

## Identifying employers in the population census in England and Wales in 1881

ROBERT J. BENNETT AND GILL NEWTON

### Abstract

This paper presents the method and first results of using the 1881 England and Wales Census Enumerators' Books (CEBs) to identify and extract employer records using occupational information. Over 230,000 employers are identified, of which about four fifths employ others. Important sub-groups are also identified of the own account self-employed, company proprietors, directors and partnerships. The paper demonstrates the feasibility of the method and suggests the wider potential for extending it to other censuses, which is now possible using the recently released I-CeM database. The paper also demonstrates some difficulties in the database for 1881, including data keying and coding errors, ranging from 0.5% to 5.5% of entries for larger businesses. Gender miscoding appears to be a systematic error of about 0.7 per 1000 people. The analysis suggests that where small or atypical sample groups are involved, users of the census database should make detailed checks with manuscript CEBs.

### 1. Introduction

Very little use has been made of the British nineteenth century census regarding the population's involvement in businesses as employers and entrepreneurs. Published census reports only systematically tabulated farmers as employers. For one year, 1851, there are published tables of all employers and their male employee numbers by occupations and region; and from 1891 there is a more limited tabulation of employers and own account workers. However, despite the lack of published reports, the census actually collected data on employers and their workforce numbers for all census years 1851-81. At the time, significant criticisms of the accuracy of the returns was used by the General Register Office (GRO) as the main argument for not taking forward any analysis. However, Schürer claims the extent of inaccuracy to be exaggerated,<sup>1</sup> and many have argued that the care of the enumerators and clerks ensured that the censuses were highly accurate, complete and consistent.<sup>2</sup> This indicates that the census might be used to extract large scale information to fill some important gaps in our knowledge of employers for an important period of change in business organisation, and offers the additional advantage of information on employers' personal and household characteristics.

---

<sup>1</sup> Kevin Schürer, 'The 1891 Census and Local Population Studies', *Local Population Studies*, 47, 1991, pp. 16-29.

<sup>2</sup> See e.g. P. M. Tillott, 'Sources of Inaccuracy in the 1851 and 1861 Censuses', in E. A. Wrigley (ed.), *Nineteenth-Century Society: Essays in the Use of Quantitative Methods for the Study of Social Data*, Cambridge, 1972, pp. 82-133; Matthew Woollard, "'Shooting the Nets": A Note on the Reliability of the 1881 Census Enumerators Books', *Local Population Studies*, 59, 1997, pp. 54-7; Matthew Woollard, 'Creating a Machine-Readable Version of the 1881 Census Enumerators' Books', in Charles Harvey and Jon Press, *Databases in Historical Research: Theory, Methods and Applications*, Basingstoke, 1997, pp. 98-101; Edward Higgs, *Making Sense of the Census Revisited: Census Records for England and Wales 1801-1901*, London, 2005, p.112.

Despite the potential value, only limited use has been made of the published census tables for 1851,<sup>3</sup> and there has been no large scale analysis of the original Census Enumerators' Books (CEBs) for the whole country. This paper uses the 1881 census as a pilot for what might be developed as a more general analysis of employers. An electronic database for this census has been constructed by Schürer and Woollard at Essex University, who derived their data from full transcripts of the CEBs hand-copied from microfilm and then keyed by the Genealogical Society of Utah (GSU).<sup>4</sup> They aligned, checked and corrected the GSU transcripts as far as possible to a common format, and supplemented the data in various ways, adding codes for occupations and for household internal structure.<sup>5</sup> This database (hereafter referred to as GSU/Essex) is used here to identify employers from the occupational descriptors recorded in the CEBs. The GSU/Essex database remains the core component for the 1881 census in the newly established I-CeM database for all the censuses 1851-1911, and hence evaluations are pertinent to this landmark new resource.

The paper first presents how the CEBs recorded employers. The paper then presents methods used to extract these records from the 1881 GSU/Essex database, and assesses the accuracy and implications of the results.

## 2. *The 1881 census records of employers*

Nineteenth century census-taking used a household schedule with a table to be filled in either directly by the household head, or by an enumerator who went from door-to-door. The individual household returns were then transcribed by the enumerator into an enumerator's book (CEB) for the entire district that they were assigned to cover. The CEBs are what has been preserved and are the source used in this paper. After completion these books were sent to the GRO in London where they were hand-processed by census clerks. The clerks followed instructions that imposed standardised definitions and classifications where local enumerators had failed to do so, relating occupations to an 'occupational dictionary' of several thousand terms which were classified to several hundred categories used in published tabulations. The final result is generally believed to have achieved a high level of consistency.

The main material used in this paper derives from the CEB occupational entries which record the responses to the census instructions to householders for:

**'Farmers** to state the number of acres occupied, and the number of men, women, and boys, employed on the farm at the time of the Census. ... Sons or daughters employed at home or on the farm should be returned – "*Farmer's Son*", "*Farmer's Daughter*". Men employed on the farm and sleeping in the Farmer's house must be described in the schedule as *Farm Servants*.'

For non-farmers the instructions stated:

---

<sup>3</sup> For England and Wales see e.g. John H. Clapham, **An Economic History of Modern Britain, Vol. 3: Machines and National Rivalries (1887-1914) with an Epilogue (1914-1929)**, pp. 33-7, Cambridge, 1938; for Scotland see Richard Rodger, 'Concentration and fragmentation: Capital, Labour, and the Structure of Mid-Victorian Scottish Industry', **Journal of Urban History**, 14, 2, 1988, pp.178-213.

<sup>4</sup> Genealogical Society of Utah, **British 1881 Census Project Transcription Guidelines**, Utah, 1994.

<sup>5</sup> Woollard, 'Shooting the Nets'; Kevin Schürer, Matthew Woollard, **National Sample from the 1881 Census of Great Britain 5% Random Sample: Working Documentation Version 1.1**, Historical Censuses and Social Surveys Research Group, Essex, 2002.

**‘In Trades, Manufactures** or other Business, Masters must, in all cases, be so designated ... Inserting always the number of workpeople in their employ at the time of the Census. In the case of Firms, the number of persons employed should be returned by the senior or some one partner only’.<sup>6</sup>

The occupational entries derived from these instructions are a mix of words and numbers, forming alphanumeric character sequences (strings) of information on employers, their business, and its workforce by gender and age category, and acreages for farmers. It is these strings which are the subject of this paper, drawn from the variable ‘occup’ in the GSU/Essex database. For operational reasons, GSU restricted the ‘occup’ variable to 80 characters, with longer entries abbreviated to fit.

GSU keying of the 1881 census into an electronic database started from the CEBs. Hence, it received no benefits of the standardisation and instructions that were followed by the GRO clerks in London. However, any marks made by the clerks are preserved in the CEBs and in most cases they have been transcribed by GSU into the database, though they were keyed in any order and might be broken up across the record, reflecting how clerical annotations appeared. This has dangers of introducing spurious and confusing information in subsequent processing.

The Essex processing started from the GSU keying and then coded occupations following as closely as possible the rationale of the census enumerating clerks, trying to impose similar consistency. Of the resulting 416 census occupational categories for 1881, 333 contain employers extracted in this paper (within which 293 were non-agricultural).<sup>7</sup> However, the GRO clerks were able to inspect occupations in the context of the surrounding record, whereas automated processing is blind to context unless its utility has been anticipated. This leads to some errors in GSU/Essex occupation codes. For example, the census instruction sought information on ‘Rank, Profession or Occupation’.<sup>8</sup> This was used by census clerks and in publications to classify rank in society as well as trade, skill or occupation.<sup>9</sup> For published tables the clerks were instructed to ignore these ranks and ‘to select for ticking that Occupation which would seem the main or more important one’. Where only rank was returned they were ‘to be ticked to the Unoccupied group’.<sup>10</sup> However, when coded by GSU/Essex some of these choices were difficult to replicate with certainty and it is clear in the subsequent analysis that many were coded to unoccupied or unknown even though information exists which allowed clerks to code to specific fields. As a result, understandably, the GSU/Essex information is subject to various constraints and needs care in interpretation and, where possible, correction.

There are some constraints on what can be achieved from extraction of employers. First, in 1881, distinct from earlier practice, clerical instructions sought to exclude household work of wives, daughters and other unpaid household workers, and any work female relatives did on a male-run family farm.<sup>11</sup> These instructions also influenced the CEBs, and makes the

---

<sup>6</sup> ‘General Instructions’, Census of England and Wales, **Householder’s Schedule**, 1881.

<sup>7</sup> Matthew Woollard, **The Classification of Occupations in the 1881 Census of England and Wales**, Department of History, Essex, 1999, pp. 6-8.

<sup>8</sup> Heading Instruction to ‘General Instructions’, Census of England and Wales, **Householder’s Schedule**, 1881.

<sup>9</sup> Woollard, ‘Classification of Occupations’, p. 2, 7; Edward Higgs, ‘Life, Death and Statistics: Civil Registration, Censuses and the Work of the General Register Office, 1836-1952’, **Local Population Studies, Supplement**, Hatfield, 2004, 56-64, 93-4.

<sup>10</sup> GRO, **Instructions for Coding Occupations**, 1881, Nos. 1 and 2, p.2: TNA RG 27/5.

<sup>11</sup> GRO, **Instructions for Coding Occupations**, 1881, No. 4, p.3: TNA RG 27/5.

published census tables and GSU/Essex tabulations non-comparable on details, especially for women on farms and for outworkers.<sup>12</sup> For example, where a farmer returned ‘hands’, ‘workers’ or ‘labourers’ which included wives, daughters and female relatives these will be counted, but where they were explicitly identified as ‘wives’ etc. they will have been excluded in published tables. In some cases we found that employers attempted to account for their outworkers, but this was not widespread.<sup>13</sup> Higgs notes that the instructions were biased towards identifying external work rather than home-workers and domestic businesses, and against recording women; he estimates perhaps 50% of homeworkers as missing or misleadingly recorded in the published tabulations.<sup>14</sup> For employers, the gender bias of the instructions to ‘masters’, may also have discouraged identification of females as employers or partners. This means that all analyses of female employers, especially farmers, have to be treated with caution.

In addition, the 1881 instructions asking farmers to state ‘the number of men, women and boys’, and non-farmers to state ‘the number of workpeople’ were inconsistent between farmers and non-farmers. They discouraged responses identifying workers by age and gender for non-farmers, leading to conflated numbers. However, the actual CEBs are often detailed, allowing age and gender breakouts, although there is also a large category of businesses where only the total number of ‘employees’, ‘hands’ or ‘other’ workers were returned.

One other major source of uncertainty derives from the wording of the instructions which led to the exclusion of wide classes of the self-employed who employed only themselves. There was no explicit instruction for enumerators to include these.<sup>15</sup> The GRO was aware of the difficulty. For example in 1851 it was recognised that ‘many employers ... in some trades nearly the whole number, omitted to attend to the instructions’ and in London ‘in some trades more than half of the whole number’ failed to record their employer or self-employed status.<sup>16</sup> Comments on 1881 by William Ogle, Superintendent of Statistics for GRO Census administration, stated that it was very difficult to use the census to distinguish between masters and men in many fields; for example ‘the great mass of tailors ... put themselves down simply as tailors, and you are left in doubt whether it is a master or a journeyman tailor’.<sup>17</sup> The same view is carried into the 1891 Census Report, which states that the householder ‘too commonly, neither cares for accuracy or is capable of it’ ... It is not wise to demand ... a result for ... which it is unsuited’.<sup>18</sup> However, Schürer suggests these claims are exaggerated, noting that even for 1891, when there may have been more inaccuracy than earlier years as a result of new instructions, the Census Report quotes only a 5.3% error in

---

<sup>12</sup> Joyce E. Bellamy, ‘A Note on Occupational Statistics in British Censuses’, **Population Studies**, 6, 3, 1953, p. 306; Catherine Hakim, ‘Census Reports as Documentary Evidence: The Census Commentaries 1801-1951’, **Sociological Review**, 28, 3, 1980, p. 557. Edward Higgs, Edward (1987) ‘Women, Occupations and Work in the Nineteenth Century Censuses’, **History Workshop**, 23, 1987, 59-80; ‘Occupational Censuses and the Agricultural Workforce in Victorian England and Wales’, **Economic History Review**, 49, 1995, 700-16.

<sup>13</sup> e.g. Henry Dishley, ‘silk manufacturer employing 19 in factory 62 at own home’; and James Smith, ‘builder employing 185 men and 16 boys unable to return number of pieceworkmen’ (GSU/Essex database): RG11/2738/12 and 819/76.

<sup>14</sup> Edward Higgs, ‘Women, Occupations and Work’, p.67

<sup>15</sup> Also some of those who did respond as masters may have been ‘outworkers’ for a ‘giving-out employer’. See Clapham, ‘Economic History Vol. 3’, p.35.

<sup>16</sup> BPP Census 1851, **General Report**, p. lxxvi; **London County Report**, p. 28.

<sup>17</sup> Ogle, Questions 138-9, in **BPP, LVIII, Report of the Committee Appointed by the Treasury to Inquire into Certain Questions Connected with the Taking of The Census**, C 6071, 1890, p. 6.

<sup>18</sup> **BPP Census Report**, 1891, Vol. III, p. 35.

ticking boxes, or failing to tick any box.<sup>19</sup> Indeed Woollard concludes that across the country these errors are not large and there is generally ‘a high level of internal consistency’.<sup>20</sup>

For those employers who employed others than themselves, however, it is likely that the returns were more reliable. For the 1851 census tabulations Clapham notes ‘the results are imperfect ... but it is safe to assume that most of these [errors] were small’; later he suggests that the trades where few returns of employees were made comprised mainly solitary masters, and those who did return employees included almost all larger businesses (albeit not in all sectors).<sup>21</sup> One sector noted by Clapham as having discrepancies in 1851 was cotton manufacture, where the number of large employers and their employees does not match the published tables of those occupied in the sector. By 1881 such a discrepancy would appear to result not from underestimation of the workforce but by some business owners not responding as employers at all. Many larger concerns were now companies and a legalistically correct response to the census instruction was that the company employed the workers. The respondent could thus describe themselves as a director who did not personally employ anyone. For this reason a range of business role descriptors such as ‘director’ and ‘partner’ must be examined.

### 3. *Extraction method*

The scale of the analysis, which requires searching the entire 1881 population of 26 million persons, meant that automated data extraction from the GSU/Essex database had to be adopted. Firstly, an algorithm was written to search the alphanumeric occupation string for each record and extract those likely to represent employers. This depended on the presence of keywords representing employee types (with spelling and gender variants), the verb ‘employ’ and its abbreviations or synonyms, and pattern-matching of typical ordering and phrasing of employee types and numbers. The terminology for employees varied widely and a range of search terms was necessary, including ‘hands’, ‘labourers’, ‘persons’, ‘assistants’, ‘apprentices’, ‘servants’, ‘clerks’, etc.

The extraction was then parsed by algorithm to split the composite employer strings into separate variables/fields. Parsing proceeded from a library of expected patterns determined by inspection and was extended iteratively to encompass the most numerous unforeseen exceptions in the un-parseable residual. By this means 81% of the strings were separated into occupational sector and number of employees by category. The residual of 28,472 complex strings (19% of the total) that could not be automatically parsed was processed by hand using research assistants and checked by the authors. Most involved systematic errors; e.g. non-employers included because of associated number sequences. These were removed. In some cases checks with the manuscript CEBs at TNA revealed coding annotations by census clerks to have been keyed by GSU as number sequences that introduced spurious employees.

For farmers the additional information available from the instruction ‘farmers to state the number of acres occupied’ was parsed more easily by algorithm. With considerably less iterations of the search terms, 90,455 acreages from 101,029 farmer strings were soon

---

<sup>19</sup> Schürer, ‘The 1891 Census’, p. 26; also see Eilidh Garrett, Alice Reid, Kevin Schürer, and Simon Szreter **Changing Family Size in England and Wales: Place, Class and Demography, 1891-1911**, Cambridge, 2001, pp. 69-71; quoting **BPP Census Report**, 1891, Vol. III, Table 5.

<sup>20</sup> Matthew Woollard, ‘Shooting the Nets’, p.57.

<sup>21</sup> Clapham, ‘Economic History Vol. 2’, p. 448, n.2; Clapham, ‘Economic History Vol. 3’, p.35.

parsed.<sup>22</sup> This left a residual of 10.5% which was processed by hand using research assistants and checked by the authors.

The purpose of the extraction was, as far as possible, to identify all economically active employers and their businesses, their employees, and in the case of farmers also the acres farmed. The presence of employees provided the usual means of identification. Employers returning no employees were more difficult to identify. For farmers the occupational coding present in the GSU/Essex database enabled all those not already extracted to be added. For other employers without employees it was insufficient to rely upon ‘master’. The use of the term ‘master’ dated from instructions used in 1851 and all intervening censuses, which in turn reflected the traditional concept of master trades with apprentices, journeymen, and other employees.<sup>23</sup> By 1881 this term was evidently becoming obsolete. It was commonly replaced simply by ‘employing’ following the occupational descriptor of the respondent, where employees were present (such as ‘tailor employing 1 man’), or otherwise no term at all because the occupation was almost entirely own account self-employed. The extraction used a wide range of search terms in addition to ‘master’. The frequency with which the main descriptors were used by respondents gives useful insights into how employers viewed themselves in 1881. As shown in Table 1 it is clear that the generic term ‘employing’ was the most widely used, and would probably have been even more frequent if the census had not prompted the use of ‘master’.<sup>24</sup>

<b>Search term (with synonyms and alternative spellings)</b>	<b>Approximate frequency of descriptor strings</b>	<b>Persons identified mainly by this descriptor</b>
Employer/farmer (including those identified solely from employees)	c. 72,000	182,892
Master (including females)	63,418	38,398
‘Spurious master’, <sup>25</sup> but with second real employer status	1,404	incl in employers
Land owner	5,027	4,819
Proprietor	4,338	incl in employers/owners
Partner	2,164	1,636
Shipowner	1,477	1,434
Landlord/landlady	1,188	incl in employers
Mine/quarry owner	808	801
Director	587	397
President of company	41	5
Head of firm	37	incl in employers
Promoter	9	incl in employers
Boss	2	incl in employers
<b>Total</b>	<b>152,147</b>	<b>230,382</b>

Table 1 Approximate frequency of terms used in strings and by individuals to describe employers, after removal of spurious or inaccurate entries.

<sup>22</sup> Excluding farmers with multiple occupations.

<sup>23</sup> C. R. Dobson, *Masters and Journeymen : A Prehistory of Industrial Relations 1717-1800*, London, 1980.

<sup>24</sup> Multiple terminology occurs in many returns leading to some double counting in Table 1; hence frequencies in the larger categories must be regarded as approximate.

<sup>25</sup> e.g. persons such as postmasters or stationmasters who were not deemed to be employers in their own right.

A range of other searches was made to ensure all employers were identified. For farmers, farm bailiff was included as this captured those operating as managers who returned employees and/or acreages who would otherwise be omitted (accounting for about 1% of farmers). Land owners were also included, although many were absentee or had been assigned to a non-farming category (usually as possessor of property or rank). Several possible additional search terms were of little utility since, although identifying some employers, they also identified many of employee status preventing them being used for simple algorithmic extraction; e.g. agent (55,811 strings), broker (5,881), factor (c 1,300), banker (955), trader (498), and governor. Other terms had too varied a response to be useful; e.g. ‘corporation’, ‘associate’, ‘trust/trustee’. ‘Mill’ yielded 167 entries, ‘company/Co/Ltd.’ 133, and ‘concern’ nine, most overlapping with other employer identifiers. ‘Owner’ identified shipowners and mine/quarry owners. Other search terms produced almost no valid respondent; e.g. ‘combine’, ‘combination’, ‘participant’, ‘consignee’ or ‘Nabob/Nawab’. An attempt was made to identify entrepreneurs acting as active investors and directors of enterprises or financial and overseas projects. The 7,789 returned in 1881 as ‘deriving income from dividends’ were analysed but almost all were pensioners or annuitants; i.e. rentier retail investors and shareholders. There was no unequivocal case of currently active entrepreneurial investors in this group. Checks on some known active investors in the original CEBs were all returned in other categories as a barrister, accountant, banker, broker, agent, etc. It is noteworthy that none were identified in the CEBs as an ‘entrepreneur’, indicating that despite the large academic literature on this topic, this term was not used in connection with occupation or to describe business roles.<sup>26</sup> The limited use of the title ‘boss’ indicates the lack of diffusion of a term widely used in the US but mainly restricted to gang leaders, especially on railways, rather than business heads in this period.<sup>27</sup> Another Americanism in its business sense, ‘president’, in the CEBs mainly referred to institutional positions, but did identify five employers, although at least two were foreign-based.<sup>28</sup>

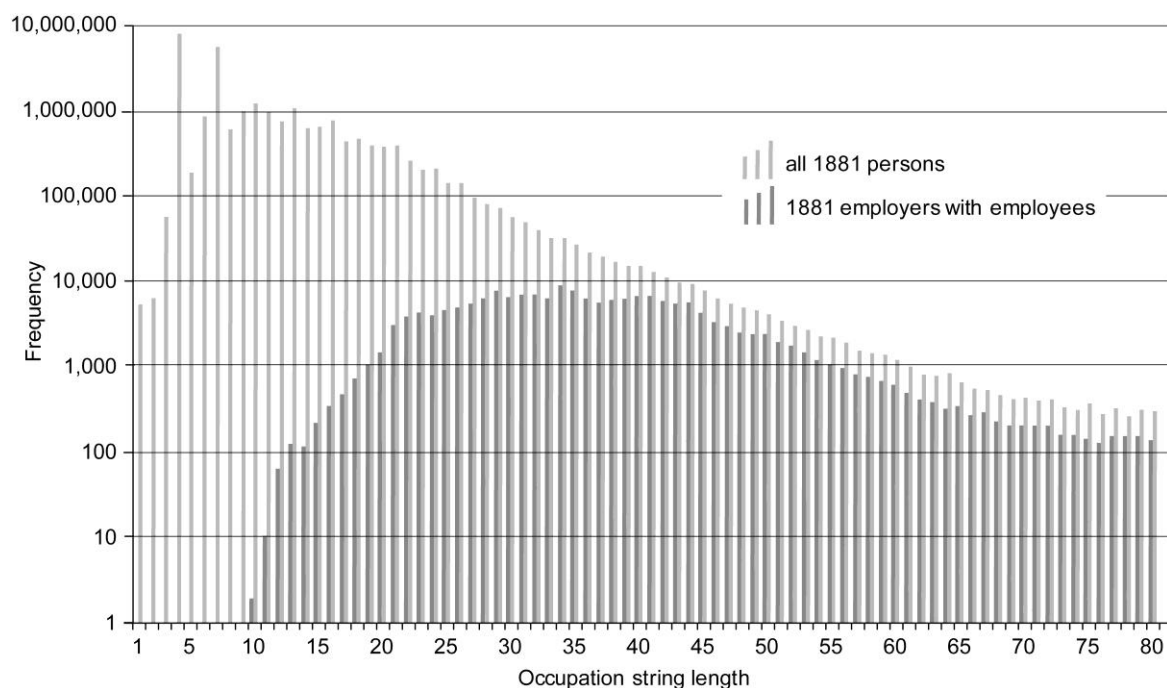
A final check on the comprehensiveness of the algorithmic extraction of employers used string length, which is usefully independent of CEB terminology. The census instructions to employers and farmers resulted in longer strings of textual description in the ‘occup’ field of the Essex/GSU database. As shown in Figure 1, mean string length for employers was 38 characters compared to 10 for all occupied individuals. For strings over 15 characters the extraction identified the majority as employers, and for strings over 49 characters over 95% were employers. A hand-check of all strings of over 49 characters, whether already identified as employers or not, yielded some unusual variants on ‘employs’ not previously encountered, which when added to other algorithmic search terms captured a further 121 employers not previously extracted.

---

<sup>26</sup> The Oxford English Dictionary documents several nineteenth century uses of ‘entrepreneur’, notably as a loan-word from French in an 1852 letter by Carlyle. In the later 1880s American political economists were using it for businesses, but the 1880 US Census individual level data had no instances of ‘entrepreneur’. Differences are more apparent in the use of ‘speculator’. In the England and Wales 1881 Census only five were ‘speculators’: all were builders also recorded with employees; whereas in the US 1880 Census more than two thousand persons were recorded as speculators. Steven Ruggles et al. ‘Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]’, Minneapolis: University of Minnesota, 2008.

<sup>27</sup> The US 1880 Census enumerates 6,592 bosses out of a total of some 50 million persons, but it returns twice as many (13,602) foremen, then the UK English corollary (Ruggles, *ibid.*)

<sup>28</sup> Like ‘boss’, ‘president’ in business use may have a railroad origin. See note 27.



**Figure 1.** String length for employers compared to other individuals

#### 4. Refining the extraction

Thorough checks were undertaken to exclude spurious employers from the extraction. These excluded non-business masters; e.g. school master/mistress, station master/mistress, riding master, master gunner, master at arms, drill master, choir master, workhouse master (and other institutional masters/mistresses). Numerous post masters/mistresses were also excluded where having no other occupation, as these should be regarded as employees/agents of the Post Office rather than employers in their own right. However, boat and barge masters, and master mariners, were not removed as they were normally own account employers. Non-employing groups that contained spurious entries were also excluded: e.g. students, clerics, domestic servants, foremen, military personnel, etc. (usually identified by the algorithm because of including ‘of/to master’). Foreign-based employers present on census night were excluded as the aim is to identify employers in England and Wales.<sup>29</sup> In addition hand checks on the database were made for errors for all of the approximately 2,500 employers who appeared to have 100 or more employees (or the words ‘hundred/s’, ‘thousand/s’, ‘score/s’, ‘tens’, ‘dozen/s’ or ‘many’).<sup>30</sup> Subsequently, checks against manuscript CEBs were also made (see later below). These corrections sought to ensure that errors were eliminated as far as possible in the large employer categories, which contain small numbers of entries and so are more vulnerable to distortion. Further checks were also made of the occupational codes against the descriptors. These were found to have over 18,000 coding errors (about 0.08%),

<sup>29</sup> There were few of these, but some were very large land owners abroad, such as Anthony C. Williams of Dursley, Gloucester, ‘Queensland farmer of 384,000 acres’; the two foreign-based ‘presidents’ excluded were of the Railway Company of Texas, and Grand Trunk Railway of Canada.

<sup>30</sup> An error in the original household statement of a tobacco manufacturer ‘employing 11600 hands’, which exceeds the entire employment of the sector, was assumed to be 1600.



mostly for cases with long strings. These were corrected by algorithm and by hand. They mainly result from complex strings where descriptors of rank have been muddled or incorrectly interpreted as occupations.<sup>31</sup>

Other checks identified 877 employers returning 'wives' as employees. A systematic hand check on these revealed that almost all were spurious, and the putative employer was herself a wife whose occupational statement duplicated that of her husband, including the employee declaration, but with the word 'wife' appended. It is possible these were husband and wife partnerships, but more likely that some enumerators returned occupations of all wives as duplicates of the husband. Only 28 employers unambiguously declared wives as employees, but these were usually other people's wives, such as 'farmer employing four labourers and their wives'. Conversely, a shoemaker employing a son or daughter, for example, especially if an apprentice, would often declare them as an employee.

A check on age removed some questionable entries of other types. It is recognised that people became employers or acted on their own account from a young age.<sup>32</sup> In general the algorithm was primed to exclude employers under 16 years old; however, if, after hand checks, younger individuals could not be ruled out as errors they were included as real employers if aged 14. Similarly 'retired' employers were excluded as no longer economically active, except where explicitly returning employees or acreage.

Within the final extracted employers some had multiple occupations. Some of these resulted from the census instruction to return 'rank, profession or occupation'. These are spurious to the purpose here since they were not multiple businesses; e.g. 'baronet and farmer', 'JP and millowner', 'captain of Salop rifles ironmaster employing about 1000 men'. These status terms were removed from employer coding. Other records had been coded by GSU/Essex to farming irrespective of the size of their other business or the order given in the returns; e.g. 'flax spinner and farmer of 350 acres employing 400 workpeople' was coded as a farmer even though spinning was clearly the major business. This reflects the instructions to census clerks but obscures the analysis of employers.<sup>33</sup> Where detected, all these cases were reassigned by hand to their presumed primary business, chiefly following the instruction in the household schedule that 'A person following more Distinct Occupations than one, should insert each of them in order of their importance'.<sup>34</sup> After adjustments about 10% of non-farm employers and 7% of farmers had two or more other businesses. However, given the number, functional importance, and large size of many of the multiple businesses, a detailed analysis of these is undertaken elsewhere.

## 5. *Reconstituting partnerships and dealing with companies*

The census recognised that many employers were heads of firms or parts of partnerships for which the instruction given was that 'the number of persons employed should be returned by the senior or some one partner only'. Responses to this instruction helped to extract employers and directors, but created two problems. Firstly, many partners return the same

---

<sup>31</sup> About a third were 'unoccupied' landowners coded to farming; many others were dual occupied 'farmers' whose principal business lay typically in manufacturing.

<sup>32</sup> Woollard, 'Classification of Occupations', p. 8.

<sup>33</sup> **BPP, Census of England and Wales, General Report**, 1881, p. 28.

<sup>34</sup> Heading Instruction to 'General Instructions', Census of England and Wales, **Householder's Schedule**, 1881.

information, resulting in duplication which, if not controlled for, will lead to overestimation of business numbers in subsequent analysis. Secondly, partners need to be identified and matched to each other as far as possible to reunite their personal characteristics with those of their business. This recovers the true business status of many employers with apparently zero employees or missing occupational/sectoral details.

The partnerships identified fall into four groups:

- (i) Responses with employee numbers from one or the senior partner.
- (ii) Responses from each partner.
- (iii) Directors/managers of named companies.
- (iv) Records where nothing is stated beyond 'partner' or 'director', or the detail given is insufficient to identify the business or other partners.

In the case of (i) the result is one record containing employees/acres. Other partners living together can usually be detected and the composition of the partnership understood because their descriptors are identical or similar. Adjacent positioning within the household is often referred to explicitly; e.g. 'partner of above'. Where partners lived apart or more distantly it is less easy to reconstitute the partnership unless they have distinctive business descriptors. Otherwise the best that can be achieved is to record them as a 'partner' but with no specific business information attributed.

Duplicate records of a business, as in (ii), resulted where more than the senior or a single partner gave information on employees. Often, but not exclusively, this occurs because of a 'ditto' in the household entry. This may cover all partners, or some of them. Fortunately, the complexity of strings for many of these records gives a very low probability of duplication of business information unless they were real partners. Partnership was assumed where a potential pair of partners possessed at least two identical features of occupational descriptor (including number of employees/acres), and were not substantively different in any other features. Thus, partners can often be readily matched irrespective of residence (more readily than many in group (i)).

For the directors in group (iii) there was no requirement to state the company name but some did, although whether all directors of the same company did likewise is difficult to determine. Detailed cross-matching against available director lists of companies may make it possible to enlarge information on this category, but has not been undertaken here.

From this discussion it is clear that the census does allow useful scope to assess partnerships and companies but this is mainly restricted to:

1. Businesses with very distinct descriptors provided that all partners gave similar descriptions of the business.
2. Businesses with less distinct descriptors whose partners are co-resident, or in close proximity. These mainly mostly represent family businesses. However, *de facto* partners who shared some business overheads (such as premises) but who ran different or overlapping businesses will be omitted; e.g. a blacksmith and wheelwright using the same workshop; a grocer and haberdasher using the same shop.
3. Directors/managers of companies, but only a proportion of the whole, within which unfortunately only a few give employee numbers.

It is noteworthy that there appear to be systematic differences in the extent of partnerships recorded. Several places have virtually no partnerships recorded, which is unlikely to reflect

reality; e.g. most of Norfolk (excluding Norwich); most of rural Suffolk, Lincolnshire and Nottinghamshire; and towns such as Nuneaton and Beverley (Yorks. East Riding). There may be some differences in propensity to have partnerships between areas, or for individuals to report them, especially for different forms of farm tenure or localised differences in manufacturing organisation. But it is most likely to reflect enumerator practice. It would have involved significantly more work for an enumerator to interrogate each householder to determine their business relationships with others, and this might have met resistance. Similarly it would be onerous for householders to take this upon themselves. It was easier to record people simply to their occupations. No attempt has been made here to compensate for such differences. However, their existence results in the partnerships recorded by the census representing only a proportion of all partnerships that existed.

## 6. Accuracy of the GSU/Essex database

The extraction of employer records provides a foundation for a wide range of subsequent analyses which are investigated in other publications. Here the purpose is to demonstrate the potential of I-CeM to act as a valuable resource for analysis of employers, for which a critical issue is accuracy of the extractions obtained. The preceding discussion has identified limits to accuracy, some of which do not appear to have been previously reported by users of the GSU/Essex database.

The largest businesses, because of their small number and complex descriptors, are particularly sensitive to errors. To assess these errors individuals were checked against the manuscript CEBs in rank order of workforce size, for the largest 303 female-headed businesses with 25 employees or greater, and the largest 166 male-headed with 850 employees or greater. Table 2 summarises the main coding errors detected. Whilst there is only one error for males, one sixth of female large businesses are coded incorrectly as employers. The most frequent issue arises from part entries of the husband's business that are incorrectly ascribed to the wife in the GSU/Essex coding. Other errors are displaced and muddled entries of particularly long occupational descriptor text strings that overspill the box provided in the manuscript *pro forma*. There are also gender miscodes.

A typical example of an incorrectly specified female business is the largest nominally identified by the extraction method. This is for Anna Shaw of Colne (Lancs.), who is shown by GSU/Essex as 'employing about 1200 workpeople & farmer 94 acres employing 2 men'. The manuscript CEB entry shows her husband Robert Shaw as 'Magistrate / Cotton Spinner and Manufacturer / employing about 1200 workpeople & / farming 94 acres, employing 2 men'.<sup>35</sup> Lines one and two are in the box on the census form against Robert, lines three and four are in the box against Anna. The GSU keying operative has separated the entry in two even though a clear large bracket links the entry in the original CEB solely to Robert. There is no evidence that Anna was a business partner (though of course she may have been unrecorded). In addition, the next line entry, for Robert a son aged 34, shows him as 'Cotton Partner with his father'. This suggests that Anna, aged 69, was indeed not a partner, otherwise she would probably have been similarly shown. Moreover, Robert senior was aged 71 so that the structure of the business was probably that he remained the owner, but his son had assumed managerial responsibility.

---

<sup>35</sup> Manuscript CEB: TNA RG11/4163/91; slashes indicate the line breaks in the hand written return.

	Female-headed business		Male-headed businesses	
	No. of people	%	No. of people	%
Business correctly shown	255	84.2	165	99.4
Business is wrongly shown as wife/husband, when no evidence of involvement	36	11.9	-	-
Business is displaced return with owner omitted altogether	4	1.3	1	0.6
Business is incorrectly coded by gender	7	2.3	-	-
Separate business descriptor muddled with another	1	0.3	-	-
<b>Total</b>	<b>303</b>	<b>100</b>	<b>166</b>	<b>100</b>

Table 2. Errors in correctly identifying businesses from GSU/Essex for the largest female and male-headed businesses, compared to manuscript CEBs.

This example demonstrates the difficulties of the census enumerators, and GSU keying operatives, in interpreting the CEB employer records which are frequently very long with the census form giving little space. The longer and more complex the occupational descriptor, the more likely entries ran over against other individuals in the following lines. As shown in Table 2, it is particularly prevalent among wives who were erroneously identified as engaged in the business. The GSU/Essex coding for the Shaws also shows a range of other problems. Robert senior and Robert junior are identifiable as employers but have no identified employees. Essex coded Robert senior and Anna to farming, whilst only Robert junior was correctly identified as a cotton spinner. The dual, and clearly primary, occupational status of Robert senior can be determined only by inspecting the manuscript CEB. For subsequent processing, to be consistent with the rest of the census records, Anna's employer status has to be removed and the correct information added to the husband and son.

Table 3 shows other errors within descriptor strings that are additional to those in Table 2. The most frequent is textual keying of the occupation or the name of the individual. Despite this, almost all these occupations have been coded correctly by GSU/Essex. There are disturbingly large numbers of errors in digits for employees, acres and age. These mostly derive from confusion of editing marks made by census clerks which have then been extracted as numbers of employees. Most number errors result in overestimates for persons with no or few employees. In addition there were cases of not identifying the occupations of partners, additional to the confusion of wife/husband status shown in Table 2.

	Female-headed business		Male-headed businesses		All businesses	
	No. of people	%	No. of people	%	No. of people	%
Employee nos.	18	5.9	6	3.6	24	5.1
Acres	5	1.7	-	-	5	1.1
Age	2	0.7	-	-	2	0.4
Occupation/code	26	8.6	6	3.6	32	6.8
Missing dual	3	1.0	4	2.4	7	1.5
Missing partners	2	0.7	2	1.2	4	0.9
Keying errors for names, occupations etc.	33	10.9	17	10.2	50	10.7
<b>Total</b>	<b>89</b>	<b>-</b>	<b>35</b>	<b>-</b>	<b>124</b>	<b>-</b>
<b>N/% of records with ≥ 1 error</b>	<b>55</b>	<b>18.2</b>	<b>14</b>	<b>9.6</b>	<b>69</b>	<b>14.7</b>

Table 3. Errors of detail in GSU/Essex coding (additional to entries in Table 2), compared to manuscript CEBs. Multiple errors for same individual mean percentages do not sum to total.

For almost all types of error there is a higher frequency for females than males, but since complex and lengthy descriptor strings are a major cause of error it is likely that the records for smaller and less complex businesses have a higher rate of accuracy. To test this, the error rates shown in Tables 2 and 3 were calculated against the size of the business's workforce. The proportion of records with errors from all causes does decline with employee size, from the very high rate of 70-80% for the largest 20 female-reported businesses, to a mean of 20-30% for those in the largest 50. The miscoding of wives/husbands error rate declines more rapidly, from 50% for the largest 10 businesses to about 10% for the largest 50, of which about half are actually in partnership with their husbands, meaning that partnership status may be missing, but the business is correctly identified. The errors of number are more erratic, but average about 10% for the largest 50 businesses. Errors of gender are also erratic by size, averaging 2.6%. However, a further check of the largest 1,000 male- and female-reported businesses against the CEBs shows that after the first 300 female and first 200 male employers by workforce size, errors settle to about 5% of female and 1% of male entries, and fall to below 1% after the largest 500 businesses for both genders. For smaller and less complex businesses, therefore, errors become relatively minor.

Some errors appear to be largely independent of business size. Some occur through abbreviation to fit the 80 character limitation of the occupation field. This is generally interpretable, but needs care. Thus the long entry of Richard Garrett of Kelsale, Suffolk, is abbreviated by GSU/Essex to 'Eng x/1 carry on bus at Leiston woks emp 500 hnds frmr & br emp 50 hnds & 30 hor', resulting in coding as farmer by GSU/Essex. The actual entry reads: 'Engineer carrying on business at Leiston works emp 500 hands, farmer & brickworks emp 50 hands & 30 horses',<sup>36</sup> with engineer and brickworks unrecognised in GSU/Essex.

However, the most important systematic error is mistaken gender. Schürer and Woollard recognised errors in GSU keying of gender and made consistency checks with relationship to head of household and other fields, and deployed a forename-sex dictionary to correct the

<sup>36</sup> Manuscript CEB: TNA RG11/1890/15.

final GSU database.<sup>37</sup> However, this appears to have been imperfect for both GSU and Essex corrections. The gender errors found through checks of the manuscript CEBs, shown in Table 2, are all detectable from the names of the individuals: at least four unambiguously male names were miscoded as female (John, Thomas, William, and Edwin). Although a female may occasionally take a male name, this was not the case in any of these cases since all had wives, and the enumerator marked the correct male column.

The extent of mis-gendered names was checked for the entire business database. In addition, for comparison, a random sample of 10,500 non-business records with similar characteristics was checked (age over 16, occupied, not retired, and with the same gender balance). In each case ambiguous first names, family names, indeterminate names and initials were excluded, as well as first names/abbreviations used for both genders or where mis-keying would confuse (e.g. Christia/en/ne, Clare/Clarence, Elisha/Elis/Elisa/Ellis, Emile/Emma/Emmanuel, Florence/Florian/Flora, Frances/is, Fred/Frederick/a, Harriet, Jessie, Julian/ne). This reduced the sizes of the samples by about 4%, mostly from the exclusion of initials, Fred/Frederick/a, and Frances/is (accounting for 80% of exclusions).

Table 4 compares male names recorded as females, and vice versa for: (i) the largest 1000 businesses checked with manuscript CEBs, (ii) the full business database, and (iii) the sample of occupied non-employers. The comparisons are not entirely like-for-like since the first is actual misclassification against the CEB, whilst the others are projected errors based solely on first names. However, the comparisons are striking. The overall error rate in the two comparison samples was similar at about 0.7 per 1000 entries, though both show a much higher error rate for females misclassified as males than vice versa. The comparison with the CEBs found a much higher error rate overall, but a reverse pattern with a much greater proportion of males misclassified as females, which mainly derives from keying the complex descriptors of large businesses that often affect the wives' entries. Hence, the significant problem found for large businesses does not extend at the same scale over the whole GSU database, but is nevertheless disturbing. Table 4 suggests that, without correction, female-headed businesses will be slightly under-estimated in general, but more significantly over-estimated among the largest businesses. It also suggests that *any* analysis of the GSU database should check for possible gender mis-coding. Particular care is needed where small sub-samples are involved, and where accurate comparisons of gender are attempted. In the extraction here all miscoded genders were corrected.

	Male coded as female		Female coded as male		Total miscoded	
	N	Per 1000	N	Per 1000	N	Per 1000
Largest businesses <sup>1</sup> : CEBs checked	27	27	-	-	27	13.5
	1000		1000		2000	
Businesses: names	112	0.58	38	2.52	150	0.72
Total N	192546		15073		207013	
Non-employer occupations: names	2	0.21	6	8.71	8	0.76
Total N	9344		689		10033	
<b>Total</b>						

Table 4. Errors in gender codes for business and non-business samples.

<sup>37</sup> Schürer and Woollard 'National Sample', pp. 19-20.

## 7. Conclusion

This paper presents the method and first results of using 1881 census CEBs to extract employer records. Over 230,000 employers are identified, of which about four fifths employ others. This is an unprecedented resource for the study of business structure and employer characteristics in a critical period of industrial activity and sectoral change and, importantly, should be replicable for other nineteenth century censuses to extend the time period covered.

The extractions are most complete and accurate for employers who have employees and, because of the nature of the census instructions, for farmers. The main omission from the extraction in terms of overall number of employers is many of the self-employed who employ no one else. For these perhaps only 10-20% are identified by the extraction method using the census instructions for returning ‘employers’ and ‘masters’ alone. Hence, any analysis will be most robust where it focuses on those employing others. The extraction also identifies an important sub-group composed of company owners and/or their directors. Comparisons are difficult and uncertain, but perhaps about one quarter of companies and one tenth of their directors are identified. A further group of business partners can be identified. Although this is perhaps only about 10% of the true number of partnerships, it provides a large sample of partnership businesses.

We have demonstrated the feasibility of using the census occupational strings to identify a large proportion of the more significant employers in England and Wales, as well as parts of important sub-groups of the own account self-employed, company proprietors, directors and partnerships. This suggests the wider potential to extend the methods to other censuses of the period, which is now possible using the recently released I-CeM database.<sup>38</sup> The 1881 component of I-CeM remains the same (using the GSU/Essex database), so that the results quoted here will apply equally to I-CeM for 1881. For other censuses that are now available electronically, the methods of extraction developed here will be broadly applicable, albeit with adaptation.

We have also identified certain specific difficulties in the GSU/Essex database for 1881, some of which probably affect other census years in I-CeM. A range of data keying and coding errors of number and gender occur, ranging up to 5% of entries of number and about 7% of occupational codes for the larger businesses. In addition extensive processing was required to identify and recode or add additional codes to dual or multiply occupied individuals in the extraction. Of course, it is understandable that some errors by GSU/Essex were encountered, particularly since the employer strings are among the most complex and unusual of the census returns. Where a large business is described, the character string given in the occupation field is often unique to one individual, whereas the greatest efforts made at Essex in rectifying mistakes in the GSU keying were mainly directed at frequent strings shared by many persons.

Our analysis suggests that users of the 1881 census database, either directly through the original GSU/Essex database, or through I-CeM, would be well-advised to exercise care in several respects. For large-scale statistical analysis the errors noted here will usually be immaterial. However, where smaller sub-samples are involved, especially where accurate comparisons of gender are needed, and certainly where embedded digits in the occupation or

---

<sup>38</sup> Edward Higgs, Christine Jones, Kevin Schürer and Amanda Wilkinson, **Integrated Census Microdata (I-CeM) Guide**, Department of History, Essex, 2013.

other fields, such as employer's employee numbers or farm acreage, are the subject of analysis, detailed checks with the manuscript CEBs are necessary, and corrections may be required. The need for corrections is particularly strong where the focus is on exceptional persons, whether it be those with the largest businesses or farms, as here, or other criteria that sample persons who deviate from the typical case.

**Acknowledgments:**

This research was financed by the Leverhulme Trust, Research Grant RG66385.

**Authors:**

Robert Bennett is Emeritus Professor of Geography, University of Cambridge and has published on small businesses, entrepreneurship and the history of business associations. Gill Newton is a Research Associate at the Cambridge Group for the History of Population and Social Structure and has worked on occupations in nineteenth century censuses.