysis of modern individual business records.

The 1851-81 censuses thus offer unique data on the whole business population for a period of industrial development where previous data having been extremely limited. However, its potential has been restricted by a number of *data-collection deficiencies* (Clapham, 1932, p. 35; Bennett et al., 2019a; van Lieshout et al., 2021; Bennett and Hannah, 2021):

First, the census administrators (the General Register Office: GRO) gave a low priority to the question (Higgs, 2005). This resulted in various defects:

1. Little analysis and tabulation was undertaken of the data collected by GRO which means we have few contemporary assessments of its quality; hence, we have a very limited basis for checking digital entry against the contemporary processing of records as they were then processed. Only a few tables of results to the question were published: in 1851 a few national summary tables for non-farm by major occupations, with some greater detail for farmers,[2] and in 1861 and 1871 (and 1881 in Scotland) tables only for selected counties only for farmers.[3]

2. The GRO gave a low priority to ensuring the quality of responses to this census question; instead the census was primarily concerned with age and risk to health from occupational hazards. This resulted in a higher level of non-response, or incomplete response, to this question than other parts of the census. The question was part of a complex question on occupation. In addition the instructions for data collection by census enumerators, its oversight by local registrars, both appointed through GRO, gave virtually no guidance on how to deal with the workforce element of the question, with almost all attention given to occupations and not to the workforce employed. Moreover this was exacerbated by supplemental instructions to householders on how to answer the question that focused entirely on different occupational groups (mine owners, professions, etc.); these instructions

[2] Census of Great Britain, 1851. Population Tables Vol 1; lxxviii, ff.

[3] 10 counties in 1861, Census of England and Wales, 1861. Vol. III General Report, 139-143; 17 counties in 1871, Census of England and Wales, 1871. Vol. IV General Report, xlvi-xlvii.

completely lost sight of the workforce issues involved meaning that employers in these specific sectors usually had even higher levels of non-response. This was further exacerbated by census administrators appearing to have given greater emphasis to collecting fuller workforce data from farms because of political priorities to track agricultural land ownership and trends in farm labour. Although, even for farm employers, there were non-responses (Mills, 1999; van Lieshout et al., 2021). This means that there are potential sources of bias deriving from:

- General non-response or incomplete response
- Differences between sectors in level of response, possibly fuller for farms
- Very few employers distinguished their workforce into men, women, boys and the enumerators' books give few indications that they attempted to overcome this deficiency.

3. There was a very imperfect GRO question design. There is no evidence it was ever piloted, nor that difficulties in eliciting the information sought were ever addressed by effective efforts at question redesign. Apart from deficiencies in the format of the question leading to non-responses, the question ignored the reality of how businesses were constituted and how this should be handled by the question structure. The question assumed an employer *was* the firm: i.e. it had a single proprietor. There was no attempt to distinguish firms from their proprietors. For *firms organised as partnerships* the 1851 question gave no guidance. In 1861-81 additional guidance was added that 'In the case of Firms, the number of persons employed should be returned by *one partner* only'.[4] But this format resulted in responses ranging from each partner leaving the instruction to the others with no-one responding, to several or all responding leading to multiple duplicate (or near-duplicate) replies. Additionally the *organisation of firms as limited companies* was completely ignored by GRO. It was not considered until the 1931 census which for the first time included directors in the occupational list, though again instructions on how this affected employer status was not tackled. In 1851-81 salaried directors or managers could consider themselves as employees of the company so that the question directed to 'employers' could correctly lead to no-one giving workforce numbers for that firm. Similarly a company director remunerated through dividends as a shareholder, or through distribution of profits or capital gains, could correctly consider the company as the employer. A question could have been directed to the manager

[4] General Instruction, Census of England and Wales, Householder's Schedule, 1861.

or controller of the firm, and in some cases a senior manager (or for farms a bailiff) did respond, but this was not very systematic. This led to a high potential for non-responses from the incorporated sector.

This means potential sources of bias from:

- Partnerships responding inconsistently, with potential duplicates, and non-responses.
- Directors or other company managers responsible for workers had a high probability of non-response;
- Since the incidence of incorporation varied by sector there could lead to other forms of sector biases.

In addition to GRO defects, second, there was incomplete digital capture of the responses the employers actually made. The original census CEBs for 1851-81 were encoded by genealogical providers contracted to The National Archives (TNA). These were converted into the digital database of the full population census for 1881 (Woollard and Schürer, 2000), and for 1851-1911 in I-CeM (Schürer and Higgs, 2014). The BBCE provides the employer responses using I-CeM as a starting point to extract, parse, and code the alphanumeric occupation strings in the CEBs by workforce size. The BBCE results from searches of the entire population data, ranging between 17.5 and 26 million people per census (Bennett et al., 2020a). The workforces stated by employers' census responses are given by the variable ETOT in BBCE for each employer (with a breakout by men, women, boys and girls, for those few that responded). The transcripts available to I-CeM varied in provenance of the transcribers used by Find My Past, and the quality they achieved, which were not fully aligned within census years, nor between years. Find My Past also has no transcripts of the occupational question containing the employer responses for 1871. BBCE infilled 1871 from another genealogical provider as well as infilling many gaps of transcriptions for other years. All transcripts have a level of error and omission, but the employer question has a much higher level of transcriber truncation than any other question because of its complexity and the length of the responses required to fully comply with the question. This single question sought information on occupation (the job performed), social rank (such as M.P., J.P., alderman), qualifications (for medics, clerics, etc.), as well as the workforce of employers and its breakdown into men, women, boys and girls. The design of the question, and instructions to enumerators also emphasised all the other elements to be recorded first, before the workforce. If a response actually included all this information, transcribers had a tendency to ignore the later information. This was particularly likely if an individual had multiple public offices and other information before they gave the workforce. It was a characteristic of many employers, particularly the larger and more prominent ones, to hold many public offices and give various other details before their workforce. This was also true of many small farmers and small businesses: even though they were small employers, they were important locally and held various offices in their locality.

This means potential sources of truncation bias that:

- Increase with firm size, and
- May differ between sectors.

Part 2: Non-response and digital data truncation

2.1 Non-response and partial response

BBCE assessments

There are important previous assessments of responses biases and deficiencies of digital capture. An important starting point is the analysis and supplementation method given in the BBCE downloads. For employers in aggregate analyses by Bennett et al. (2020b, 2021) found significant differences between the number of employers giving workforce data for the 1851-81 question compared to what would be expected from the number of employers declared in later censuses, from 1891 onwards. The 1891 census question sought a much simpler response: to state if an individual was an employer, own-account proprietor, or a worker. This question also had deficiencies of design and data collection by GRO, which require adjustments. However, the 1891 responses as 'employer', when adjusted, provide a means to compare with the aggregate number of employers who responded in 1881 as identified by stating a workforce number. The comparison for 1881 can also be made for earlier years 1851-71. This comparison demonstrates that (*for E&W*) only 44.4% of employers responded in 1851, and this proportion declined to 34.1% by 1881, requiring an uplift of 2.935 (Bennett, et. at. 2021, Figure 1). Indeed the actual *number* of responding employers remained almost the same over each census 1851-81, at about 200,000; but the number of employers as a whole increased. Hence, the *proportion of respondents declined*. This may have been a result of increased numbers of partnerships and incorporated businesses that have potentially higher non-response rates, declining motivation by GRO or enumerators over time, or other factors. There are also indications in this analysis of sector biases, suggesting that professions, mine owners, ship owners, and perhaps some retailers were less likely to respond – as expected from the deficient question design. The estimate of a response rate of 34.1% for 1881, requiring an uplift of 2.9, is the best measure that currently exists for the 1881 data: indicating a non-response rate across all sectors of 65.9%

The BBCE estimates, *for non-farmers*, use two methods to *supplement* the reported employer numbers, which vary according to whether aggregate numbers, or the individuals themselves, are identified by supplementation. The results from the two supplemented estimates are given in BBCE as, respectively, EMPSTATUS_NUM and EMPSTATUS_IND (Bennett et. al., 2020). Both were constructed separately for 843 of the 844 I-CeM occupational and sector categories to allow for differences in non-response or digital capture by occupation/business type. The method identifies response/non-response characteristics of proprietors by occupation compared to workers from the later censuses 1891-1911 (given in Bennett et al., 2020b) to estimate the supplemented responses in 1851-81. Although entrepreneurial characteristics may have changed, the characteristics of responding and non-responding proprietors were similar over time. The method reallocates individuals to their correct ‘employment status’ - as employers, own-account proprietors (who had no employees), workers, and economically inactive.

For farmers, the 844th I-CeM occupation, supplementation provided in BBCE in EMPSTATUS_NUM and EMPSTATUS_IND is based on a specific farming supplementation model. This uses the additional information sought from farmers in the census question 1851-81: to give the size of acres occupied. Acres, together with a range of other information specific to farming, were used to give highly specific supplements that varied by part of the country, land capability, latitude, longitude, distance to main local market, as well as the farmer’s household structure and demographic variables (Montebruno et al., 2019a).

Comparisons of BBCE and large-scale secondary data

The methodology to prepare BBCE supplementation seeks to counteract the joint effects of non-response and truncations. This involved extensive checking of secondary and primary sources to assess potential numbers and trends between census years, separately for employers and own account operators by sector. This was mainly focused on the chief sectors that had major impact on assessments: i.e. the largest and those that were most volatile or subject to major changes (as a result of economic cycles or innovation and technology). These sectors were selected from the fine mesh of 844 census occupational categories and disaggregate sectors estimates used for the full estimation (Bennett et al., 2021). As there are no sources that give these data for all sectors on the scale required across the whole country this was inevitably approximate and relied on sector case studies, local case studies, parliamentary and other reports. This was referred to as an 'intelligence-led' approach: this defined the BBCE supplement EMPSTATUS_NUM. It was complemented by a more automatic statistical approach that defined the BBCE supplement EMPSTATUS_IND (Bennett et. al., 2021), and the farm estimates. It is believed that the BBCE supplementary data at aggregate level are the best that can reasonably be achieved at the scale required, although all the working steps and data are provided so that other researchers can replicate and improve the method in the future (see downloads available through www.bbce.uk). Hence, BBCE supplementation already embeds and uses comparisons with the main secondary data sources.

An additional test compared the BBCE supplementations against trade directories. Two comparisons were made for the 1881 supplementations: for aggregates for selected sectors against Kelly national sector directories; and by matching actual individuals against local trade directories. The national directory sector comparisons confirm approximately similar trends in proprietor numbers, especially steep growth in some sectors over 1870-80 (Bennett et al., 2021; Figure 4 and Table 4). For the local directory individual comparisons the numbers identified in the directories and census were very close, with some discrepancies occurring because differences between trading and residential address, and because directories were not up-to-date (individuals were inactive or no longer proprietors in the census). Generally the census shows more employers not listed in directories, especially where proprietors were women. However, for matched individuals results were much more variable, as to be expected. For some large sectors the use of the same occupational descriptors by both proprietors and workers limits the accuracy of matching the correct individuals in BBCE supplementation: e.g. for 'blacksmiths', many building trades such as carpenters, painters, or plasterers, which have good numerical matches but poor identification of individuals. The largest difficulties were cases of industrial workers such as 'weavers' and 'spinners' which cannot be differentiated in census responses between workers and proprietors (Bennett et al., 2021; Tables 5 and 6).

Despite the expected limitations, the BBCE supplements give generally robust sector estimates compared to national directories, with local directory comparisons generally confirming estimates of proprietor numbers, but with the matching for specific individuals varying in quality. Also female supplementation, which is expected to give poorer estimates, matched well.

These checks confirm that the BBCE supplements provide a generally good base for numerical estimation. The estimates for NUM and IND simultaneously compensate for non-response and truncation deficiencies in digital capture. They allow the total digital records, by occupational category, to be supplemented to give estimates of what would be expected if *fully* captured by the original question and accurate digital capture. Although they take no account of differences in employer workforce size they provide an important starting point for checking that any size-adjusted weighting scheme matches the estimated number of employers in aggregate. Hence, they provide the main starting point below.

2.2. Estimation of non-response: tests by firm size

A second important starting point is further analysis by Bennett and Hannah (2021) that attempts to assess non-responses bias by firm-size and sector. There are two bases for this assessment: 'truth data', and record-linked responses.

'Truth data'

In general there is no source of 'truth' data giving information on employment size for all firms; this is why the BBCE is so valuable in being the first source to approach a full coverage. However, it is possible to construct 'truth' information for a limited sample of firms that have complete coverage as a result of archival preservation of employer information detected by intensive combination of all available records. In general this can only be achieved for the largest firms that have attracted sufficient attention for their workforces to be recorded or reported in primary and secondary sources, and which generally have better chance of archival survival. Leslie Hannah has attempted to construct such a data base for *the largest firms of 1,000 or more employees in manufacturing* (on a broad definition) in 1881. He used BBCE as a starting point and then supplemented this with intensive search of all secondary sources (contemporary local and national newspapers, parliamentary papers and reports of factory visits by professional societies, Grace's Guide, dictionaries of biography, a broad range of reliable internet sources, with some from company histories and secondary research). These secondary searches also allowed supplementation of transcripts by checking the records of all BBCE ≥1,000 employee firms against the original CEBs. This allows separation of truncations and non-responses. A summary of the method is given in Hannah and Bennett (2021) with an appendix and supplementary material listing the 438 UK manufacturing firms with their employees. This list of firms is as close to a 'truth' database that is possible at this historical distance; any omissions are likely to be very limited – there were few or no large firms that did not occasion a report in local and national newspapers or other record. Of Hannah's UK list, 401 were located in Britain that had ≥1,000 employees in 1881 (i.e. excluding Ireland), of these 399 had been identified at the time of the analysis of non-responses.

Comparison between BBCE and the Hannah 'truth' data for these 399 firms show *37% were contained in BBCE*, even after the BBCE transcripts were made complete by including the full census responses in the CEB (i.e. eliminating transcriber truncations). Whilst this is disappointing, it is close to the 34.1% response rate estimated by BBCE supplementation across all sectors (see Section 2.1). Hannah's list also includes 25 Statutory and Chartered companies, and 120 Public and Private Registered companies. Given that directors and managers of companies were completely ignored in the census instructions it is expected that few replied. Indeed no census response related to Statutory and Chartered companies, and only 15% of Public and Private Registered Company responded in 1881 giving workforce numbers. Overall the response rate from companies was 12%. Hence, legal form had major influence on non-response rates. For partnerships the response rate was 49% and sole proprietors 63%; 50% for the combined non-corporate sector (Bennett and Hannah, 2021, Table 1). This is better for manufacturing than the 34.1% response rate across all sectors estimated from BBCE supplementation, which includes infill of the original digital records.

A more extended analysis of non-response demonstrates that legal form was the dominant and *only significant factor* explaining non-response for these large manufacturing firms. Both sole proprietors and partnerships were significantly more likely to respond, with little difference in significance between them. Tests of covariates for business sector, location by region, size of firm, and years of incorporation were all insignificant for explaining non-response, with minimal additional significance from any interaction effects (Bennett and Hannah, 2021, Table 2 and 3). Also Scotland, despite its different laws of partnership and incorporation, was not differentiated in any way; it also had legal form as the only significant factor explaining non-response. To test robustness, the corporates were removed from the statistical tests: there were then no significant factors or covariates explaining non-response; i.e. non-response was random for these large *non-corporate* manufacturing firms.

These results mean that for large manufacturing firms BBCE *non-corporate coverage* can be re-weighted to compensate for non-response to give acceptable estimates for most statistical analyses. However, for *corporations*, and comparisons between them and the non-corporate sector, other sources of data need to be used to infill the systematic gaps in BBCE, by similar methods used by Hannah to construct ‘truth’ data. However, this is unrealistic since the vast majority of firms were small and have no secondary data available at the scale required for ‘truthing’. This conclusion covers 1881, but preliminary analysis by Hannah suggests a similar conclusion for 1851, and by implication the intervening years 1861 and 1871 as well.

Record-linked responses

Since ‘truth’ data can be constructed only where secondary sources allow, which is essentially usually restricted to the largest firms, an alternative is required to develop understanding of non-response characteristics of the great mass of small and medium-sized employers. One way to assess these was developed by Bennett and Hannah (2021) by using record linkage to compare the same employer’s responses across successive censuses. This makes the assumption that an employer of an ongoing business with a workforce identified in one year should have made a census return in adjacent censuses for as long as remaining the proprietor. The census data give four possible years, 1851-81, with ‘sandwiched non-response gaps’ identifiable where an employer in one or two years had a non-response gap between years where responses were given. These sandwiched gaps were the primary target for testing consistency. An additional sample was constructed with a non-response gap for an earlier or later census adjacent to a year where the employer gave a workforce; this is less reliable as an identifier of non-response than sandwiched gaps, since the proprietor could correctly make a null response before entering business, or after exiting; there is no large-scale source to remove these falsely identified non-responses. However, this ‘adjacent gap’ sample was used as a possible further comparator of non-responses. To limit false identification of gaps the samples were restricted to employers with 10 or more employees, since firms of this size were more likely to be actually continuing employer businesses, whereas very small firms might go in, out and in of employer status over successive censuses. The record-linked data sample provided with BBCE is used as a starting point, with statistical tests showing the sample to be representative of the underlying employer response population at $p \geq 0.05$ or greater (Montebruno and Bennett, 2020).[5] The sample combines forward- and backward-linkage, thus overcoming most of the typical representativeness problems of record linkage studies. The samples provided an N of 968 sandwiched gaps, and 1,567 gaps in adjacent census. To restrict attention to non-response, both samples were after checking and infilling BBCE against the original CEBs, thus eliminating any truncations from digital capture. Table 1 summarises the results.

Non-response increases with firm employment size for both samples: ‘sandwiched’ and adjacent gaps. Tests of significance show that *firm size is the main explanatory factor*, with proprietor age as the only other significant covariate. Tests of locational differences (between counties), wealth and status effects (as indicated by N of servants), and localised spatial variation by population density were all insignificant. The order of recording of farmer in occupational strings is not significant, nor is the presence of a non-occupational title or rank, nor the sector. This suggests that a possible weighting scheme to compensate for non-response is possible, applied to firm size, with age as a possible additional or selection

[5] There is weaker representativeness by sex; but the female sample is very small, and only one female proprietor occurs with response gaps; generally female employers of businesses over the 10 employee threshold operating over long periods were very rare.

variable. Table 1 gives the raw results of the non-response rates by firm size category and aggregate sectors. These provide an option for re-weighting schemes.

Table 1. Rate of employer non-response by employee size and sector, as percentage of record-linkages; row 7 aggregates the ≥800 category. Note the N is very small for the larger firms, especially the 600-799 size category. Source: based on Bennett and Hannah, 2012, Table 7.

Size group (employees)	**Sandwiched gaps**				**Gaps between adjacent censuses *(all sectors)***
	All sectors	**Manufacturing**	**Non-manufacturing (excl. Agric)**	**Agriculture**	
10-15	2.4	5.7	3.2	2.1	4.4
50-55	3.6	1.4	3.5	10.0	7.7
100-150	4.9	3.9	9.4	0.0	6.8
300-599	9.6	5.9	21.1	-	11.8
600-999	19.2	17.6	22.2	-	-
≥1,000	18.4	17.8	20.0	-	-
All ≥800	18.7	17.8	21.1	-	12.6
Average	4.0	5.9	6.4	2.4	7.3

2.3 Digital data truncation

2.3.1 Comparison with published tabulations

One way of assessing the extent of truncation is to compare the original response levels in the CEBs between the digital capture in BBCE and the tabulations by GRO clerks for those few tables GRO published. van Lieshout et al. (2019a, 2021) provide comparisons with the most useful GRO published tables.[6] In E&W these cover 1851-71 for farmers, and only 1851 for

[6] van Lieshout et al. (2019a) covers E&W; it is updated and extended to Scotland in van Lieshout et al. (2021)

non-farmers. No GRO tabulations were made for 1881 in E&W, but limited farm tabulations for 1881 were made in Scotland.

England and Wales (E&W) farmers

In 1851, which has the most complete GRO tables for E&W, there is a very close relation between the published farmer numbers given by GRO as 249,431, and 249,251 in BBCE/I-CeM, with also similar numbers by sex. Both these figures include those stating ‘retired’ or ‘former’ farmer. However, the identification of farm ‘employers’ derived from transcribed census responses of those who gave a workforce has poorer match: 133,620 employers tabulated by GRO, and 93,547 in BBCE/I-CeM: a 30% difference, with deficits concentrated in Lancashire and Yorkshire West Riding (van Lieshout et al., 2021, Table 1). By employee size almost all size categories have deficits in BBCE/I-CeM, with 95% of the missed employers in the range 1-6 employees. In contrast, the largest size class (60 employees and over) has 25% more in BBCE/I-CeM than the GRO tables (van Lieshout et al., 2021, Table 4). There are also some systematic differences between regions. Hence, whilst GRO must be taken as reasonably accurate as an estimate of what was recorded in census responses, it is imperfect and contains some errors of clerical tabulation. Nevertheless it is clear that BBCE/I-CeM has data truncation for employers within the farmer category.

For 1861 the GRO tabulated farmers by size for 10 English counties. However they did not tabulate in the same way as 1851 and included in farm size ‘the farmer himself must be added, and frequently the farmer’s sons at home’.[7] Moreover, it is not clear how this was actually done. Hence, for 1861 there is no exact way of calculating digital truncations since the smallest farms had additional workforce that is not included in the extractions made for BBCE, which are only for those with *declared* workforce; and there is no accurate way of including what census clerks included. Nominally, the 1861 GRO table indicates a deficit in BBCE/I-CeM of 35.8% identifiable as employers. But it is probably similar to the 1851 (van Lieshout et al., 2021, Table 6).

For 1871 GRO tabulated 17 ‘representative’ counties in England. This shows a 26.3% deficit for employers in BBCE/I-CeM compared to the GRO (excluding farms giving no employees

[7] 1861 Census England & Wales, General Report, p. 29.

included by GRO). This is fairly uniform across the size categories of 1-49 employees. As in 1851 and 1861, BBCE/I-CeM identifies larger numbers of employers for larger farms of 50 or more employees, with numbers particularly higher for 50-55 (van Lieshout et al., 2021, Table 7).

Scotland farmers

For Scotland in 1851 farm employers are 8.2% under recorded in the digital records, mainly in the smallest farms. However, as in E&W, the largest size class is larger than the published: 3 times as many as in the published tables – 33 compared to 11 (van Lieshout et al., 2021, Table 5).

For Scotland in 1881, the only other published table of farm employers, there are difficulties with the Scottish data resulting from lack of clarity of definition of what was included by census clerks regarding crofters, and because the tabulation was for men only. It appears that the overall deficit was similar to 1851, concentrated in farms of 1-4 employee; farms over 40 employees have greater numbers in BBCE/I-CeM (van Lieshout et al., 2021, Tables 8 and 9). These comparison tables also breakout the response's for men, women boys and girls, showing the major deficiencies of the census process for capturing this breakdown of the workforce.

E&W non- farmers

1851 is the *only year* in which GRO gave tabulations in E&W for non-farm employers. There are difficulties for reproducing exactly how GRO tabulated the data since they quote tabulations of 'masters' employing 'men' which are smaller categories than all identifiable employers. If only those employers explicitly declaring men are included, whether or not they state 'master', there is a 7.3% deficit in BBCE/I-CeM: from 87,270 to 80,921; this is greatest in London and the NW, but with some divisions higher than GRO (van Lieshout et al., 2021, Figure 2 and Table 2).

As for farms, the main under-estimation is for the smallest employees, again mainly 8 employees and less, with a strong concentration of under-estimates in London and the NW.

However, excluding these 2 regions the BBCE/I-CeM has 12.6% more non-farm employers than GRO tabulated. This includes the effect of any discrepancies between GRO and BBCE handling of partnerships and family employees; these cannot be pinned down because of the lack of precise detail of how GRO accomplished their tabulations.

But for the largest firms the BBCE data in total contain 17-22% more responses than GRO tables. After excluding London and the NW, which have large archival losses, all size classes have more employers than GRO – ranging from 5% for the very smallest with 1 employee, up to 2.5 times more for the largest (though for the largest the number of responses involved is very small). The under-estimate by GRO is roughly scaled as a quadratic with increasing size. This indicates two factors. First there were clearly some of the same problems with in interpreting the CEB responses for GRO clerks as those encountered by modern transcribers and largely overcome in the BBCE digital processing – probably mainly due to missing employee numbers in the very complex returns of many large employers, which were either squeezed into the small box available on the census form or spread over several lines against other people! Second, and probably more important, is the difference between what GRO tabulated in 1851 and the full responses included by BBCE. Assessed in van Lieshout et al. (2021, Figure 2), did GRO tabulate only those explicitly stating ‘masters employing men’ (as stated in the 1851 tables, but clearly incorrect as that gives far too few when compared with BBCE – less than one third in all regions); did it include men and ‘other’ (which slightly exceed the GRO tables in all cases (except London and NW, and marginally below in Eastern), though they are relatively close in most regions and almost identical in Wales and Northern; or did it, as in BBCE, include any employee of any type which are about 5% greater than men and ‘other’ in most cases? It appears that GRO probably worked with a definition of men and ‘other’ and missed some responses. Also critical is that the comparisons demonstrate the impact of some missing archive records that GRO had available when making the published tabulations, but which have now been lost - mainly for London and the NW. This is important in how BBCE is used and is the most important feature that affects the re-scaling method used here.

Some effect of the difficulties of understanding GRO tabulations can be seen from the estimates of average firm size. The GRO reported that for non-farmers 87,270 masters employed one man or more who had a total of 727,468 men in their employ, or an average of

8.33 men each.[8] The BBCE database has 88,364 employers who employed a total of 870,370 people, giving an average of 9.85 employees, or 9.04 (if calculated in the same manner as the GRO, which calculated the average within the larger employee-category ranges, not the actual employees numbers of each individual employer)(van Lieshout et al., 2019a, p.23). The digital records thus will tend to give higher estimates of mean firm size; this is a result of having smaller proportion of the smallest firms, larger proportion of larger firms, and by using individual firms rather than calculating averages from categories as used by GRO.

Scotland non- farmers

In 1851 these were tabulated by the Scottish census only for the 9 major Burghs, although this is where most non-farm businesses were located (van Lieshout et al., 2021, Table 3). Comparing all employers in BBCE/I-CeM with this tabulation, defined as employers of any employee (not exclusively 'men' as defined by GRO) indicates there were 9.7% more employers in BBCE. These were mainly in the smaller firms of 1-5, but also in the medium sizes of 10-49 employees.

Conclusions on comparison of digital capture with published

The comparisons are ambiguous and difficult to interpret. This is because the exact method used by GRO is not fully clear in their published notes, and they were not consistent in method between years, or the parts of E&W or Scotland covered. It is evident that tables were not actually restricted to men only or masters, as GRO stated. In Scotland, they sometimes included at least some crofters, but probably not all. Some years included imputations from family workforces as employees of the head, but this does not seem to have included all such family with similar occupations, and it is not transparent how decisions by GRO were made (if indeed it was actually feasible for GRO clerks to consistently apply a set of inclusions rules to these records which are very difficult to interpret).

Hence the comparisons give only general guidance on how any re-weighting can be applied to compensate for truncation when researching firm-size data:

[8] 1851 Census England & Wales, Population Tables I, Vol. II, p. lxxviii.

1. Most clear is that there is a probable deficiency of digital capture for the smallest firms, mainly under about 10 employees, but it is unclear from the GRO tables if this applies across each year in a consistent way or not.
2. It also clear that GRO clerks failed to capture some for the largest firms, with counts especially low for the largest non-farm businesses in both E&W and Scotland.
3. There are limitations of BBCE/I-CeM digital data deriving from archival losses that have occurred from the TNA records now available to be digitised (see Section 2.4).

2.3.2 Truncation estimated from record-linked samples

Previous analysis of truncation using the record-linked sample of *sandwiched* non-response gaps reviewed above required prior detection of inadequate or truncated transcriptions. These were used to infill the BBCE data using the original manuscript CEBs. The sandwiched-gap sample thus offers a means to compare truncated BBCE records with the CEB originals responses. This sample can be used to assess how far transcription truncations and defects are systematic across the BBCE database.

Using this sample Bennett and Hannah (2021, Table 9) show that *firm size and sector are highly significantly associated with the probability of digital data truncation*: larger firms and some sectors are systematically more truncated. All firm size classes are affected for the sample (which is for 10 employee firms and above). Agricultural processing and refreshments are the only sectors not affected by truncation among the 13 sectors in the BBCE EA17 classification. However, truncations *within size and sector categories* are largely random.

Other tests show more minor effects: for multi-sector portfolio businesses, where farming was recorded first, any other sector is likely to be truncated, (though not within the category of non-portfolio farm responses). Among other covariates there is a positive association of truncation with the number of servants, although at a low significance level: larger households are more likely to be truncated. This is probably arises because many had 'servant' recorded in the relationship column leaving the occupational column empty so that employer's details spilled over it and became misattributed. Servants could perhaps be used as a selection variable, but the significance value is weak. However, the age of the respondent, string length of the response and local population density are not significant.

An important finding in Hannah and Bennett is that truncation shows no significant county effects in England and Wales (but Scotland was not estimated in the RL data). This means that BBCE infills have mostly corrected the original truncations and data losses in I-CeM which van Lieshout et al. (2019a, 2021) show were highly concentrated geographically effects in England and Wales (as noted above, mainly affecting London and the NW), whilst Scotland had no data significant losses and its truncations appear to be random.

It is noteworthy that over 88% of the England and Wales variance is unexplained so that truncation is clearly complex. This is probably the result of TNA using different external contractors to transcribe the data who managed the process in different ways, with transcriber variation in error rate, as well as any database errors in the capture by I-CeM and in BBCE.

The Hannah and Bennett analysis is consistent with the expectations from the census instructions and the small size of the boxes on the census form in which respondents had to write the required information. As a result the CEB entries from employers with larger firms, which often had to give more information, were either very cramped, or spilled over into the box or boxes below which were supposed to be used for other people. Transcribers often ignored all the additional detail of complex employer responses with a tendency for some complex responses to be not transcribed at all: or they did not perceive the information in lower boxes on the form when working line-by-line down the page. Transcribers working to tight budgets and timetables simple ignored these troublesome responses or missed them in sequential keying. As a result Bennett and Hannah found that firm size that explains this effect most effectively, with string length itself also significant but at a much lower level. This seems to be because smaller firms responses usually fell either on one line in the census form, or the response was cramped into one box of 1 or 2 lines; small firms rarely spilled into another box. This means that firm size can be used as a principle reweighting variable.

The Hannah and Bennett analysis also shows that using the record-linked sample to estimate compensation weights is not straightforward. Weight calculations by employee size and sector have wide standard errors and low significance levels, which will create more errors than they correct in most applications. A much larger sample is required to obtain reliable estimates of weights using record-linkage. This is difficult to achieve, and the later analysis in Section 5 of weights derived from the record-linked sample indicates that they contribute only a minor element of the re-scaling of the BBCE employer responses required. As a result, a hybrid method is suggested, based on re-scaling of the BBCE IND estimates for the majority of firms (under 1,000 employees) combined with large firm 'truth' data.

2.4. Archival loss

Whilst the inconsistencies between GRO and BBCE give only weak guidance on possible adjustments for truncations of employer responses, archival losses can be more reliably estimated. This is because the GRO processes gave far more priority to getting populations counts correct: this was the main objective of the census. Hence, published population numbers can be used to check and re-weight modern digital records. Weighting schemes have already been widely used for the 1861 data which have the highest losses. Here weights are evaluated for 1851-71; for 1881 archival loss is far too small for adjustments to affect subsequent analysis (see van Lieshout et al., 2019a, 2021).

For the general population archival losses (i.e. not attacking the problem of truncations and non-responses by employers) data can be weighted to compensate to match the published population figures. This can be achieved in aggregate, but since the archival losses are concentrated locationally, and the lowest level for which the I-CeM and BBCE data are most accurate is at the RSD level, they are best developed for RSDs. Weights should be applied to the population present in I-CeM, and have been created for each RSD. This means that if a certain RSD has 50% of its population missing, records of the remaining population will be weighted up by x2. This method means that RSDs that are fully missing cannot be weighted individually, though the whole population can be weighted up to compensate for these missing RSDs.

Weights can be either adjusted to within the 5% level, or all population numbers can be adjusted (even those RSDs within 5% of the published levels) to match the published figures exactly (see Jaadla, 2019). This is a matter of user choice. The weights given in BBCE downloads linked to van Lieshout et al. (2019a) are for all years 1851-71, separately for men and women, and for the total. The weighting by gender extends the weights in Jaadla (2019), which adjust only for missing women. Separate weights by gender may be important for assessing entrepreneur populations which vary strongly by location and gender. In the majority of RSDs the proportion of men and women missing is similar, but a few RSDs with missing data have gender imbalances. However, for weighting purposes it can generally be assumed that all missing data are random. This allows the different proportions of men and women missing compared to the GRO published population to be used for weighting.[9]

Note that for supplementation purposes in Bennett, et al. (2019a) *The Age of Entrepreneurship,* adjustments for archival loss using weights were made only for 1861. The difficulty of weighting up the entrepreneur population in 1851 where whole RSDs are missing was deemed too inaccurate to be used. Entrepreneurs are very variable in spatial distribution, particularly between urban and rural areas, so that inferring from neighbouring RSDs to a missing whole RSD could lead to major distortions of analysis.

For this paper, because most missing data can be assumed to be random, the re-scaling method should generally work well to re-weight up to the estimated total employer numbers (from the BBCE supplemented data). However, there will be deficiencies where data loss affects locations with high concentrations of employers with firm-sizes that differ for the GB average. Since archival loss in 1881 is very minor and locationally random this is not important for the pilot estimates in this paper which use only 1881 data. However, for extension of the method to other years the effects of archival loss will need to be evaluated, with 1851 likely to need major re-weighting method since the main areas of loss in London and NW have higher concentrations of the larger firms than most other regions.

Part 3: Methodology for estimating the firm-size distribution

Previous analyses of BBCE, summarised above, demonstrate that GRO tabulations and estimates using record linkage are both inadequate to fully estimate non-response rates. This

[9] For a small proportion of the population (overall, less than 1 per cent), I-CeM and hence BBCE was unable to code gender (recorded as gender unknown, ‘U’), usually due to ambiguity between a person’s name, their relationship to the head of household, and the gender column where their age was recorded, as well as some mis-coding. The published census reports have no unknown sex (everyone was assigned by clerks), meaning that those with unknown gender in I-CeM cannot be weighted using gender-specific weights so that non-gendered full population weights have to be used. ‘U’ could be corrected by users checking original records.

section presents a method of full estimation for GB firms and workforce by firm size by sector for 1881. It uses the an up-scaling of the census information provided by employers in the census extractions in BBCE, recognising at the outset that the estimates possible are approximate, given the definitional ambiguities in the census and data limitations. The method developed is referred to as *data re-scaling*. It is akin to other methods used in modern censuses for post-survey data editing and quality control.

For estimating by sector a key constraint is that the GB census was based on occupations. These can be interpreted as industry sectors in many cases, but the match is poor for some sectors, and especially for maker-dealers in industries such as apparel and shoe making where businesses often made the product and sold it directly to consumers. In addition the coding of the census occupational data in BBCE/I-CeM is not accurate enough in many cases to identify detailed sub-sectors: it was inconsistently applied by the original census respondents to their actual firms, and the methods of BBCE data supplementation (although developed for 844 sub-sectors) cannot achieve high levels of robustness for highly disaggregated sectors. These challenges can be overcome by the use of more aggregate sector coding where fine distinctions are no longer required. That used here is the EA17 coding which was developed in the BBCE to give an approximate mapping from occupation to industrial sector (WP 5, Bennett et al., 2017).[10] This allows assessment for the whole economy across all sectors.

3.1 Estimation stages: Method of data re-scaling

Stage 1. The GB census was derived from BBCE separately for the England and Wales (E&W) and Scottish results, each coded to EA17.

Stage 2. The total employer numbers in 1881 were derived from (i) the 'extracted' which are those employers who responded by giving a workforce size (termed ETOT in BBCE); and (ii) the supplemented data in BBCE by the two alternative methods: NUM and IND. Estimates

[10] Note the coding of the census responses here uses I-CeM as a starting point; but the versions V1 and V2 avail bale have numerous occupational coding errors. BBCE corrects most of these, but checks were not made for all occupational strings where the N of strings was low and did not include any indictors of employer or OA status that were critical to BBCE: see BBCE *Guide* and Bennett et al. 2019a, Chapters 3-4.

by both supplementations are compared below for E&W, but for Scotland only IND is used (see WP 20, Smith et al., 2019).

Stage 3. This is the key stage of *data re-scaling*. Initially we assume the size distribution for the 'extracted' employers applies to the supplemented employer numbers and allocate by applying the proportions of extractions by employer size across all employers given by the supplementations. For example, if 10% of extracted firms in a sector employed 20 people, it was assumed that 10% of the supplemented employers in that sector had firms employing that size of workforce. This assumes that non-response was random except for the factors allowed for in the BBCE estimation of the supplemented data (sector, demography, household relationships, location, etc.). We know from previous analysis, discussed above, that *non-responses* by size and sector were random; only legal status (incorporation) affects non-response in a systematic way. However, *truncations* in the BBCE data are only random within size and sector categories; although the effects were small, there were systematic differences between sizes, being greater for larger sizes (over about 300 employees), with some variation between sectors. These effects should be already included in the supplementation process which gives aggregate compensation for the combination of non-response and truncation. The data re-scaling was then applied to the size distribution in each EA17 sector. For each sector, the re-scaling factor was applied to the difference between the extracted and supplemented employer numbers. For example, if 10% of a sector's extracted employers had a workforce of 20 workers, then 10% of the difference between the total extracted employers and the total supplemented employers were assumed to have employed 20 workers; hence, if in this case, the number of extracted employers was 50 and the total number of employers suggested by the supplementation method was 200, we would assume that 10% of 150 had 20 workers, 0.1 x 150 = 15. This up-scales the estimated deficit between the 'supplemented' data using a multiplier that varies by firm size and sector**.**

Employers at every reported workforce size are used (i.e. there is no aggregation into size categories); this ranges from 1 to 8,000 in 1881 census 'extracted' responses. The re-scaled are added to the actual extracted employers for each size, for each sector separately, to control for different size distributions by sector. This gives the size distribution for all employers and their workforces for 'all' employers. Two estimates are possible - for number of firms by size, and workforce by size. And each can be estimated from the two supplementation methods NUM and IND.

Note that the size distribution is calculated across all sizes (no categories), *for reporting comparisons, however, summary tables use size categories*. However, note that because only the existing responses are available for the 'extracted', all increments to the data are applied to existing firm at a particular size. This is unproblematic for small firms: e.g. in England and Wales all sizes from 1 to 177 (no firm responded with exactly 178 employees), and then with other small gaps up to 400 employees. But after this the gaps for sizes with no respondents become more numerous, and for sizes over 800 very frequent. This means that actual firms that did not respond at a given size, such as 178 employees, are instead estimated at the available other sizes. Since there is bunching in the original data at 10s, 100s and 1000s this means that existing bunching is increased. Various methods could be used to overcome this to give a broader spread of firm sizes for the re-scaled data, but at this preliminary stage these are not developed. The effects should, be small. But for subsequent use of the data bunching has to be taken into account (e.g. by using clustered estimation methods; as developed in section 4)

Stage 4. The workforce from these calculations can then be compared with the total workforce estimated in BBCE. BBCE gives three employment status variables: employers (used in Steps 2 and 3), workers, and own account. Each is derived by the same methods as NUM or IND. The estimated workforce should match the estimated total of workers. This is subject to evaluation as a test of accuracy below.

Stage 5. The estimates from data re-scaling give decimal values for the number of firms. These can be rounded to the nearest whole number, so 1.4 is rounded to 1 and 1.6 to 2, or they can all be rounded down, with both 1.4 and 1.6 can be rounded to 1. For large firms the data re-scaling rarely gives an increase in whole numbers for a given firm size; hence, if rounding down is used very few additional large firms are estimated. This does not accord with the expected reality, and tests show it does not fit Hannah's 'truth' data. Hence all estimates used are rounded to the nearest whole number, not exclusively rounded down. Different estimates from the BBCE NUM and IND estimates were also assessed. Some comparisons are made here and generally the differences are too small to matter; but all final estimates are based on IND.

Stages 6A. Comparison is made for worker estimates against Booth and Factory Returns, where available.

Stage 6B. Comparison is made for firm-size estimates against Hannah's 'truth' data for manufacturing with 1,000 or more employees.

Subsequent stages: The effect of incorporation is included in interpretation of the numbers after the estimation stages, where it is generally concluded that the re-scaling method mages to include these (implicitly); but for the largest firms adjustments are needed to the re-scaling method.

Own account proprietors (sole proprietors with no employees) (OA) are a separate category in the BBCE estimates; they are excluded from the workforce estimates here since each is a single-worker firm. But it is important to bear in mind that, as well as workers and employers, there were self-employed individuals with no employees. They are reported in the firm-size tables, by sector, as a separate element.[11]

The calculations are undertaken s*eparately* for England and Wales, and Scotland; then combined to give GB totals. No individual returns from the Irish census survive for this period in significant numbers so that this paper uses only GB data. Irish data has been added for the largest firms in a separate exercise for manufacturers by Hannah to give UK estimates (a version of the main data is available in appendices to Hannah and Bennett, 2021); this is not replicated here for the rest of firms to estimate re-scaled UK totals.

[11] Note, as above, BBCE has some employers and own account in the *worker-only* EA17 sectors 14-17 because of occupational coding errors in I-CeM that have not been corrected. They are referred to as 'unattributed' in the following analysis.

3.2 Estimates of workforce numbers

A key test of alterative estimates is how far workforce by firm size matches the BBCE estimates of total workers. In Table 2 the BBCE this estimate of workers is in the left hand column; this is used as the ‘actual’ number, but is an estimate.

Ideally workers should be equal to the estimates (for NUM cols; and for IND cols). IND usually gives lower estimates (as expected). In each case the Rounded down gives lower estimates of total workers because of the large number of small decimal estimates involved, as expected. There are also issues for unattributed workers.

For interpreting these tables we expect differences from:

1. Any discrepancies of the census ‘extracted’ size distribution from actual. We are assuming under-reporting and truncation is proportionally the same for all sizes. But this may be incorrect: e.g. retail, professional services, and finance and commerce look over-reported, or over-weighted, compared to most sectors. This may be because the response rates here were actually very complete compared to other sectors. Also major gaps in reporting in manufacturing may come from absence of many Ltd Co., mainly affecting textiles, iron/steel, ships, railway engineering. There are also certainly coding errors of the sector attributions.

2. Classification issues for a ‘sector’ vs. an occupation. For these the extracted firms will be smaller than the number implied by the total workers in that sector; e.g. many blacksmith workers were employed in businesses that were not blacksmithing firms so the extracted workforce numbers will be smaller than the actual blacksmith workers; similarly for construction. The same will apply in other sectors; e.g. construction, with many carpenters, painters etc. employed in non-construction industries. It is also very difficult to differentiate textile occupations such as ‘weaver’ or ‘spinner’ between worker, own account and employer. This is a key issue in the estimates noted later below. Other classification issues also arise from how occupations were treated in I-CeM.

3. Sectors with a large OA component. We do not know how far the OA, and the 1 or 1-3 size category that may contain family of other relatives that were sometimes reported or not,

overlap or interact and therefore create problems for the estimation. Some OA are probably workers in BBCE, and others should be in the 1-employee category. Also if some OA are misstatements and were are actually employees then the NUM or IND worker totals will be too low and the workforce returns from employers should be higher and the estimates generated from them will be too high, and vice versa. This may occur, for example, when individuals were gang leaders and masters who sub-contracted labour. This may explain some discrepancies between sectors that have different census reporting or workplace organisation.

Table 2.1. Workforce numbers, E&W; highlighted where estimates match reasonably well

EA17		Actual Supplemented IND	BBCE NUM		BBCE IND	
			Rounded	Rounded down	Rounded	Rounded down
1	Agriculture	1151784	819419	795393	778785	775145
2	Mining & quarrying	475071	462416	424780	331686	305612
3	Construction	640362	645445	617844	479638	478150
4	Manufacturing	2145650	2660401	2656440	1974752	1914405
5	Maker-dealers	678563	535809	534760	383570	354207
6	Retail	198632	454858	452627	431240	408990
7	Transport	610928	157128	157017	90726	88905
8	Professional services	269359	883598	846845	814195	813231
9	Personal services	431902	140652	140652	83648	81451
10	Agricultural processing	75525	137344	134698	128011	127624
11	Food retailing	214408	255819	246016	214889	204453
12	Lodging & refreshment	106596	103488	101631	105118	104924
13	Finance & commerce	89186	429061	418088	428791	418023
14-17	Unattributed	2428777	2820	2820	2776	2776
	Total	9516743	7688258	7529611	6247825	6077896
	Total excl. unattributed	7087966	7685438	7526791	6245049	6075120

Table 2.2. Workforce numbers, Scotland; only possible from BBCE by IND method

EA17		Actual Supplemented	BBCE IND	
			Rounded	Rounded down
1	Agriculture	202441	145710	144708
2	Mining & quarrying	74179	12367	9040
3	Construction	87762	151051	150059
4	Manufacturing	343579	420032	312390
5	Maker-dealers	86214	70806	64646
6	Retail	23789	124223	122439
7	Transport	74330	17799	17728
8	Professional services	38162	34489	34384
9	Personal services	33562	8104	8080
10	Agricultural processing	8762	21659	20486
11	Food retailing	31273	54833	54426
12	Lodging & refreshment	8898	20339	20255
13	Finance & commerce	13637	69709	64969
14-17	Unattributed	312002	619	619
	Total	1338590	1151740	1024229
	Total excl. unattributed	1026588	1151121	1023610

Table 2.3. Workforce numbers, GB; only possible on consistent basis from BBCE for the IND method

EA17		Actual Supplemented	BBCE IND	
			Rounded	Rounded down
1	Agriculture	1355411	924495	919853
2	Mining & quarrying	550591	344053	314652
3	Construction	756285	630689	628209
4	Manufacturing	2507648	2394784	2226795
5	Maker-dealers	816021	454376	418853
6	Retail	218951	555463	531429
7	Transport	698941	108525	106633
8	Professional services	308647	848684	847615
9	Personal services	475974	91752	89531
10	Agricultural processing	83117	149670	148110
11	Food retailing	235095	269722	258879
12	Lodging & refreshment	108742	125457	125179
13	Finance & commerce	104020	498500	482992
14-17	Unattributed	2740763	3395	3395
	Total	10960206	7399565	7102125
	Total excl. unattributed	8114554	8836559	7098730

Also if all OA are thought of as 1-employee firms they should be added to the employers; then those employer estimates will be increased reducing any under-estimation. The combined group of 1-employee and OA is very large, whichever way they are counted. They are kept apart in subsequent discussion but a proportion can be added if desired; some comparisons are made below. The analysis here suggests they are not a major influence on the aggregate but strongly affect some sectors and this is likely to explaining some of the over-estimation in maker-dealer sectors, retailing and some manufactures.

4. Additionally there are issues arising from multi-sector (portfolio) firms that have employees in various 'industries' but are captured in employer responses under what they primarily list. What has been used here is the BBCE coding of 'main' activity as recorded in

the census, after removing as far as possible any census bias to put farming first (see Bennett et al., 2019a, Chapters 3 and 11). However, Hannah's checks on large firms indicate that the main activity allocated is occasionally different from what it should be on a rational assignment based on workforce numbers of the 'main' business. Understandably the census, and BBCE/I-CeM, allocations can sometimes be imperfect, and they may differ from how a proprietor saw their firm or chose to respond to the complex occupation question in the census. There is not much that can be done about this, but it confirms that there will be fuzziness in sector allocations.

The same issue applies to businesses that integrate across several stages of production and sales; e.g. in Hannah's data on firms of 1,000 employees and over in manufacturing, a lot of employees were not in manufacturing industry, but in mining, warehousing or commerce. Hannah's rough estimate for firms he classed as significant manufacturers, is that those reporting over 2,400 employees included 26% of employees in other sectors (principally mining, with most of the rest in distribution), but for those employing 1,000-2,400 this declined to 8%. It is probable that the smaller the company down the size distribution the more negligible the effect. This was a major phenomenon in iron & steel (with backward integration to coal and iron ore, and forward integration to engineering products). It is a problem that occurs in all classification systems, and also bedevils SIC coding in modern times. It means that overall all the sector codings have to be treated as fuzzy at the margins, with Hannah's data suggesting the problem is probably significantly greater the larger the firm, because they use less outsourcing and self-supply activities like raw materials and warehousing or transport, as well as commercial activities. Because of the uncertainness of making appropriate adjustments no attempt is made in the re-scaling used her to compensate for sector misallocations; the EA17 sectors are treated as approximate and fuzzy.

3.3. Estimates of firm size: aggregate data (all sectors)

The main aim is estimating the firm size distribution; but firm size also helps to interpret workforce estimates. Later in the paper, the assessment of cross tabulations by size and sector need to focus on the largest firm categories which make most difference to the workforce estimates; these are also most important in subsequent interpretations about industry concentration and development of the economy, and are most sensitive to method.

Estimates were made across *all* firm sizes not by classes, but to simplify presentation the tables are presented in 14 size categories. Table 3 compares the total firm size estimates *across all sectors* with the actual total number of employees (ETOT) in BBCE from the self-identified 'extracted' employers. Ratios of differences in the estimates between the rounded IND method and ETOT are shown in the final column. The fit is closest to the average for the firms of 5-10, or 5-20 employees, with the largest firms generally having the largest differences.

Table 3.1. Firm Size England and Wales

Firm size E&W	Actual Extracted ETOT	BBCE NUM		BBCE IND		Weight R vs ETOT
		Rounded	Rounded down	Rounded	Rounded down	
1	34419	102533	102529	86481	86472	2.51
2	32337	96575	96565	81403	81394	2.52
3	34751	68086	68079	57764	57758	1.66
4	16535	47447	47444	40379	40374	2.44
5-9	38007	107019	106979	91214	91179	2.40
10-15	15984	45598	45550	38777	38736	2.43
16-19	4613	13326	13295	11353	11331	2.46
20-49	10677	34838	34655	28937	28757	2.71
50-99	2881	10917	10726	8835	8642	3.07
100-199	1483	5774	5638	4626	4455	3.12
200-249	354	1415	1382	1125	1066	3.18
250-499	624	2360	2289	1826	1710	2.93
500-999	309	1353	1315	1075	1030	3.48
≥1000	122	575	545	454	431	3.75
Total	193,098	537,816	536,991	454,249	453,335	2.35

Table 3.2. Firm Size Scotland ; IND only

Firm size Scot	Actual Extracted ETOT	BBCE IND		Weight R vs ETOT
		Rounded	Rounded down	
1	5834	15575	15569	2.67
2	7285	17892	17884	2.46
3	6073	13815	13807	2.27
4	4817	10695	10689	2.22
5-9	9694	22133	22103	2.28
10-15	2942	7418	7385	2.52
16-19	823	2135	2118	2.59
20-49	1748	4892	4773	2.80

50-99	498	1269	1189	2.55
100-199	270	688	609	2.55
200-249	81	209	171	2.58
250-499	147	321	257	2.18
500-999	85	214	162	2.52
≥1000	27	73	54	2.70
Total	40,324	97,329	96,770	2.41

Table 3.3. Firm Size GB; IND only

Firm size GB	Actual Extracted ETOT	BBCE IND		Weight R vs ETOT
		Rounded	Rounded down	
1	40253	102056	102041	2.54
2	39622	99295	99278	2.51
3	40824	71579	71565	1.75
4	21352	51074	51063	2.39
5-9	47701	113347	113282	2.38
10-15	18926	46195	46121	2.44
16-19	5436	13488	13449	2.48
20-49	12425	33829	33530	2.72
50-99	3379	10104	9831	2.99
100-199	1753	5314	5064	3.03
200-249	435	1334	1237	3.07
250-499	771	2147	1967	2.78
500-999	394	1289	1192	3.27
≥1000	149	527	485	3.56
Total	233,420	551,578	550,105	2.36

Key points to note are:

1. For **E&W** the total number of firms estimated between NUM and IND by 20%, but differs little within them by rounding or rounding down. Rounding down has most effect proportionally for the larger firms because more of the estimates produced for those larger firms have small decimal proportions removed by rounding down. However, the choice of rounded or rounding down has little consequence for the estimates of the number of small firms, especially those under 20 employees; though of course there remain embedded issues about the response rate and inclusion of the smallest of 1, or 2 and 3 employees which are not dealt with explicitly by the estimation method. In terms of the effect on workforce estimates, the differences that have most impact differ between small or large firms. For small firms, and sectors with a higher proportion of small firms, the main differences in estimates derive from the choice between NUM and IND methods and not rounding.

2. Substantial re-scaling is needed to take account of truncation and non-responses. The final column of the tables show the up-weighting required to align those 'extracted' in BBCE to give ETOT worker numbers equal to the estimates by size based on the IND method with rounding. For GB, the estimates suggest an under-response/truncation in the census that ranges requiring an average up-weight of 2.36, but varying by size. The small firm sizes generally are close to the average, but for firms over 50 employees the up-scaling ranges from about 3 to 3.5. This is slightly lower than the supplemented non-response estimate of 2.9 for E&W (Section 2.1) because of using rounding and for GB. The patterns are similar in E&W and Scotland.

3. There are two size categories that are exceptional. First, at the smaller size, firms of 3-employees in GB and E&W, or 3-9 employees in Scotland, need a lower up-scaling than average; very much lower in E&W and hence in GB. This marked contrast with the other smaller size categories suggests that this is affected by misallocation of partners for firms vs. the 1- and 2-employee firms in E&W. In Scotland this effect is less marked, but still exists and appears to affect a wider range of size categories upwards to include the 4 and 5-9 sizes.

The second exception is the 250-499 size class. While the re-scaling generally increases as firm size increases, this size class has re-scaling lower than the adjacent sizes. It is difficult to believe that those firms of 250-499 employees were better at responding than those of 200-249 or 500-999. It is likely to derive from differences between sectors that are particularly concentrated in this size class compared to those above and below.

The exceptions in these results are a warning that the re-scaling the data has as far as possible to take account of the unique features of employer responses and data capture by sector and firm size, with disaggregation into small categories by sector and size inherently difficult.

3.4. Estimates of firm size by sector

One of the key outputs form this paper is estimating firm size by sector. Using the BBCE supplemented estimates allows most of the main differences in non-response and truncation to be dealt with. At this point, given the results above, we restrict attention to IND and use rounded-up estimates. The main concerns are for the 6-7 largest firm categories of over 20 or

50 employees because these have most impact on the workforce estimates; large firms are also most important in subsequent interpretations about industry concentration and development of the economy.

Own-account self-employed without employees (OA) estimates are given at the foot of each table. A small number of residual employers and own account that cannot be coded directly from BBCE, because they are wrongly coded in I-CeM/BBCE and have *not* been corrected, are incorrectly in the worker-only sectors 14-17 (these were grouped as ‘unattributed’ in the previous tables). Those in sectors 14-17 are not numerous, but they contain some medium and larger firms of great significance. They have been mis-coded because of the complexity of their occupational descriptor strings. Mis-coding also occurs in all sectors so that the tables can only be taken as a preliminary estimate. Estimates are particularly sensitive to mis-coding of the largest size classes which need re-coding in final estimates

The estimates of firm size by sector are given in Table 3.4, with totals for each sector for all firms employing anyone else, all OA proprietors employing only themselves, and the total of all firms that were employers and OA. The tables provide estimates of all economically active entrepreneurs (as defined in BBCE). For definitions of economically active see Bennett et al. (2017, 2019a, 2019b).

Table 3.4.1. England and Wales estimates of N of firms by size by EA17 sector

Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	total	14-17
1	26838	266	5753	5700	11983	5309	1205	1400	3420	1964	14961	6445	1237	0	0	0	0	86481	0
2	23975	260	6563	6817	10445	5870	1242	2367	2581	2128	12684	5100	1371	0	0	0	0	81403	0
3	17636	255	5362	5344	6178	4773	791	1756	1646	1638	7606	3273	1506	0	0	0	0	57764	0
4	12630	204	4293	4140	3821	3525	568	1378	956	1112	4359	2264	1129	0	0	0	0	40379	0
5-9	29342	719	10602	10409	7375	8941	1146	4102	2092	3039	7094	3554	2796	1	0	0	2	91214	3
10-15	11950	385	4603	5351	2754	4052	594	2279	839	1432	1789	1323	1424	0	0	0	2	38777	2
16-19	3312	119	1352	1792	684	1208	181	778	180	489	371	403	484	0	0	0	0	11353	0
20-49	5438	421	3673	6338	1963	2889	448	2976	649	1114	722	445	1859	2	0	0	0	28937	2
50-99	495	201	920	3142	604	752	104	1306	142	261	107	99	701	1	0	0	0	8835	1
100-199	69	133	289	2233	270	283	41	728	32	66	44	33	405	0	0	0	0	4626	0
200-249	11	63	53	525	100	74	16	243	0	9	4	0	27	0	0	0	0	1125	0
250-499	11	124	80	1027	102	115	32	220	33	12	16	0	54	0	0	0	0	1826	0
500-999	3	63	40	501	33	20	25	220	0	0	8	0	162	0	0	0	0	1075	0
≥1000	4	75	8	167	9	10	5	110	0	3	8	0	54	2	0	0	0	454	2
All firms	**131714**	**3288**	**43591**	**53486**	**46321**	**37821**	**6398**	**19863**	**12570**	**13267**	**49773**	**22939**	**13209**	**6**	**0**	**0**	**4**	**454250**	**10**
OA	97919	1878	27227	67376	207079	114963	13204	21905	119816	10704	137991	98485	12070	16	4	0	4	930641	24
Firms + OA	229633	5166	70818	120862	253400	152784	19602	41768	132386	23971	187764	121424	25279	22	4	0	8	1384891	34

Table 3.4.2. Scotland estimates of N of firms by size by EA17 sector

Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	total	14-17
1	3353	15	2146	720	3253	1225	232	243	294	315	2108	1245	417	4	5	0	0	15575	9
2	5224	38	2409	802	2958	1176	169	258	222	414	2640	1028	549	4	1	0	0	17892	5
3	5274	32	1752	583	1598	886	116	350	139	334	1786	637	322	2	3	1	0	13815	6
4	4393	20	1652	472	913	580	125	198	108	221	1185	420	405	1	1	1	0	10695	3
5-9	8334	75	3812	1351	1711	1695	192	516	185	600	2009	970	669	6	6	2	0	22133	14
10-15	1882	42	1795	758	557	716	70	260	56	200	540	287	252	2	1	0	0	7418	3
16-19	412	6	574	305	143	285	30	15	20	46	140	86	72	0	1	0	0	2135	1
20-49	526	108	1332	897	357	566	67	211	95	250	193	85	204	1	0	0	0	4892	1
50-99	34	21	248	496	109	206	16	15	10	43	35	0	36	0	0	0	0	1269	0
100-199	2	24	51	354	43	77	12	60	10	10	21	0	24	0	0	0	0	688	0
200-249	0	0	16	113	23	21	4	0	0	5	3	0	24	0	0	0	0	209	0
250-499	1	3	6	243	3	46	0	0	0	0	6	0	12	1	0	0	0	321	1
500-999	0	3	0	139	6	31	8	15	0	0	0	0	12	0	0	0	0	214	0
≥1000	0	0	3	41	0	10	4	0	0	0	3	0	12	0	0	0	0	73	0
All firms	**29435**	**387**	**15796**	**7274**	**11674**	**7520**	**1045**	**2141**	**1139**	**2438**	**10669**	**4758**	**3010**	**21**	**18**	**4**	**0**	**97329**	**43**
OA	33738	261	4146	16311	27656	14181	6162	2253	9799	907	15773	6653	2145	21	18	4	0	140028	43
Firms + OA	63173	648	19942	23585	39330	21701	7207	4394	10938	3345	26442	11411	5155	42	36	8	0	237357	86

Table 3.4.3. GB estimates of N of firms by size by EA17 sector

Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	total	14-17
1	30191	281	7899	6420	15236	6534	1437	1643	3714	2279	17069	7690	1654	4	5	0	0	102056	9
2	29199	298	8972	7619	13403	7046	1411	2625	2803	2542	15324	6128	1920	4	1	0	0	99295	5
3	22910	287	7114	5927	7776	5659	907	2106	1785	1972	9392	3910	1828	2	3	1	0	71579	6
4	17023	224	5945	4612	4734	4105	693	1576	1064	1333	5544	2684	1534	1	1	1	0	51074	3
5-9	37676	794	14414	11760	9086	10636	1338	4618	2277	3639	9103	4524	3465	7	6	2	2	113347	17
10-15	13832	427	6398	6109	3311	4768	664	2539	895	1632	2329	1610	1676	2	1	0	2	46195	5
16-19	3724	125	1926	2097	827	1493	211	793	200	535	511	489	556	0	1	0	0	13488	1
20-49	5964	529	5005	7235	2320	3455	515	3187	744	1364	915	530	2063	3	0	0	0	33829	3
50-99	529	222	1168	3638	713	958	120	1321	152	304	142	99	737	1	0	0	0	10104	1
100-199	71	157	340	2587	313	360	53	788	42	76	65	33	429	0	0	0	0	5314	0
200-249	11	63	69	638	123	95	20	243	0	14	7	0	51	0	0	0	0	1334	0
250-499	12	127	86	1270	105	161	32	220	33	12	22	0	66	1	0	0	0	2147	1
500-999	3	66	40	640	39	51	33	235	0	0	8	0	174	0	0	0	0	1289	0
≥1000	4	75	11	207	9	20	9	110	0	3	11	0	66	2	0	0	0	527	2
All firms	**161149**	**3675**	**59387**	**60760**	**57995**	**45341**	**7443**	**22004**	**13709**	**15705**	**60442**	**27697**	**16219**	**27**	**18**	**4**	**4**	**551579**	**53**
OA	131657	2139	31373	83687	234735	129144	19366	24158	129615	11611	153764	105138	14215	37	22	4	4	1070669	67
Firms + OA	292806	5814	90760	144447	292730	174485	26809	46162	143324	27316	214206	132835	30434	64	40	8	8	1622248	120

Part 4: Comparisons with large scale contemporary data

Two source for comparisons with the BBCE estimates are available at sufficient scale to constitute equivalent national GB coverage, though both have deficiencies in the way they cover the employers and firms we are interested in: Booth (1886) and the Factory Returns. Both can be used to assess comparisons of workforce and firm size.

4.1. Workforce comparisons with Booth

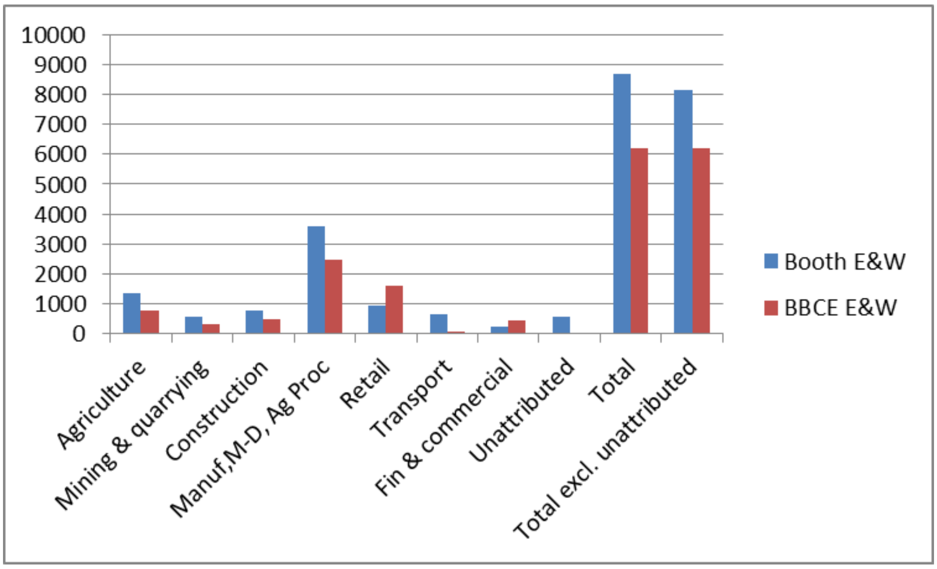
An important historical comparator on a sector basis is Booth's analysis (1886, pp. 414-5 for E&W; pp. 418-9 for Scotland). However, Booth does not attempt to separate employers from the rest of the working population in a sector; this means that OA in particular affect the accuracy of comparisons possible. Also a comparison of his sectors is approximate, as Booth clearly uses different definitions of some sectors, mainly of retail and commerce, and attributes more of the unattributable than attempted here. He must have included own account and some employers within his totals, insofar as they were included in the GRO published tabulations, but the extent is unclear. He excluded those of property and rank (Booth, 1886, p. 323), as done in BBCE (unless they reported workforces, data which Booth would not have been able to access). Those BBCE employers and own account remaining in the worker-only EA17 sectors 14-17 (because of occupational coding errors in I-CeM) differ from the unattributed in Booth; but workers in Booth and BBCE that are unattributed should be identical (general labourers, clerks, etc.). Booth also gives estimates only to the nearest 00s.

Table 4.1 and Figure 1 give the comparisons. The OA make a large difference in some sectors. Comparisons of the BBCE data table and in the figs with Booth show that the estimates accord generally quite well in relative terms, but overall Booth has larger numbers in all but two categories, especially in the large categories of agriculture and manufacturing (which also include all maker-dealers (M-D) and agricultural processing - EA17 sectors 5 and 10); he has smaller numbers in retail and finance & commerce (mainly EA17 6 and 13). This is consistent across E&W and Scotland. The differences are mainly because the estimates here from BBCE take no account of OA numbers, but the higher estimates in BBCE are surprising in the cases of retail and finance & commerce which also have many OA. There appear to be some questions about how Booth managed to allocate OA in these as well as the other categories.

Table 4.1. Comparison of Booth and BBCE workforce estimates using IND; E&W

EA17		Booth E&W	BBCE E&W
1	Agriculture	1371	779
2	Mining & quarrying	562	332
3	Construction	796	480
4, 5, 10	Manuf, M-D, Ag Proc	3599	2486
6, 8, 9, 11, 12	Retail	924	1613
7	Transport	654	90
13	Fin & commercial	225	429
	Unattributed	560	3
	Total	8691	6212
	Total excl. unattributed	8131	6209

Figure 1.1. Booth and BBCE compared: E&W



Other differences derive form how occupations are allocated by Booth, compared to BBCE. Perhaps the major sector affected is railways and some other transport. Railways workers often are listed as labourers, some specifically ‘railway labourers’, ‘road labourers’, etc. in the census; and many other occupations such as ‘plate layers’, ‘paviours’, etc. Booth attempted to allocate these to transport categories, and also attribute general labourers to these categories as far as feasible. BBCE did not attempt reallocations of general categories where there was insufficient information; and a category such as ‘railway labourers’ would be attributed to Occode 132 in BBCE/I-CeM, the category entitled ‘Railway labourer (not railway contractor’s labourer’), this occode category is coded to EA17 sector 7 ‘Transport’. This has major consequences. Booth puts 139,500 employees in EW railways in 1881, and the 1884 Rail Return has 312,017. Booth puts platelayers employed by railways in “roadmaking” and probably those shipping and restaurant/hotel staff are also somewhere other than in the railways that often employed them (which is also the same in BBCE). As a result, Booth underestimates railway employment considerably, and BBCE underestimates by even more (though it should be neutral between shipping and rail within EA17 sector 7).

Table 4.2. Comparison of Booth and BBCE workforce estimates using IND; Scotland

EA17		Booth Scot	BBCE Scot
1	Agriculture	265	146
2	Mining & quarrying	84	12
3	Construction	111	151
4, 5, 10	Manuf, M-D, Ag Proc	557	513
6, 8, 9, 11, 12	Retail	123	187
7	Transport	85	18
13	Fin & commercial	34	70
	Unattributed	68	1
	Total	1327	1098
	Total excl. unattributed	1259	1097

Figure 1.2. Booth and BBCE compared: Scotland

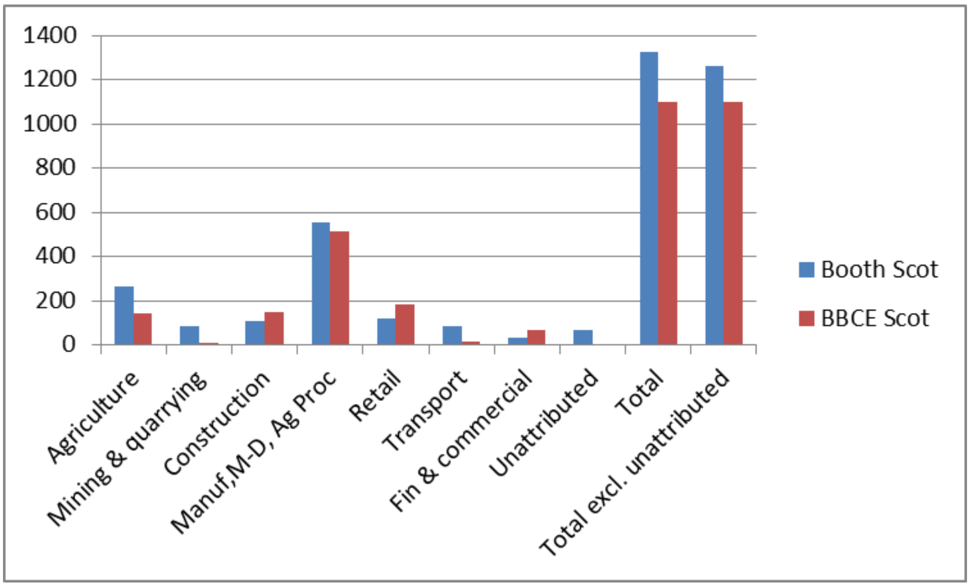
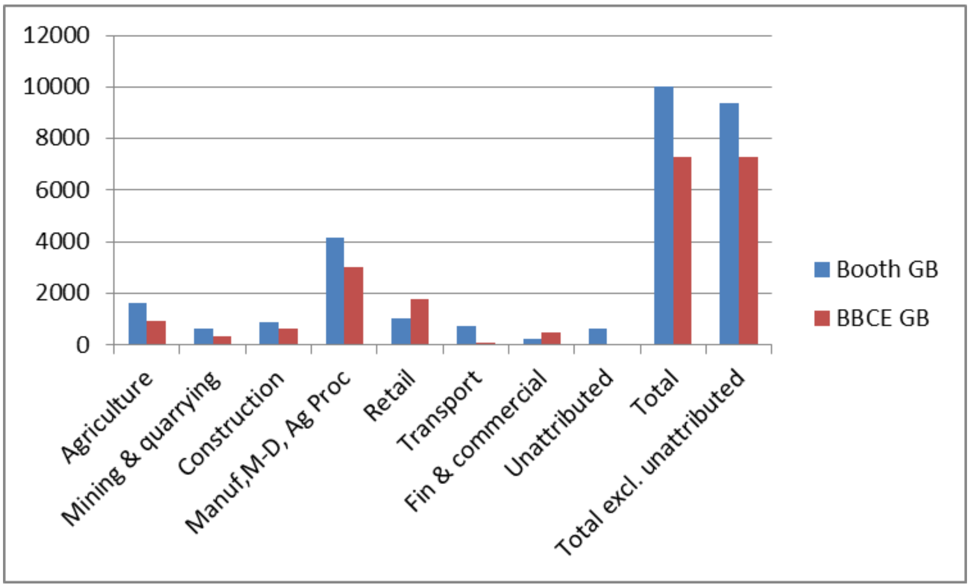


Table 4.3. Comparison of Booth and BBCE workforce estimates using BBCE IND; GB

EA17		Booth GB	BBCE GB
1	Agriculture	1636	925
2	Mining & quarrying	646	344
3	Construction	907	631
4, 5, 10	Manuf, M-D, Ag Proc	4156	2999
6, 8, 9, 11, 12	Retail	1047	1800
7	Transport	739	108
13	Fin & commercial	259	499
	Unattributed	628	4
	Total	10018	7310
	Total excl. unattributed	9390	7306

Figure 1.3. Booth and BBCE compared: GB



If the OA are added to the EA17 estimates to compare with Booth we get a closer match, especially in Scotland (central columns in Table 2 and Figure 2). This is a good confirmation that the basic estimation method for the workforce works quite well. The main deficiency in estimation of absolute numbers for GB is under-estimation for agriculture, manufacturing and transport, and over-estimation for retail. The comparisons after removing the unattributed from Booth are better. It appears that the largest discrepancy with Booth arises from the treatment of dealers that could be assigned to retail or in sectors such as manufacturing as maker-dealers; with a smaller secondary effect from how unattributed workers are allocated.

This is to be expected. Later attempts at making this division (Lewis, Feinstein etc.) have agreed that Booth underestimates manufacturing.

Table 4.4.1. Comparison of Booth and BBCE workforce estimates using IND; E&W

EA17	Booth E&W	BBCE+OA	BBCE+OA&EMP
Agriculture	1371	876	1007
Mining & quarrying	562	333	336
Construction	796	507	551
Manuf,M-D, Ag Proc	3599	2772	2885
Retail	924	2106	2249
Transport	654	103	109
Fin & commercial	225	441	454
Unattributed	560	3	3
Total	8691	7141	7594
Total excl. unattributed	8131	7137	7591

Table 4.4.2 Comparison of Booth and BBCE workforce estimates using IND; Scotland

EA17	Booth Scot	BBCE+OA	BBCE+OA&EMP
Agriculture	265	180	328
Mining & quarrying	84	12.3	84.6
Construction	111	155	131
Manuf,M-D, Ag Proc	557	558	623
Retail	123	236	198
Transport	85	24	101
Fin & commercial	34	72	66
Unattributed	68	1	68
Total	1327	1238.3	1599.6
Total excl. unattributed	1259	1237	1531.3

Table 4.4.3. Comparison of Booth and BBCE workforce estimates using IND; GB

EA17	Booth GB	BBCE+OA	BBCE+OA&EMP
Agriculture	1636	1056	1927
Mining & quarrying	646	345.3	650.6
Construction	907	662	998
Manuf,M-D, Ag Proc	4156	3330	4621
Retail	1047	2342	1758
Transport	739	127	774
Fin & commercial	259	513	316
Unattributed	628	4	628
Total	10018	8379.3	11672.6
Total excl. unattributed	9390	8374	11043.3

Figure 2.1. Booth and BBCE compared after adding OA to BBCE: E&W

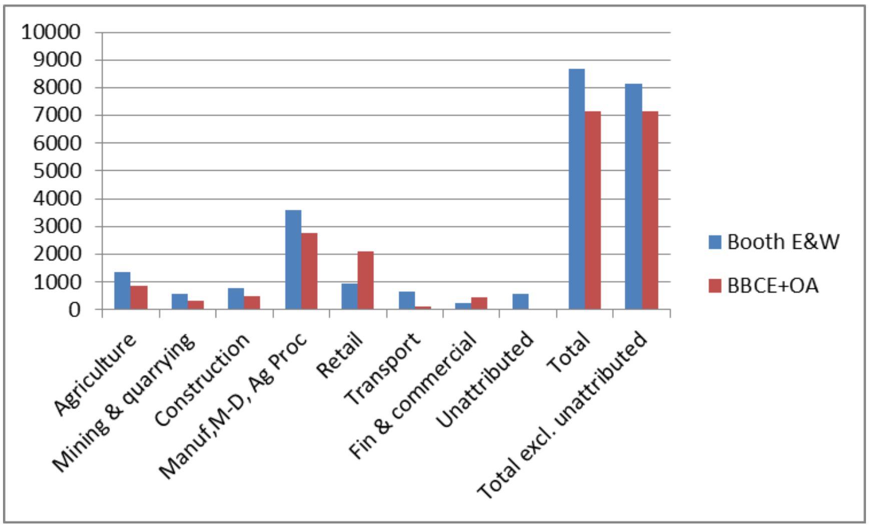


Figure 2.1. Booth and BBCE compared after adding OA to BBCE: Scotland

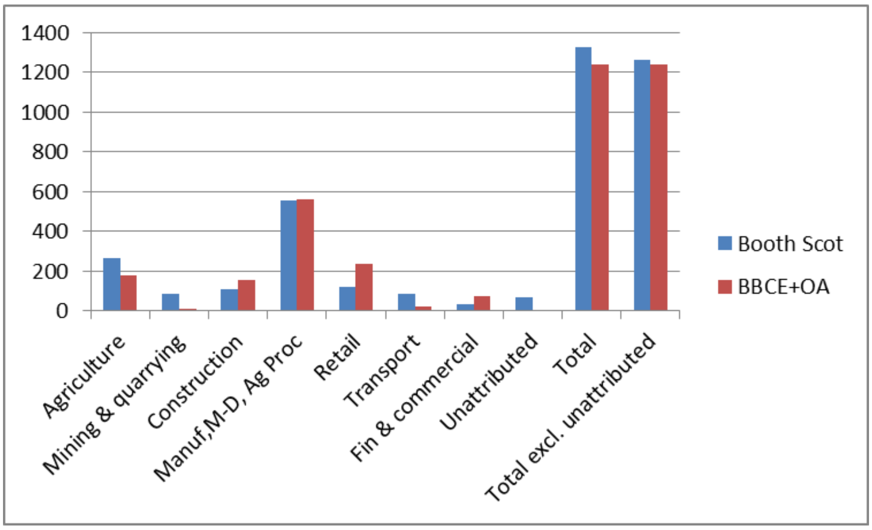
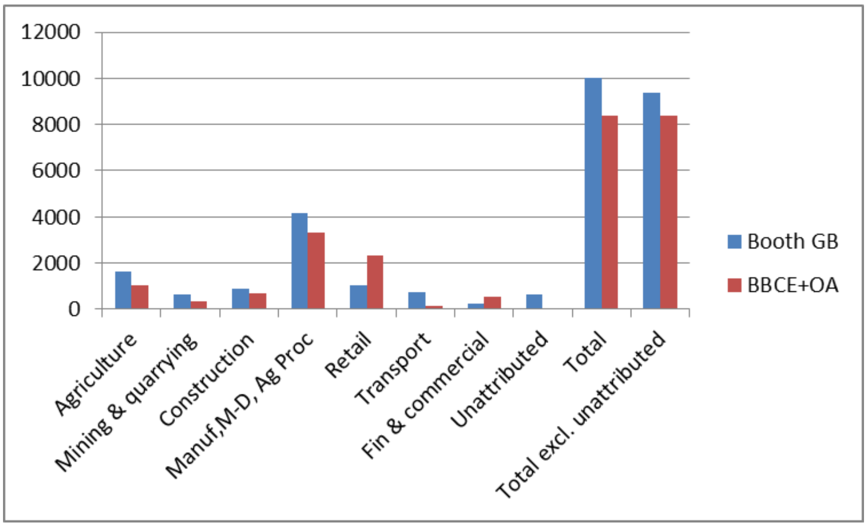
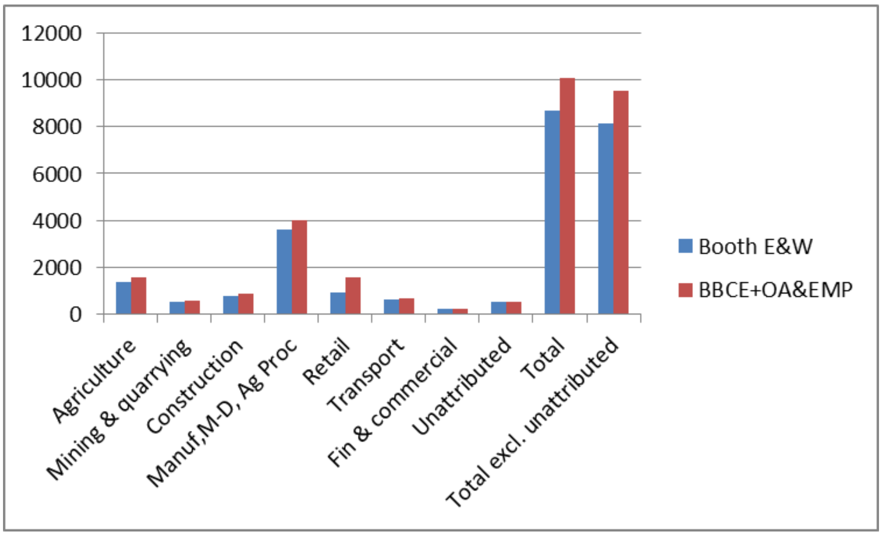


Figure 2.1. Booth and BBCE compared after adding OA to BBCE: GB



If we also add employers as well as OA to the BBCE EA17 estimates to compare with Booth (right hand cols of the tables above) we get an even closer match. Now BBCE numbers exceed Booth on average and in almost all sectors. The main absolute difference is now between retail, and to a lesser extent agriculture and manufacturing. These differences again are likely to derive mainly from the maker-dealer issue.

Figure 3.1. Booth compared after adding OA and employers to BBCE: E&W



The patterns are now almost identical between E&W and Scotland, which may result from the difficulties of assigning between E and OA in Scotland (see WP 20, Smith et al., 2019). Overall now the BBCE estimates give higher numbers in total, mainly deriving from larger numbers in agriculture, manufacturing and retail. This is an excellent confirmation that the basic estimation method for the workforce works well. It suggests that elements of the Booth estimates contained in his last table entries should be reassigned to the economically active. This is probably the elements he lists as law, medicine, art and amusement, literature and science, education, property owning, and indefinite. These are all allocated in the BBCE EA17 sectors for those individuals who are running private business rather than working state hospitals, schools, etc. which Booth does not seek to differentiate; this is actually quite major limitation in the value of Booth's estimates for workers vs. proprietors at the 1881 date.

Figure 3.2. Booth a compared after adding OA and employers to BBCE: Scotland

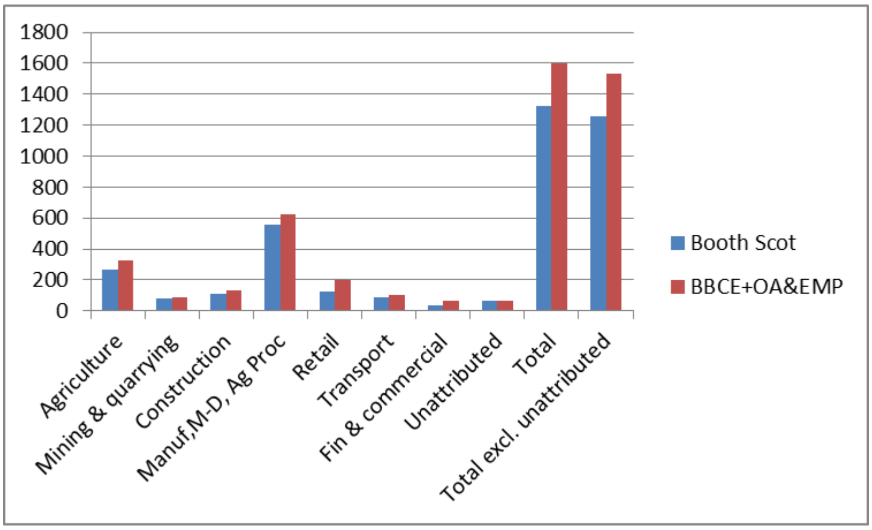
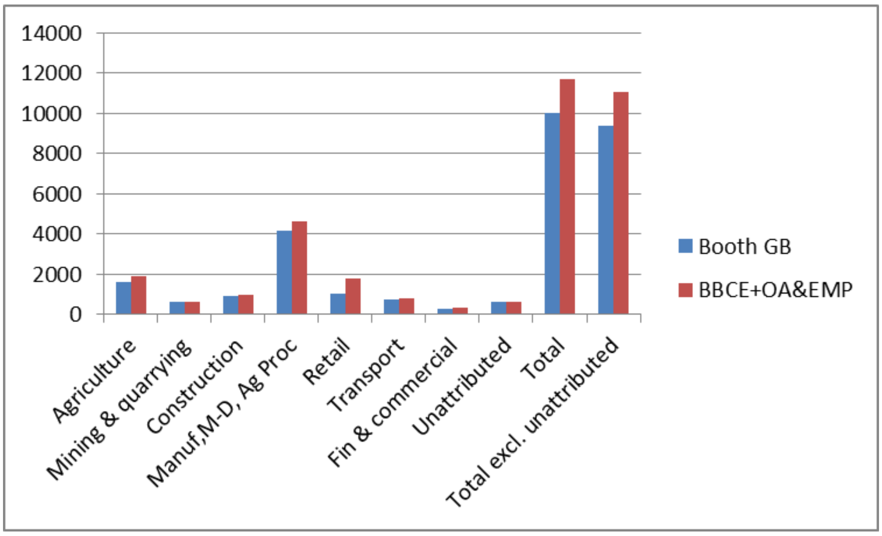


Figure 3.3. Booth compared after adding OA and employers to BBCE: GB



In the comparison exercise for manufacturing alone, the effect of corporates is difficult to discern and is ambiguous. From the result here it appears that *the supplementation method used in BBCE to allocate between employers, own account and workers, when combined with the estimation of firm numbers by size, gives effective workforce estimates that handle the deficiencies of responses by corporate employers in the same way as non-corporates*. In effect the method of supplementation and estimation combines the corporate and non-corporates employers, by assuming that the employers who did respond by size are representative of the size distribution across both business forms. The Bennett and Hannah (2021) analysis suggests that this should not be satisfactory, especially for the largest businesses. As a result an approach to how best to include the largest corporates is developed in Part 5 below.

4.2. Workforce comparisons with Factory Returns (FR)

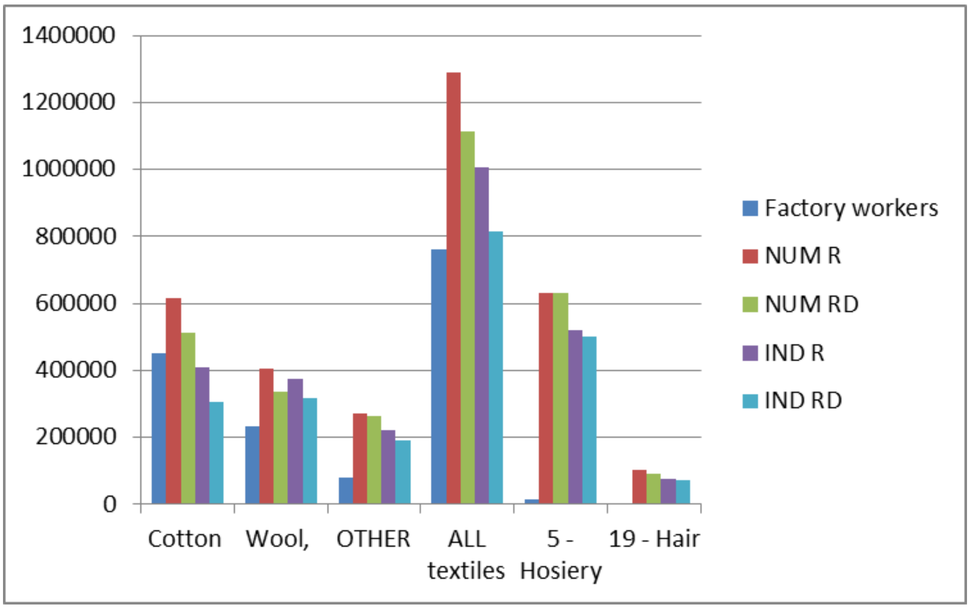
The Factory Returns (FR) are useful as they identify firms (and a mix of workshops, enterprises and sites), something not reliably available from the census. The FR are probably nearer to a definition of plants than firms, and BBCE nearer to firms than plants. The 1885 FR are probably nearer to plants that the 1878 FR. Also BBCE includes some outworkers as employees. Generally outworkers do not appear to have been recorded in the census, though this is inconsistent between respondents and difficult to fully detect; BBCE as far as possible tried to exclude them (Bennett et al., 2019a, p.72-3). FR attempted to completely exclude them.

The FR for 1878 are closest to 1881, but the FR for 1885 are also used below (PP, 1879, 1885). They give data on workforce numbers that allow comparison of worker estimates. Although they are only available for various textiles sectors,[12] they are a valuable guide since this is an important group of industries that constitute 19% of firms ≥10 employees in the BBCE for E&W for all sectors, and 47% of those with ≥100 employees. Hence, they are one of the most significant groups of sectors where firms are spread across the medium and larger size classes [though the BBCE data show they had declined compared to 1851 where textiles were 24% of firms ≥10 employees, and 57% of those ≥100 employees: Bennett et al., 2019a, Chapter 5].

We know that the FR misses workers in small units and in early reports had partial coverage for some areas. In the comparison exercise here there is some general accordance for cotton, wool, but a poor match for other textiles because of FR coverage; with very limited utility of comparison for hosiery and hair which FR hardly covers at all (Figure 3.4). Including OA leads to greater discrepancies and has only small effects except for hosiery. The FR gives slightly better coverage of wool and other textiles in 1885 than 1878, but are very similar.

[12] Earlier FR coverage includes more manufacturing sectors and can be used for comparisons with earlier censuses for 1851-71, but are too distant in time to be useful here.

Figure 3.4. Estimates of workers compared to 1878 Factory Returns E&W only

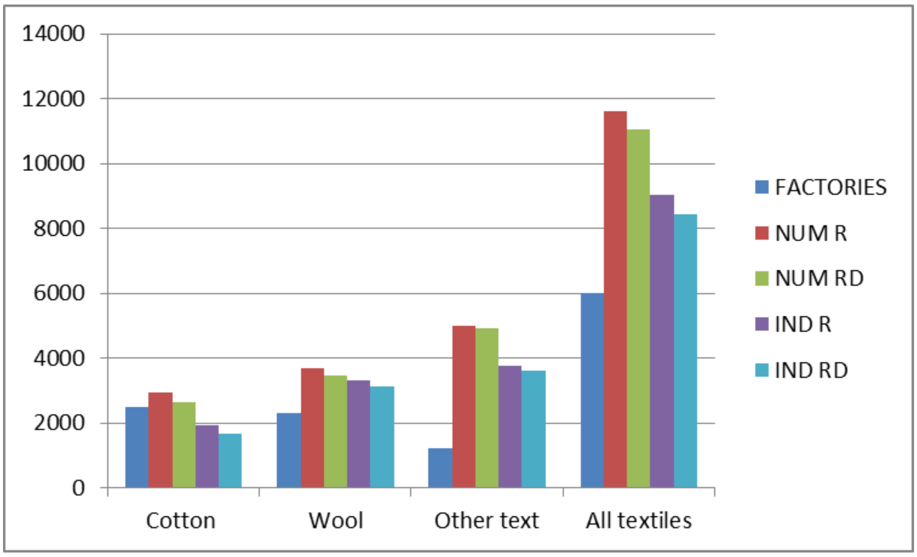


Given that the Factory Returns at this date are known to give partial coverage, the match with estimates here would be expected to produce under-estimates by FR compared to BBCE – which is generally the case. For workforce numbers, however, these comparisons are close enough to confirm generally good accordance between FR and BBCE by number of firms estimated by the methods used here, especially in the light of the close comparisons already shown with the Booth data.

4.3. Firm size comparison with Factory Returns

The Factory Returns for 1878 and 1885 may be more useful for *total* firm numbers than for workforce as they are one of the few sources on number of firms in larger enterprises and workshops. The comparisons in Figure 3.5. below for manufacturing alone shows that NUM usually over-estimates, and IND is closer in numbers. Inclusion of OA has only small effect in these sectors, though largest for other textiles. This accords with the other conclusions for the textile sectors: for hosiery and hair the FR comparisons are of limited utility. As with workforce numbers, the comparisons of firm numbers look good enough to confirm the accordance with what the Factory Returns can show. The IND R being used in this paper gives one of the closer fits.

Figure 3.5. Estimates of firm numbers compared to 1878 Factory Returns E&W only



Mean firm size in the FR can be compared from the sector breakouts in manufacturing. The initial difficulty of comparison is, however, what size to assume the FR fully covered in each sector. The figures show how the FR relates to the BBCE estimates (IND Rounded) assuming different size thresholds for FR inclusion. A factory of ≥10 employees is needed to match the FR figures for cotton, but between 10 and 50 employees for wool, ≥50 for other textiles, ≥10 for all textiles, but about ≥250 for apparel, and leather.

Figure 3.6.1. Mean firm-size estimates compared to 1878 Factory Returns for different size thresholds: E&W

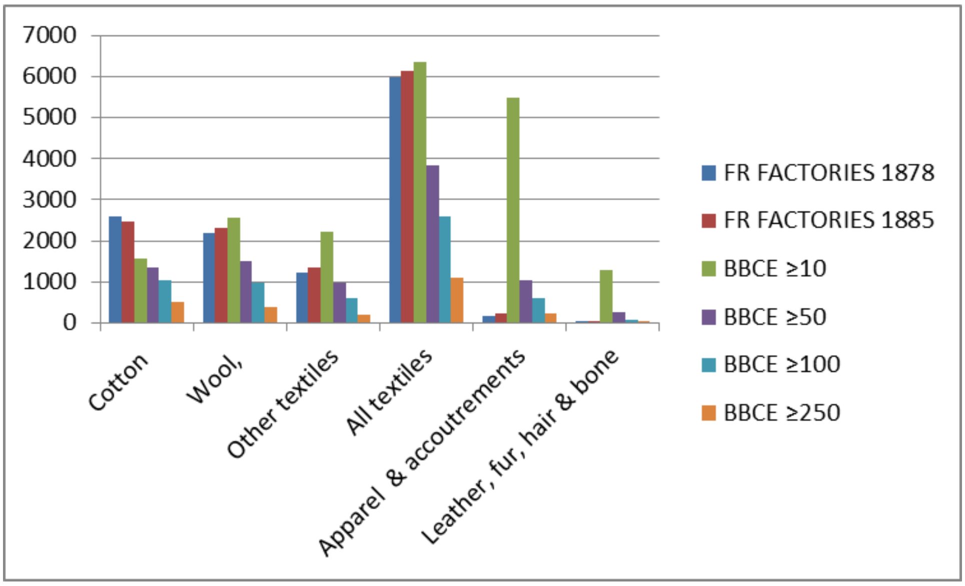


Figure 3.6 shows mean firm size comparisons using different thresholds. There is little difference for comparisons with the FR 1878 or 1885, though for all textiles the estimates are closer to the 1885 FR for all textiles. The match for *all textiles* with a factory size of ≥10 fits best overall, in both E&W and Scotland, and overcomes the issue of how wool, cotton and other textiles were differentiated in practice in both the FR and in the census. However, for hosiery (apparel) and hair (leather, fur etc.) there is no good fit except for the largest firm sizes, because the FR was surveying only a sub-sector within the industry category used here, especially in E&W.

Figure 3.6.2. Mean firm-size estimates compared to 1878 Factory Returns for different size thresholds: Scotland

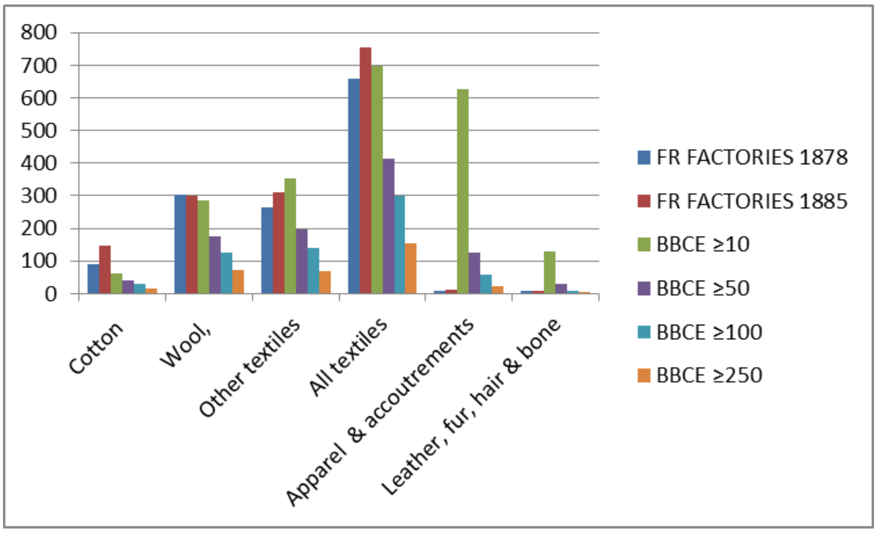
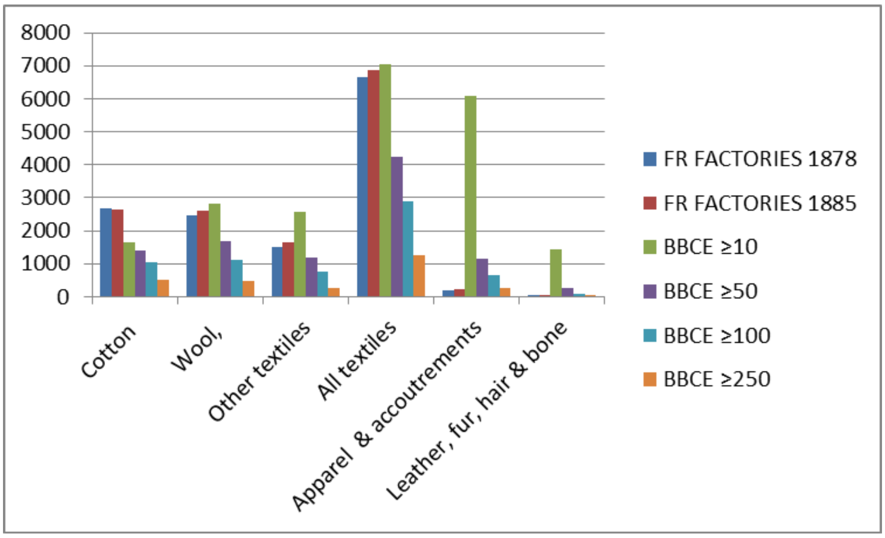
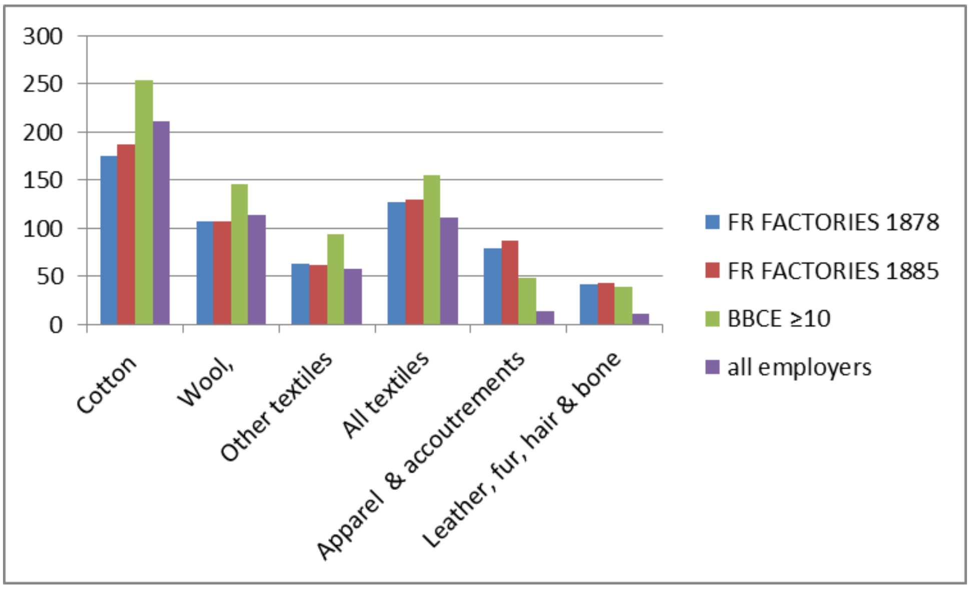


Figure 3.6.3. Mean firm-size estimates compared to 1878 Factory Returns for different size thresholds: GB



The analysis of *mean firm size* has to make assumptions about what the Factory Returns were using as a definition of factories and workshops. It appears that ≥10 is appropriate in general for the FR, and indeed there are firms down to this size which can be identified where the number of returns is only one, as for some Scottish sectors where the FR returns are very small. Using a threshold of 10 for a factory/workshop, as appears to be done for FR, seems a useful mid-point for analysis, but more varied thresholds could be used (see Figure 3.7).

Figure 3.7.1. Mean firm-size estimates compared to 1878 Factory Returns with ≥10 threshold: E&W



The results of the calculations using BBCE with ≥10 employees in these Figures show that only all textiles is a useful comparison because of the incomplete coverage of the FR in other sectors. For all textiles the matches are close between FR and the BBCE estimates, in both E&W and Scotland, but operate in different directions, with E&W slightly overestimated against the FR, and Scotland slightly underestimated. If we take the FR as correct, this probably arises either from (i) small differences in the way in which the BBCE estimates are constructed in E&W and Scotland (Smith et al., 2019, 2021), or (ii) differing definitions between what constituted a 'factory' in the FR, or (iii) how census respondents viewed the census question between E&W and Scotland - at the enterprise or the plant/factory level. Hannah's large firm analysis suggests they mainly referred to their specific business under their management, which could include several plants. However, for GB as a whole, the BBCE data with a ≥10 employee threshold slightly overestimate both the 1878 and 1885 FR estimates, which must mean that the census estimates are more inclusive than FR at this size threshold.[13] Overall BBCE can be taken as good estimates for all textiles in Scotland, Wales and England.

Figure 3.7.2. Mean firm-size estimates compared to 1878 Factory Returns with ≥10 threshold: Scotland

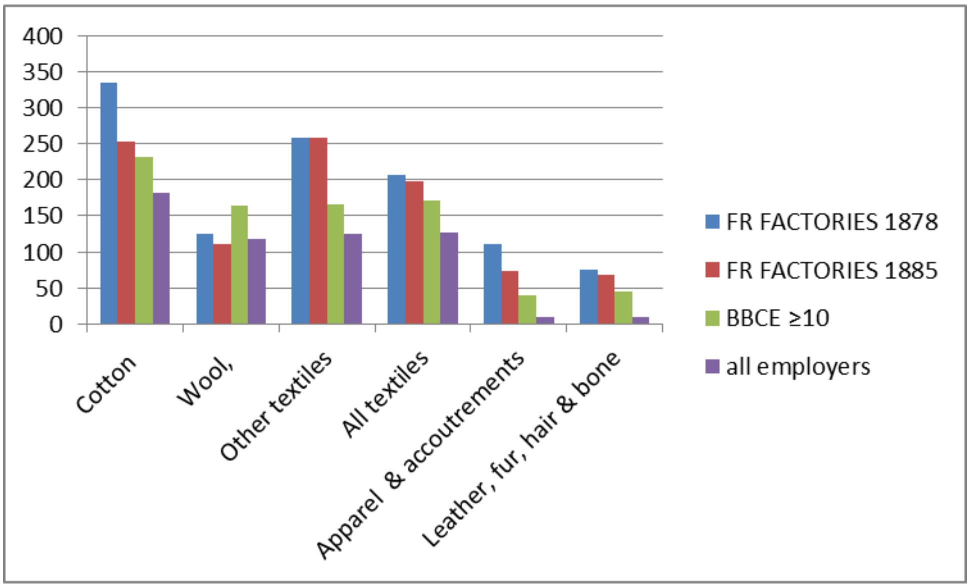
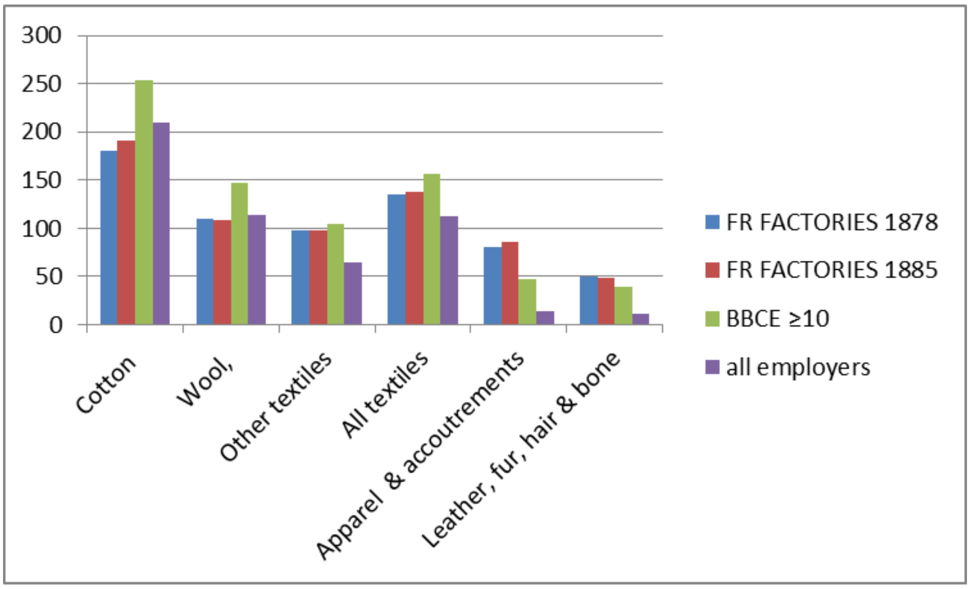


Figure 3.7.3. Mean firm-size estimates compared to 1878 Factory Returns with ≥10 threshold: GB



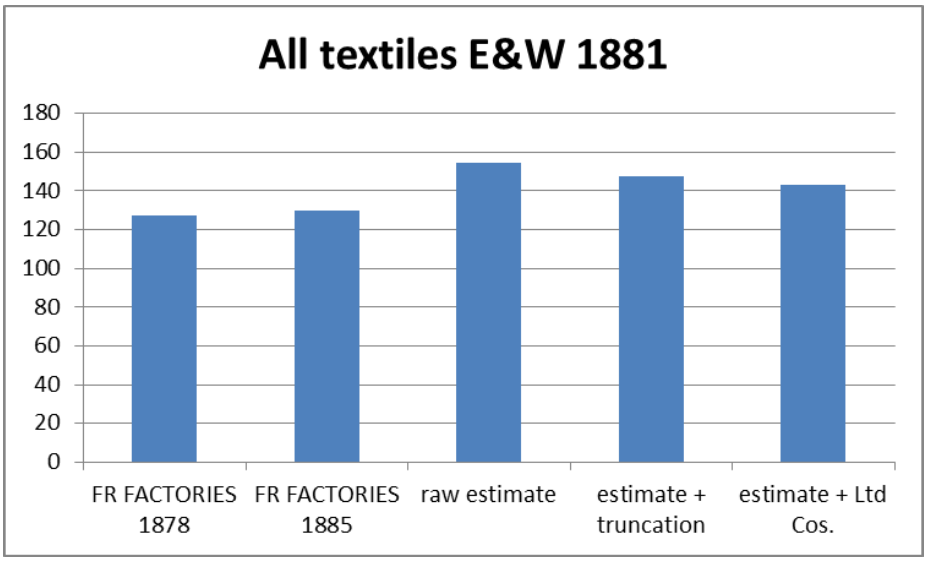
[13] It is also possible that 1878 and 1885 were depressed years with many factories stopped or on reduced workers and not counted by FR, while 1881 was more prosperous: as suggested by Jenkins (1978) evaluation of the FR. Also from 1875 onwards the definition of a factory changed from employing 50 or more to using power. So some BBCE workshops with more than 10 employees, and all hand workers, would be excluded by FR in later years.

However, non-response and truncation will affect these estimates if it differs by firm size. The more firms there are the lower the sizes the lower the mean. For example, if the BBCE had either approximately 10-15% more firms of over 300 employees, based on 2-3% more from those under 300 employees, this would lower the mean. The Bennett and Hannah (2021) calculations for all firms under 300 employees for E&W for all textiles give 1,190 firms of ≥250, and 6,629 firms of 10-249. Hence, for the larger firms truncations may raise numbers by 119-178, and for smaller firms by 133-200. Taking the midpoints as 149, and 167, a re-calculation increases the E&W numbers to 6,655 compared to 6,339. This lowers the mean from 154.6 to 147.2. This is closer to the FR of 127.3 for 1878, and 129.8 for 1885. Hence, truncation alone does not account for most of the differences.

Another factor is that the corporate sector is not fully included in BBCE. However, all the workers for companies are included, and the number of corporations is fairly small. In all manufacturing operating mainly in GB there were 398 manufacturing companies in the 1881 DoD data (van Lieshout and Bennett, 2019b); general manufactures containing textiles had 207, of which textiles was the majority. Hannah found 143 with ≥1,000 employees. If 90% of all general manufacturing was textiles companies, this would be 186. Assuming all these had over 10 employees, they can all be added to the business numbers. This gives a mean size of 143, which is closer to the FR estimates, but still indicates that the FR probably did not cover all workers that were relevant in the size category of 10 and over (e.g. some outworkers were not counted, and/or some smaller workshops were missed and/or the FR size category was defined as larger than 10 in some cases for what constituted a factory); alternatively, the BBCE has too many people assigned to worker categories who should be OA or employers. Bennett et al. (2021) show that using trade directories, the categories of textile workers such as ‘weaver’ or ‘spinner’ are among the most difficult occupational descriptors to assign between worker, own account and employer because the census made no effective effort to demark these different statuses.

Figure 3.8 shows that overall the estimation process confirms the methodology, at least for all textiles. It appears to fit the FR well, and the differences can be mainly explained by the effects on the estimates of employer numbers from truncations and exclusion of most corporates from BBCE. The remaining discrepancy between FR and the estimates appear to derive from slightly different definitions between them, presumably mainly related to the size or characteristics of what constituted a ‘factory’ to be surveyed by the FR compared to the enterprise as viewed through the eyes of census respondents. There will also be any errors in the supplementation process used in BBCE which may result in over-estimating employers vs. OA and workers, and hence affecting mean firm size calculations. Given all these constraints the conformity of the BBCE estimates and FR is remarkably close, and hence very reassuring that the estimation method is reasonably accurate, at least for all textiles.

Figure 3.8. Mean firm size 1881 in all textiles compared to FR, with adjustments: E&W



4.4. Firm-size comparison with the Hannah large manufacturing firm data

Comparisons with the Hannah ‘truth’ data for manufacturing firms of ≥1000 employees can be used to test sensitivity for the largest size class. Though constrained to manufacturing this remains a valuable assessment as this was a major sector, and the one with the largest number of the largest firms and the largest number of corporates – hence it is most sensitive to different estimation assumptions and to truncation and non-response deficiencies. Hannah takes a wider definition of manufacturing than the BBCE EA17 sector 4: he also includes some industries that in BBCE are under maker-dealing (EA17 sector 5) and agricultural processing (EA17 sector 10). He also takes a wider definition than some previous analysis (e.g. Feinstein) by including the utility industry gas production and its by products, which is already in BBCE EA17 sector 4 (manufactures). To achieve comparability the Hannah manufacturing definitions were applied to the BBCE by extracting the wider manufacturing categories using Hannah's definitions.

Table 4.5 shows that overall the re-scaling method underestimates the number of firms by 38%, identical to Hannah's data estimate of the BBCE non-response rate (see Section 2.2), and close to the 44% estimate of the BBCE response rate from supplemented data (Section 2.1). The re-scaling is almost exact or very close in some sectors, such as *all* textiles. Some other sectors are closer when aggregations are made, especially across machinery, engineering and metals. The Hannah manufacturing sub-sector classification is not easily matched with BBCE in some cases. This is particularly true of railway engineering, which is mainly included in transport in BBCE, but Hannah has attempted to separate out railway manufacturing workshops using secondary data. Similarly, Hannah has attempted to separate textiles in a different way. Most textiles differences derive from how cotton and woollen workers are described in the census, and how distinct or overlapping the two sectors of manufacturing firms were. Hannah used secondary data to make fine distinctions between textile subsectors which the census did not achieve. He also treated textiles outside cotton and wool in a slightly different way; in BBCE these are not as well distinguished as Hannah was able to achieve. Because of these and other differences, BBCE has 20 large firms which are not assigned to Hannah sectors because of uncertain information, or mixed or complex structures that Hannah allocated to a specific sector. These are shown as 'other manufactures' in the Table). The exact firms could be allocated out individually, but the purpose of the comparisons is to test the robustness of BBCE sector coding because this has to be applied across all firm-size classes where little secondary data exist.

The sectors with some of the largest discrepancies are mainly where corporates are significant and not fully included in BBCE census responses: Iron & steel, Tool & weapons, Shipbuilding, and Chemicals & utilities. Many firms in these sectors were incorporated by 1881, but in most other sectors the incorporated element was smaller and for these the fit with BBCE is usually better. There was a moderate level of incorporation in textiles, but for all textiles BBCE using statistical supplementation is as good as Hannah's truthing. Elsewhere BBCE supplementation methods do not fully adjust for the small number of firms in some of the largest size categories which are disproportionately where truncations and effects of corporates are concentrated.

Table 4.5. Firms of ≥1000 in BBCE estimates and Hannah database GB: Hannah fine categories aggregated to final Hannah codes; differences over 40% highlighted

Hannah codes	Manufacturing sector classification	E&W BBCE N	Scot BBCE N	GB BBCE N	Hannah GB N	Deficit N BBCE - Hannah	Diff N %	Hannah: mean size
1	Cotton	58	2	60	84	-24	-28.5	1750
2	Wool, worsted, blankets	48	4	52	30	22	92.6	1334
3,19	Other textiles & upholstery, furniture	19	8	27	27	0	58.8	1383
	All textiles	**125**	**14**	**139**	**139**	0	8.6	1658
4-6	Apparel , accoutrements, shoes	**10**	0	10	22	-12	-54.5	2320
7	Iron & steel, bolts etc.	**6**	3	9	84	-75	-89.3	3157
8	Machinery	24	6	30	15	15	100	1840
9	Textile machinery	1	0	1	12	-11	-91.7	2300
10-12, 15,22	Tool & weapons	4	0	4	17	-13	-76.5	1845
13	Shipbuilding	8	4	12	30	-18	-60	2585
14	Railway engineering	0	0	0	27	-27	-100	2342
16	Bricks, earthenware & glass	6	0	6	9	-3	-33.3	2168
17-8	Chemicals & utilities	3	2	5	34	-29	-85.3	1744
21	Printing, publishing & paper	4	0	4	11	-7	-63.6	1541
23-4	Food, and agric. processing	9	2	11	12	-1	-8.3	1430
25	Other manufacturers	20	0	20	0	20		-
	TOTAL	**219**	**31**	**250**	**444**	**-24**	**-37.6**	**19524**

A plot of the BBCE data for EA17 manufacturing sectors 4, 5 and 10 vs Hannah's data for the firms of 1,000 employees and over shows a close relationship in terms of general pattern and trend lines (Figure 3.9). All are based on the data at size data-points: N of firms of a particular size. The important differences are: BBCE has fewer large firms over 5,000 employees, and the Hannah data contain all Ltd. companies as well as non-corporates. The BBCE has census respondents re-scaled to match the number of employers estimated by EMPSTATUS_IND. For both data sources the effects of bunching at 1000s and 100s is evident and similar, though bunching is more pronounced for BBCE. The regression is for frequency at a given size data-point (e.g. 10 firms with 1,500 employees). As a result the N of firms in BBCE is reduced to 90 size data-points, and to 75 in Hannah's data.

A similar closeness of the estimates is shown in Figure 3.10 for log firm size. In both figs the small difference in regression coefficients is accounted for by the longer upward tail for

Hannah's data, which has 7 firms larger than the largest firm (of 8,000 employees) in the BBCE manufacturing data, 13 firms larger than the next largest in BBCE (of 7,000 employees), and 19 over the next largest of 6,000 employees. This difference is less crucial for the log comparison. The comparison exaggerates the BBCE shortfall since the Royal Dockyards (the largest firm), London & NW Railway workshops (the 6th largest firm), Great Western Railway workshops (12th largest), and Royal Ordnance Factories (15th largest) were not be included in BBCE (the Dockyards and Ordnance were treated by BBCE as state industries, and railway workshops were not disaggregated from the rest of transport).

Figure 3.9. Plot of Hannah and *manufacturing* BBCE firms of 1,000 employees and over (GB) vs firm size; with simple linear regressions

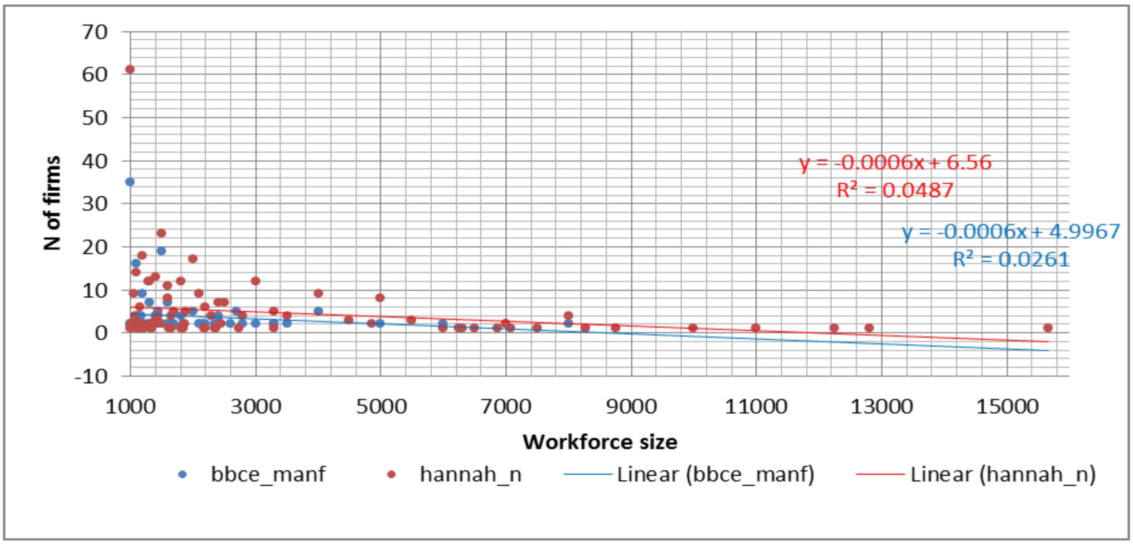
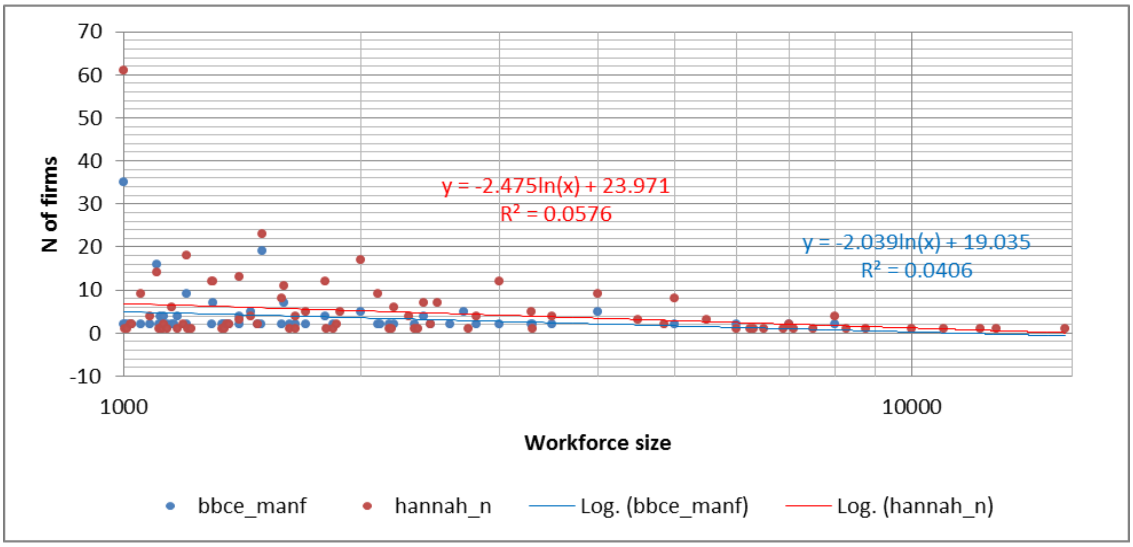
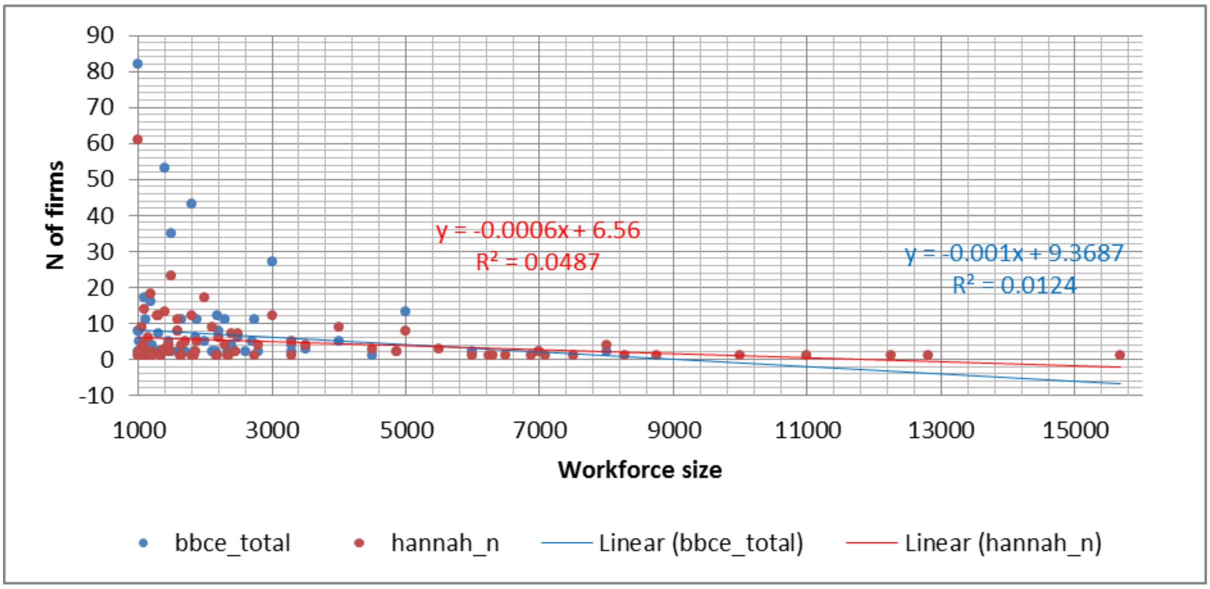


Figure 3.10. Plot of Hannah vs BBCE *manufacturing* firms of 1,000 employees and over (GB) vs log firm size; with log linear regressions



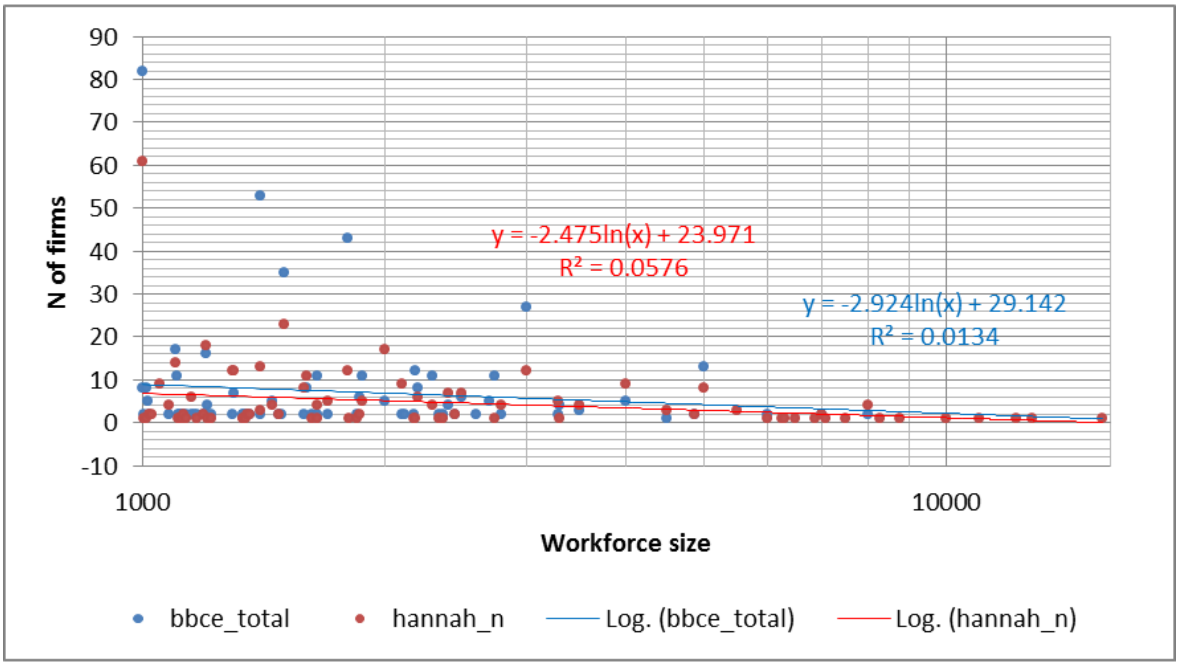
A plot of the BBCE data for *all* EA17 sectors vs Hannah's data for manufacturers of 1,000 employees and over shows similar relationships (Figure 3.11), except that the regression coefficient for BBCE is now larger. Bunching at 1000s and 100s is again evident, and bunching is relatively more pronounced for the BBCE all firms' data than manufacturing. Results are also similar for logged firm size for all BBCE firms (Figure 3.12). Note the upper tail does not change between all and manufacturing BBCE firms since all the largest firms in BBCE are manufacturers.

Figure 3.11. Plot of Hannah manufacturing and BBCE *all* firms of 1,000 employees and over (GB) vs firm size; with simple linear regressions



These results are somewhat encouraging. Since the Hannah definition of manufacturing is wide, and the exact attribution of sectors in BBCE EA17 is fuzzy, given sector coding issues, the similarity of the relationship with the Hannah large firms' data for BBCE manufacturing sectors 4, 5, and 10, and all BBCE sectors indicates that all large firms can to some extent be inferred, or even surrogated, by Hannah's manufacturing firms (at least until equivalent 'truth' data is available for large non-manufacturing firms). It also suggests the need to scrutinise the sector coding of the BBCE data, particularly at the upper tail.

Figure 3.12. Plot of Hannah manufacturing vs BBCE *all* firms of 1,000 employees and over (GB) vs **log firm size; with log linear regressions**



Various estimates of the regression coefficients are compared in Table 4.6. This shows re-scaled BBCE and Hannah estimates without and with clustered estimation to allow for bunching at size data-points. The BBCE data are estimated first for manufacturing firms alone, and then for all firms. The all-firm estimates allows comparison of manufacturing and other firms, but in any case manufacturing forms a large proportion of all the largest BBCE firms. The table also compares the Hannah data for the effect of excluding the Royal Dockyards (RD) which are something of an outlier, and arguably not part of the same economic system.

The upper section of the table is for un-logged firm size; the second section is for logged data which takes account of the non-linear form of the firm-size distribution; and the third section is for the Hannah data after exclusion of Ltd companies at each size data-point. The unlogged and logged data have close similarities of estimates for both re-scaled BBCE and Hannah. Clustering has little effect on the coefficient estimates but marginally increases significance levels by reducing spurious variance. Excluding the Royal Dockyards from Hannah's data has little effect. The intercepts and slope coefficients remain similar for all estimates in the same sections of the table. Exclusion of Ltd companies at each size data-point changes the coefficients, but they remain similar to estimates with Ltd companies included.

Table 4.6. Regression estimates of BBCE and Hannah data vs firm size; note data are not firms but size points (the number of observations at a specific size).

	Coeff	prob	Cons	prob	R²	F	prob
BBCE *manuf* firms vs size	-0.0006	0.226	4.99	0.000	0.026	1.50	0.226
BBCE *manuf* firms vs size *clustered*	-0.0006	0.081	4.99	0.000	0.026	3.15	0.081
BBCE *all* firms vs size	-0.0010	0.349	9.37	0.001	0.012	0.35	0.348
BBCE *all* firms vs size *clustered*	-0.0010	0.160	9.37	0.001	0.012	2.02	0.159
LH vs size	-0.0006	0.048	6.56	0.000	0.048	4.04	0.048
LH vs size *clustered*	-0.0006	0.006	6.56	0.000	0.048	8.04	0.058
LH vs size ex RD	-0.0006	0.050	6.71	0.000	0.048	3.97	0.049
LH vs size ex RD *clustered*	-0.0006	0.007	6.71	0.000	0.048	7.71	0.007
BBCE *manuf* firms vs log size	-2.04	0.130	19.03	0.060	0.040	2.37	0.129
BBCE *manuf* firms vs log size *clustered*	-2.04	0.116	19.03	0.066	0.041	2.54	0.116
BBCE *all* firms vs log size	-2.92	0.330	29.14	0.197	0.013	0.96	0.330
BBCE *all* firms vs log size *clustered*	-2.92	0.286	29.14	0.179	0.013	1.16	0.287
LH vs log size	-2.47	0.031	23.97	0.008	0.058	4.83	0.031
LH vs log size *clustered*	-2.47	0.032	23.97	0.015	0.058	4.75	0.032
LH vs log size ex RD	-2.52	0.036	24.28	0.010	0.055	4.54	0.036
LH vs log size ex RD clustered	-2.52	0.039	24.28	0.018	0.055	4.43	0.038
LH vs size ex Ltd	-0.0004	0.289	4.11	0.001	0.019	1.14	0.289
LH vs size ex Ltd *clustered*	-0.0004	0.114	4.11	0.002	0.019	2.57	0.114
LH vs log size ex Ltd	-1.52	0.207	14.69	0.109	0.027	1.63	0.206
LH vs log size ex Ltd *clustered*	-1.52	0.210	14.69	0.136	0.027	1.61	0.201

Five important conclusions derive from these estimates, which are also clear from Figures 3.9-3.12. First, none of the estimates has high significance and can only be taken as a guide for comparison between the two data sets. They are both affected by the specific patterns of bunching, and whether zero observations at a firm size data point are set to missing, as here. There are also sector coding issues. They must be taken as a preliminary test; further tests of the two data distributions are being developed. Second, the Hannah data confirm the difficulties of the re-scaling method in coping with the small number of large firms at many of the data points, with no BBCE firms over 8,000 employees, whilst Hannah found seven. This indicates different approaches may be needed beyond the BBCE maximum size point.

The third conclusion is that, apart from the firms over the BBCE maximum size available, the re-scaling method achieves good matching between BBCE IND and the ‘truth’ data. All regression coefficients are similar, certainly within the bounds of the data uncertainties involved. Wide standard errors, and low or non-significance, in most cases indicate that most alternative estimates are not strongly distinguished. Clustered estimates of BBCE manufacturing firms are marginally better than non-clustered for unlogged data, but slightly lower for logged, since logging reduces the variance for the upper sizes. BBCE estimates for manufacturing firms are poorer than for all firms, but the differences are small and BBCE estimates are rarely significant. For manufacturing, these indicate that the re-scaled BBCE data is something of an indicator of the whole firm-size distribution for the largest firms that is likely to have existed in 1881. But greater sector controls and checks on coding are needed as it is likely that the coefficients vary by sector.

Fourth, the effect of the exclusion of most corporates from the BBCE data generally does not undermine the capacity to estimate the firm-size distribution. The Hannah coefficients with corporates excluded are much closer to the BBCE regression coefficients, which is reassuring (although IND re-scaling is only imperfectly compensating for large-sized corporates, as expected for the very largest firms). Nevertheless IND seems to give a useful approximation. This is possible because, as previous analysis of both the record-linked and Hannah ‘truth’ data indicate, non-responses were largely random by size and enough non-corporates replied in the relevant size class for re-scaling to cover most corporate non-respondents (except for the very largest sizes). Indeed, the Hannah data for manufacturing businesses of 1,000 or more employees for 1881 include some incorporated companies with employee numbers recorded in the census responses, for both private unlisted and listed, although their response rate was significantly lower than non-corporates (Bennett and Hannah, 2021). This was inevitable given that the design of the census question did not guide directors or others who might have made responses for workforces of Limited companies.

Fifth, as noted when comparing Figures 3.9-3.10 with 3.11-3.12, the differences are quite limited between the BBCE estimates of manufacturing firms alone, and all firms. Hence, although *all* firm estimation needs a check on sector coding and controls for sector differences in later developments, the effect of greater sector detail is unlikely to give major improvements of the re-scaling model used here or the inferences about the firm-size distribution as a whole. This is not surprising. The Hannah definition of manufacturing is wide, and the exact attribution of sectors in BBCE EA17 is fuzzy given data coding issues. Hence, the similarity of the relationships between *all* BBCE firms and manufacturing, and between the BBCE estimates and the Hannah large manufacturing firms, is to be expected. Until equivalent 'truth' data are available for large non-manufacturing firms, the expectations of the BBCE data re-scaling for *all* large firms might be inferred to be similar to Hannah's manufacturing firms.

These simple OLS regression estimates are an experiment. They are perhaps inappropriate for these data as the upper values are truncated for BBCE. Both BBCE and Hannah also have a maximum feasible value (for the largest firm). Table 4.7 gives equivalent truncated regression estimates using the restricted part of the population. This makes the assumption that the data have a truncated normal distribution, which can be scaled upward so that the distribution integrates over the restricted range. This is estimated for both BBCE and Hannah with their maximum observed firm sizes. These estimates should be unbiased by allowing for the truncations. In this case the slope coefficients for BBCE and Hannah are almost identical, and the intercept coefficients are close, indicating the close match of the two data sets. However, the truncated regressions have almost no difference from the OLS regression shown in Table 4.6. This indicates that truncation effects alone have little effect on the estimation process. Arguably a Poisson estimate, probably with allowance for truncation would be preferable. However, the data are not really currently event rates and would have to be re-cast. Preliminary experiments with the Poisson were unstable.

Table 4.7. Truncated regression estimates for firms in BBCE and Hannah data vs firm size.

	Coeff	prob	Cons	prob	Wald Chi²	prob
BBCE *manuf* firms vs size	-0.0006	0.130	4.28	0.000	2.29	0.130
LH vs size	-0.0006	0.042	6.56	0.000	4.15	0.041

Clearly the main drawback of the re-scaled estimates is that corporates, having economies of scale and other advantages, accounted for all but one of the very largest firms over 8,000 employees, and almost all the largest firms over 5,000 employees. Hence, although the effects of corporates can probably be generally catered for by the re-scaled IND estimates for smaller firms, for the largest a different approach would be preferable.

One possibility is to replace re-scaled BBCE with Hannah's 'truth' data after a specified size point cut-off. This could be for all firms over 1,000 employees; or could be just for the upper tail where the BBCE census responses are absent or infrequent. Inspection of the data suggests this could be at various points between 1,500 and 5,000, but to avoid the effect of bunching in either BBCE or Hannah it is preferable to choose a size where this effect is minimised. This could be tested by trial and error or other estimation methods, but as an experiment the use of 2,400 shows the potential of this approach. Table 4.8 compares the regression coefficients for BBCE IND re-scaled and Hannah's data for those firms with 1,000 – 2,400 employees. Although the N of size points is reduced to 59, this is still large enough to indicate likely outcomes. In this case the slope coefficients are almost identical - indicating that re-scaled BBCE and Hannah have near-equivalent information on the size distribution for this part of the firm-size range. The differences are almost entirely in the intercept coefficients, although even these are small. In both the un-logged and logged estimates this indicates that re-scaled BBCE compared to Hannah's data is very close.

A conclusion from this experiment is that, for firms of 1,000 – 2,400 employees, a small re-weighting could be used to align the coefficients of re-scaled BBCE with 'truth' data. However, the differences are small compared to the standard errors of the estimates from both data sets, so that such an approach may be unjustified.

Table 4.8. Regression estimates for firms at or below 2,400 employees in BBCE and Hannah data vs firm size.

	Coeff	prob	Cons	prob	R²	F	prob
BBCE *manuf* firms vs size	-0.0025	0.247	7.83	0.047	0.029	1.38	0.247
BBCE *manuf* firms vs size *clustered*	-0.0025	0.247	7.83	0.047	0.029	1.38	0.247
LH vs size	-0.0024	0.451	9.41	0.059	0.011	0.58	0.451
LH vs size *clustered*	-0.0024	0.483	9.41	0.130	0.011	0.50	0.482
BBCE manuf firms vs log size	-4.148	0.229	34.23	0.172	0.033	1.49	0.228
BBCE manuf firms vs log size clustered	-4.148	0.278	34.23	0.229	0.033	1.20	0.278
LH vs log size	-4.055	0.415	35.34	0.330	0.133	0.68	0.418
LH vs log size *clustered*	-4.055	0.502	35.34	0.432	0.133	0.46	0.502

To test sensitivity a series of other experiments were made to compare the effect of size ranges for firm-size cut-offs above and below 2,400. These have results consistent with using 2,400: almost identical slope coefficients, but with intercept coefficients more unstable depending on the data points included - generally but not always slightly higher for the Hannah data at lower firm-size cut-offs, and slightly lower with higher cut-offs, but all with high standard errors. This indicates that the choice of firm-size point for switching between re-scaled BBCE IND and 'truth' data is somewhat arbitrary up to about 3,000 as a cut-off. Given that the results are almost identical in each size range, re-weighting is probably unjustified, since the low significance of all the estimates indicates that BBCE and 'truth' data as essentially similar distributions as far as can be tested in this preliminary way.

An alternative approach is to recognise that most firm-size literature indicates a log-log distribution of firms by size. Indeed this has been estimated from the original BBCE data for 1881 (and other years 1851-71) by Montebruno et al. (2019b). Table 4.9 gives log-log estimates for BBCE and Hannah. These estimates are again very close between the two data sets, especially using the truncated regression estimator, though as in previous tables all the estimates are not statistically significant, or only marginally so.

Table 4.9. Regression estimates for log BBCE and log Hannah data vs log firm size.

	Coeff	prob	Cons	prob	R²	Wald Chi²	F	prob
Log BBCE manuf firms vs log size	-0.242	0.141	2.826	0.023	0.038	-	2.23	0.038
Log BBCE manuf firms vs log size *truncated regression*	-0.242	0.129	2.826	0.018	-	2.31	-	0.128
Log LH vs log size	-0.372	0.014	3.810	0.001	0.074	-	6.36	0.014
Log LH vs log size *truncated regression*	-0.292	0.187	3.908	0.029	-	1.74	-	0.187

Efforts to estimate the log-log model for firm-sizes up to 2,400 are frustrated by the large number of size-points with only one firm, which makes the estimation unstable. This could be overcome by grouping the data into size categories, but this loses detail of the distribution. Alternative estimators are being explored to take this analysis further.

The different approaches to estimation indicate that, overall, the two data sets give equivalent coverage of the upper part of the firm-size distribution. However, further estimation methods and experiments are required to decide on the most appropriate way to make a switch between re-scaled BBCE and 'truth' data to give proper estimates of the workforces of the largest firms. Nevertheless, at this stage, it seems feasible to suggest that the two data sources can be combined to give a valid coverage of the total firm-size distribution, using Hannah's 'truth' data above an appropriate size-point, without further adjustment.

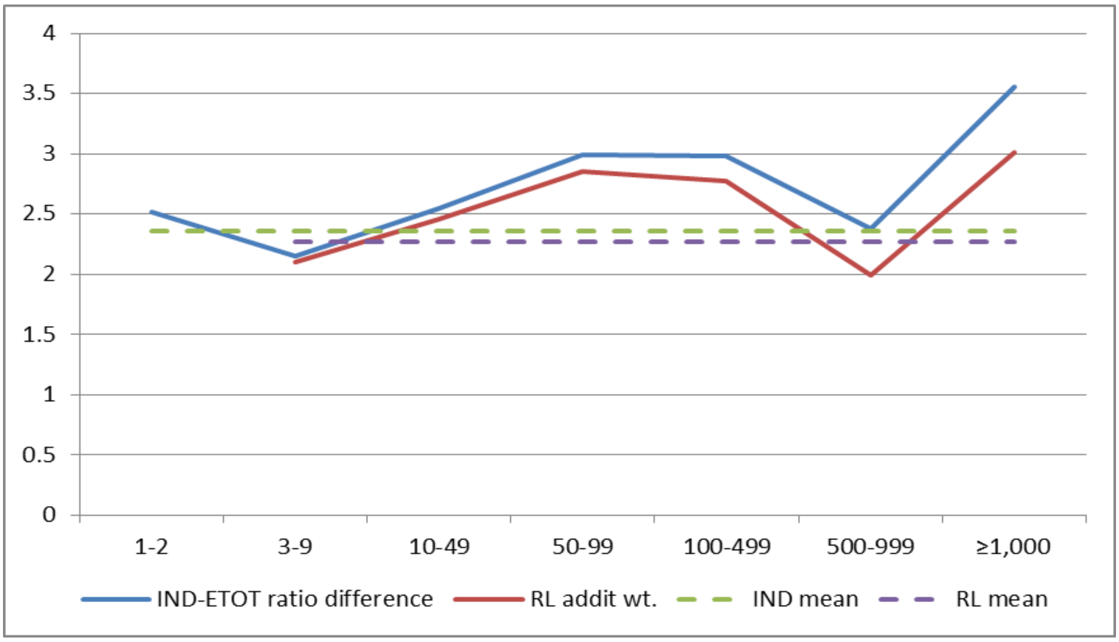
Part 5. Re-scaling / Weighting: Next steps

5.1. Strength of the re-scaled IND estimates compared to RL

The re-scaled IND data are the estimates generated by the method developed in section 3 of this paper. A check can be made by comparing with what would be generated if the results of the record-linked (RL) estimates of non-response are used for re-scaling. This allows a check of the re-scaling method against the only pervious analysis to assess non-response for the main firm-size distribution (from 10 employees upwards). The implied weights from IND and RL are shown in Figure 5.1. The size classes used here are those used in earlier tables to summarise the distribution; the RL estimates shown in Table 1 need slight adjustments to accord with these classes, but these have only minor effects on calculations.

The blue line in the Figure is the ratio between the original ETOT and the re-scaled BBCE IND data, as developed in section 3 (Table 3.3) above. The re-scaling weight ranges from 2 for the smallest firms to 3.5 for the largest firms. The mean weight is 2.36. The mean is heavily influenced by the large number of very small firms (under 10 employees). Note the very smallest size class (1-2 employees) is subject to more uncertainties because of fuzziness in the data about inclusion of partners, family members, and own account proprietors. It is shown here for completeness and has not been estimated for the RL data.

Figure 5.1. GB comparison of up-scaling using IND and RL estimates, both for 1881.



The red line in Figure 5.1 is the reduction in the ratio using re-scaling method estimates with IND if weights derived from the record-linked estimates of non-response are used first (based on Table 1). These IND weights are reduced, but by only a small amount. The RL red line is lower than the re-scaled IND estimates by the weighting for occasional non-responses (people who respond in some censuses but not in others). The RL re-scaling is insufficient compared to the IND method. The comparison demonstrates that non-response compensation, as estimated by record-linkage, provides only a small part of the adjustments needed. There is no way to estimate the additional re-scaling needed solely from the RL data - RL alone is an inadequate adjustment method. In addition the re-scaled IND has the advantage that it includes both non-responses and truncations, but most importantly it also ensures the data fit to aggregate estimates of employer numbers rather than relying solely on responses in the original CEBs. It is thus a more complete way of making estimates. It is a more practical and effective way of attempting to get a form of 'truth' without having the actual records of firms available (as was possible for Hannah's large firms).

Given that all the estimation methods contain various assumptions and approximations, only the general patterns from these comparisons can be treated as relatively robust, and small differences should be ignored. What can be concluded is that efforts to estimate non-response and truncation at the micro level through RL comparisons with CEBs are likely to under-estimate the population of firms, with increasing under-estimation with firm size for the medium and larger size classes. The method developed here, of re-scaling the BBCE supplemented estimates EMPSTATUS_IND, provides estimates that are likely to be closer to the actual firm size distribution.

5.2. What does the analysis of Hannah's 'truth' data add?

The analyses of Hannah's 'truth' data indicates that for the large manufacturing firms the re-scaled BBCE IND generally works well to estimate actual firm numbers by firm size – including many corporates. Its major limitation is the inability to re-scale for the largest firms where no census respondents replied (in 1881, all firms over 8,000 employees, and almost all the largest firms over about 5,000 employees).

For 1881 the poorer performance of the re-scaled IND estimates for the largest firms is not necessarily important, since there are 'truth' data available. The experiment with using a 2,400 size-point to switch between them two suggests that the main limitation of the re-scaling method can be overcome by substituting the 'truth' for the re-scaled data for the largest sizes. However, the existing 'truth' data cover only manufacturing firms (on Hannah's wide definition). Although manufacturing incudes most large firms in 1881, use of wider 'truthing' data requires the sectoral coverage to be extended.

This suggests that the best way to overcome similar deficiencies in earlier census years would be to develop similar ‘truth’ data for 1851, 1861 and 1871, extend the ‘truthing’ to large non-manufacturing firms, and identify equivalent size-points to switch between re-scaled BBCE and ‘truth’ data. This is a considerable challenge as secondary data are often less available for the earlier years.

On the other hand, the deficiencies of the IND estimates should not be exaggerated. In 1881 an important part of the BBCE shortfall for the largest firms derives from two firms BBCE excludes as state industries (the Royal Dockyards, the largest firm, and the Royal Ordnance Factories, the 15th largest), and railway workshops are included in transport in BBCE and not part of a manufacturing comparison with Hannah. Also in the earlier censuses the challenges for the IND method at the largest sizes is reduced: there are less corporates, the largest manufacturing factories (over 1,000) are smaller in number than in 1881, and many of the largest firms in 1851 will have relied on outworkers for scale, which BBCE and most researchers exclude from firm size estimates (since these were external contractors).

6. Conclusion

This paper has developed a methodology to estimate the firm-size distribution of Britain in 1881 using *data re-scaling* of the BBCE database to compensate for census non-response and various biases in digital data preparation. This has been applied to sector distributions based on BBCE EA17 sector codes. It prepares the ground for extending the methodology to earlier census years 1851-71, and making other comparisons.

The BBCE database provides estimates of number of firms for both employers and own-accounts proprietors using two methods of data supplementation: EMPSTATUS_NUM and EMPSTATUS_IND. These methods already re-scale firm *numbers* to compensate for non-response, inaccurate response between employments status categories (misallocation bias between employers, own-accounts, workers and inactive), and digital data truncations and miscoding. EMPSTATUS_IND is used here as the main starting point to estimate the *distribution of firms between different sizes* of workforce.

Previous assessments of the BBCE data for firm-size have demonstrated that it is superior to GRO tabulations in offering a more inclusive coverage of all employers who returned their workforce data. Indeed BBCE is significantly superior to GRO tabulations. Hence, re-scaling for non-response using GRO tables is inadequate, even where available. For *non-farmers*, BBCE generally has a much higher proportion of the firms of the larger size classes (by including *all* employees), and also a larger number of small firms than GRO tabulated (once data losses from the archival record are allowed for).[14] For *farm proprietors* the position is more complex, but BBCE again includes more large firms. However, neither BBCE nor GRO have responses that employers should have given, but failed to provide. The challenge for the paper is how data re-scaling can be used to estimate the number of all firms of workforce different sizes.

Similarly using the record-linkage results to re-scale by size is inadequate to infill the non-responses required because occasional non-respondents are relatively few. The key problem for estimation is that a large number of employers who never responded.

The re-scaling method uses the BBCE supplemented data for 1881 based statistical estimation of the characteristics of responding employers (compared to non-respondents) from later censuses. This gives total firm numbers by sector. Here, these have been compared with actual extracted census responses and the difference scaled up by a different proportion within each size class and sector. The BBCE IND data already include compensations for non-response, misallocation bias, and digital truncations and miscoding. The re-scaling method attempts to extrapolate these compensations across the firm-size distribution.

Comparisons of the re-scaled estimates have been made with contemporary data from three sources: Booth’s estimates of workforces, Factory Inspectorate Returns, and ‘truth’ data (developed by Hannah from BBCE and available secondary sources). There is good correspondence between the re-scaled estimates and these contemporary historical data on the number of firms and their workforces. However, the comparisons with Factory Returns and ‘truth’ data make clear that a major challenge for re-scaling is the very low response rate in the census from directors, managers or controllers of corporate firms, especially for the largest. Addressing the census question that gathered workforce data to these firms was not

[14] Using the 1851 GRO tables - the only year for which GRO made full tabulations.

considered by GRO, there was no specific direction or instruction for them to respond, and it is not surprising that many were non-respondents. Despite this deficiency of census design, the re-scaled BBCE data in most cases gives a good estimate of workforce size for the distribution of firms up to about 1,000 employees. This is possible because non-responses were random by size within each business legal form, and it appears that enough non-corporates replied in the relevant size class for re-scaling to cover the corporate non-respondents. The estimation works well even though, as found in previous analyses, the non-response rate was generally significantly lower for corporates than non-corporates. The non-corporates in effect compensate for corporates, and this gives effective estimates because IND data already include compensations for non-response, misallocation bias, and digital truncations and miscoding, that *embed the effects from low corporate non-response*. Hence, the re-scaling method includes these compensations across the firm-size distribution (at least for firms up to about 1,000 or to about 2,000 - 3,000 employees).

While re-scaled IND generally works well to estimate actual firm numbers by firm size, including many corporates, it has limitations for the largest firms. For many of these there were no census respondents; hence there are few or no extracted firms that can be re-scaled. In 1881 all firms over 8,000 employees had no census respondents. The substitution of the re-scaled data by Hannah's 'truth' data for the largest firms has been evaluated by experiment. This looks a feasible way to combine the general value of the IND re-scaling method for the vast majority of firms with 'truth' data for the firms over about 2,400 employees or similar size cut-off.

The same approach can be used to overcome deficiencies in earlier census years, through the cut-offs will probably differ. The BBCE database already gives supplemented IND data for these years. These can be used for re-scaling, which can then be combined with 'truth' data for 1851, 1861 and 1871 for large manufacturing as well as non-manufacturing firms. There is a considerable challenge for developing such 'truth' data as the secondary sources to achieve this are often less available than for the earlier years, though the number of firms over 1,000 employees is smaller. An additional challenge for the earlier censuses is archival loss. For 1881 it was possible to ignore archival loss because that census year had minimal losses. However, for earlier years archival loss is significant. Previous comparisons with GRO published tabulations indicate that for earlier censuses BBCE will need re-weighting by region or preferably Registration District or Sub-District; this will need to be combined with the re-scaling methodology. It is particularly challenging for 1851 where the two regions with the largest archival loss (London and the NW) have the largest concentrations of large and medium-sized firms.

Acknowledgments:

This data prepared for this research was originally supported by the ESRC under project grant ES/M010953: **Drivers of Entrepreneurship and Small Businesses**; pilot data extraction for 1881 by Leverhulme Trust grant RG66385: **The long-term evolution of Small and Medium-Sized Enterprises (SMEs)**; and Scottish data extraction by Isaac Newton Trust Grant, 18.40(g) **Business proprietor succession and firm size change 1851-1881**.

References

Bennett, R. J. and Hannah, L. (2021) British employer census returns in new digital records 1851-81; Consistency, non-response and truncation - what this means for analysis. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*; forthcoming

Bennett, R. J., Smith H. J., van Lieshout, C., and Newton, G. (2017) *Business sectors, occupations and aggregations of census data 1851-1911*. Working Paper 5. https://doi.org/10.17863/CAM.9874

Bennett, R., Montebruno, P., Smith, H., & Van Lieshout, C. (2018). *Reconstructing entrepreneur and business numbers for censuses 1851-81*. Working Paper 9. https://doi.org/10.17863/CAM.37738

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

Bennett, R., Montebruno, P., Smith, H., & Van Lieshout, C. (2019b). *Reconstructing entrepreneur and business numbers for censuses 1851-81: a tailored logit cut-off method*. Working Paper 9.2. https://doi.org/10.17863/CAM.43891

Bennett, R. J., Smith, van Lieshout, C., Montebruno, P. and Newton, G. (2020a), *The British Business Census of Entrepreneurs 1851-1911 (BBCE)*: *User Guide*. https://doi.org/10.17863/CAM.47126

Bennett, R. J., Smith, H., Montebruno, P. (2020b) ‘The Population of Non-corporate Business Proprietors in England and Wales 1891–1911’, *Business History*, 62, 1341-72. https://doi.org/10.1080/00076791.2018.1534959

Bennett, R.J., Smith, H., Montebruno, P. and van Lieshout, C. (2021) Changes in Victorian entrepreneurship in England and Wales 1851-1911: Methodology and business population estimates. *Business History*, https://doi.org/10.1080/00076791.2021.1894134

Booth, Charles (1886) Occupations of the People of the United Kingdom 1801–1881, *Journal of the Royal Statistical Society*, 49, 2, 314-435.

Clapham, J.H. (1938) *An Economic History of Modern Britain, Vol. 3: Machines and National Rivalries (1887–1914) with an Epilogue (1914–1929).* Cambridge: Cambridge University Press.

Hannah, L. and Bennett, R.J. (2021) Large-scale Victorian Manufacturers: Reconstructing the lost 1881 UK employer census, *Economic History Review*, forthcoming; preliminary version as LSE WP 330 https://www.lse.ac.uk/Economic-History/Working-Papers/Working-Papers-2021

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

Jaadla, H. (2019) *Weights to adjust for the number of missing women by Registration Sub-Districts in the I-CeM database, 1851–1911* [Dataset]. https://doi.org/10.17863/CAM.45290

Jenkins, D. E. (1978) The Factory Returns: 1850–1905, *Textile History*, 9, 1, 58-74.

Mills, D.R. (1999) Trouble with farms at the Census Office: an evaluation of farm statistics from the censuses of 1851-1881 in England and Wales, *Agricultural History Review*, 47(1), 58-77.

Montebruno, P., Bennett, R.J., van Lieshout, C., Smith, H. and Satchell, M. (2019a) Shifts in agrarian entrepreneurship in mid-Victorian England and Wales. *Agricultural History Review*, 67, 71-108.

Montebruno, P., Bennett, R.J., van Lieshout, C. and Smith, H. (2019b) A tale of two tails: Do Power Law and Lognormal models fit firm-size distributions in the Mid-Victorian era? *Physica A: Statistical Mechanics and its Applications*, 523, 858–875.

Montebruno, P., & Bennett, R. (2020) *Inter-census record-linked entrepreneurs and non-entrepreneurs 1851-91 using BBCE and I-CeM: database structure, assessment, downloads and User Guide*. WP 25. https://doi.org/10.17863/CAM.50180

PP (1879) *Return of the Number of Factories authorised to be Inspected under the Factory and Workshops Acts … inspected 1878*, Parliamentary Papers.

PP (1885) *Return of the Number of Factories authorised to be Inspected under the Factory and Workshops Acts … inspected 1885*, Parliamentary Papers.

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

Smith, H., van Lieshout, C., Montebruno, P. and Bennett, R., J. (2019) *Preparing Scottish census data in I-CeM for the British Business Census of Entrepreneurs (BBCE).* Working Paper 20. https://doi.org/10.17863/CAM.43504

Smith, H., Bennett, R., J. van Lieshout, C., and Montebruno, P. (2021) Entrepreneurship in Scotland, 1851-1911, *Journal of Scottish Historical Studies*, 41 (1), 38-64.

van Lieshout, C., Bennett, R. J. and Smith, H. (2019a) *Extracted data on employers and farmers compared with published tables in the Census General Reports, 1851-1881*. Working Paper 13. https://doi.org/10.17863/CAM.37165

van Lieshout, C., Bennett R. J. and Montebruno, P. (2019b) *Company Directors: Directory and Census record linkage*. Working Paper 14. https://doi.org/10.17863/CAM.37166

van Lieshout, C., Bennett, R.J. and Smith, H. (2021) The British Business Census of Entrepreneurs and firm-size, 1851-1881: New data for economic and business historians. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. https://www.tandfonline.com/doi/full/10.1080/01615440.2019.1707140

Woollard, M., Schürer, K. (2000) *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version).* [data collection]. Federation of Family History Societies, Genealogical Society of Utah, [original data producer(s)]. Federation of Family History Societies. SN: 4177, http://doi.org/10.5255/UKDA-SN-4177-1

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

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

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

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

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

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

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

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

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

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

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

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

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

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

WP 23: Bennett, Robert, J., van Lieshout, Carry and Schürer, Kevin (2020) *Missing in the Census 1851-1911: The 'lost', 'missing', and 'gaps' in I-CeM and BBCE, with weights to adjust RSD populations*. http://doi.org/10.17863/CAM.50179

WP 24: Newton, Gill and Bennett, Robert J. (2020) *Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM. http://doi.org/10.17863/CAM.50178*

WP 25: Montebruno, Piero and Bennett, Robert J. (2020) *Inter-census record-linked entrepreneurs and non-entrepreneurs 1851-91 using BBCE and I-CeM: database structure, assessment, downloads and User Guide. http://doi.org/10.17863/CAM.50180*

WP 26: Bennett, Robert, J., van Lieshout, Carry, Smith, Harry and Montebruno, Piero (2020) *Supplement to BBCE User Guide: Website definitions, downloads, Atlas of Entrepreneurship, and linkage to I-CeM. http://doi.org/10.17863/CAM.50181*

WP 27: Bennett, Robert, J., Smith, Harry and Hannah, Leslie (2021) *Estimates of GB firm-size by sector in 1881 using BBCE sector definitions (EA17 sector codes),*

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

also see
https://www.bbce.uk/