Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vhim20

Geo-coding addresses in historic British census data: An open methodology

Joshua Rhodes

To cite this article: Joshua Rhodes (2025) Geo-coding addresses in historic British census data: An open methodology, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 58:1, 31-53, DOI: 10.1080/01615440.2024.2431491

To link to this article: https://doi.org/10.1080/01615440.2024.2431491

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
2025, VOL. 58, NO. 1, 31–53
https://doi.org/10.1080/01615440.2024.2431491

Geo-coding addresses in historic British census data: An open methodology

Joshua Rhodes[a,b]

[a]History, Durham University, UK; [b]The Alan Turing Institute, London, UK

ABSTRACT

This article introduces *CensusGeocoder*, an open-source python package for automated address geo-coding, and *AddressGB*, an openly available dataset of c. 121 million individuals with geo-coded addresses from I-CeM (digitized historic British census data for 1851 to 1911). *AddressGB* enables address and street-level GIS analysis of I-CeM data, a much higher spatial resolution than existing approaches that aggregate individuals to parishes or census registration sub-districts. This also opens new opportunities for linking historic census data to other spatial datasets. The article details the methodology underpinning *CensusGeocoder* and showcases the scale, accuracy, representativeness, and future applications of *AddressGB*.

KEYWORDS

Census; geo-code; historical GIS; Britain

Introduction

Digitized individual-level historic census data have transformed quantitative research on nineteenth- and early-twentieth-century Britain. Previously, historians undertaking large-scale analysis relied on contemporary printed census reports, which recorded aggregated demographic data such as age, gender, and occupation by county (Lawton 1978; Marsh 1965). But in 2014, Schürer and Higgs (2014, 2015) released I-CeM (Integrated Census Microdata) – an individual-level, standardized dataset of over 180 million people enumerated in British censuses between 1851 and 1911. In 2024, a revised version of this dataset including 1921 census data was released (Schürer, Higgs, and FINDMYPAST LIMITED 2024; Schürer, Wakelam, and FINDMYPAST LIMITED 2024a, 2024b). I-CeM contains transcriptions and data extracted from the original census records, including (among many other variables), people's names, addresses, ages, marital status, occupations, and place of birth. Researchers are therefore no longer limited to the summary statistics in the contemporary printed census reports and can now work with census data from the individual or household up to national level.

Concurrently, boundaries of key historic census administrative units – parishes, registration sub-districts (RSDs), registration districts (RDs), and registration counties – have been digitized, geo-referenced, and linked to I-CeM (Day, et al. 2016; Roughley and Anderson 2019; Satchell et al. 2017, 2018; Southall et al. 2022).[1] This has enabled researchers to analyze census data dynamically over time at different levels of spatial aggregation. This has spawned a wave of historical geographies of occupational structure, demography, and economic development based on I-CeM (Bennett et al. 2019; Bogart et al. 2022; Day 2020; Jaadla et al. 2020; Philips et al. 2022; Reid et al. 2018; Smith, Bennett, and van Lieshout 2022).

However, there are important reasons – and significant scope – to move spatial analysis beyond these administrative units. Aggregating census data to these units precludes and obfuscates certain research questions, since the areas they delineate are arbitrary and often bear no relationship to meaningful groupings of people or communities. In a US context, for example, increasing the resolution at which individuals are geo-located in historic census data beyond wards or census tracts has been central to new research on highly localized, neighborhood-level racial segregation (Shertzer, Walsh, and Logan 2016; Notter and Logan 2022). It has also enabled the creation of important interactive public history tools, such as the historic New York Digital Atlas (Baics et al. 2021).

Aggregation also renders relationships with other geo-spatial datasets imprecise because distances must

CONTACT Joshua Rhodes Joshua.m.rhodes@durham.ac.uk History, Durham University, UK

be measured from polygon centroids rather than from where individuals lived (Bogart et al. 2022). This encourages the use of I-CeM as primarily a tool for macro-economic analysis. The smallest, most common unit for these types of analyses – Registration Sub-Districts (RSDs) – are large aggregations that flatten important differences between localities and neighborhoods. The median population of an RSD in 1851 was 6,096.5 and in 1911 this had risen to 9,159. The median area remained approximately 6,523 hectares.[2] In contrast, the smallest geographical units used in modern British censuses, known as Output Areas, contain no more than 625 people and on average cover an area of approximately 100 hectares (Office for National Statistics).

The prospect of greater spatial resolution and a more meaningful unit of analysis is offered by the buildings, streets, and places recorded in the census addresses stored in I-CeM. The challenge is translating these addresses to their real-world locations – a process known as geo-coding or geo-locating. Two different methods for geo-coding I-CeM addresses have previously been outlined by Walford (2019) and Lan and Longley (2019, 2021). Walford (2019) demonstrated the potential of a semi-automated approach by geocoding just under 80,000 addresses for 260,000 individuals from six case-study areas in London and Middlesex in 1901 and 1911. He automatically matched 37% and 45% of his 1901 and 1911 sample to a database of modern addresses. The remainder he geo-located manually by identifying the location of the historic addresses in secondary sources and creating new point data at those locations. This approach geo-coded 100% of addresses and the linking was highly accurate because the majority of addresses were manually geo-located. However, the time-consuming manual process makes this method suitable for targeted case-studies but not as a scalable way to geo-code census addresses across Britain.

Lan and Longley (2019, 2021) developed an automated method capable of geo-coding I-CeM addresses across Britain from the 1881, 1891, and 1901 censuses. They matched historic addresses to a modern address database, matching street names and (where possible) individual house numbers. If no match could be made to a modern address, they attempted to match to a gazetteer of late nineteenth- and early twentieth-century placenames and streets (GB1900 GAZETTEER). With this dual approach, Lan and Longley geo-located the addresses of 66% of the British population in 1881, 73% in 1891, and 77% in 1901. This was a major step forward in geo-locating individuals nationally but has two key limitations. The first is that the geo-coded address data and the code used to generate it have not been made openly available, which means the method cannot be reproduced nor the data re-used. Secondly, the accuracy of their geo-coding method is unclear because details of the validation process (if any) were not reported.

In response to these limitations, this article introduces *CensusGeocoder* (Rhodes 2024a), an open-source python package for automated address geocoding, and *AddressGB* (Rhodes 2024b), an openly available dataset of geo-coded addresses from the 1851 to 1911 censuses.[3] *CensusGeocoder* offers several major advantages over existing methods of geo-coding I-CeM addresses. It is highly configurable. By default, *CensusGeocoder* links census addresses to modern road line vector data (OS Open Roads) and historic placename point data (GB1900 GAZETTEER). These are the best alternatives in lieu of historically accurate comprehensive GIS datasets of nineteenth- or early twentieth-century addresses. New opportunities presented by machine learning and computer vision for extracting street geometries and labels from historic maps at scale will hopefully soon create these much-needed datasets (Li et al. 2020; Jiao, Heitzler, and Hurni 2021; Kim et al. 2023). Users can use new GIS datasets (as and when they are created) to geo-code addresses with no changes to the code base required. Lastly, making the tool and resulting dataset openly available ensures reproducibility and enables continued community development. This raises the prospect of improving *CensusGeocoder*'s accuracy and coverage of British census data, as well as its application to censuses in other countries.

AddressGB radically lowers barriers to working with street-level historic British census data. Currently, researchers must use a highly restricted 'Names and Addresses' supplement to I-CeM (Schürer and Higgs 2020) that requires a special license from the UK Data Service. Most researchers only have access to the anonymized version of I-CeM (Schürer, Higgs, and FINDMYPAST LIMITED 2024), which contains the census variables for each individual except their name and address.[4] *AddressGB* allows users of the anonymized I-CeM dataset to map census data at address level without releasing any restricted data that is part of the 'Names and Addresses' I-CeM. In doing so, it opens new possibilities for researchers without extensive institutional support to work with census data at address level, for example making viable new Masters and PhD projects, which might not have been feasible owing to existing access restrictions.

The remainder of this article is structured as follows: firstly, the sources and methodology underpinning *CensusGeocoder* are set out. The article then

introduces *CensusGeocoder*'s first output, *AddressGB,* reporting the number of addresses geo-coded in each census to GB1900 and OS Open Roads respectively. The accuracy of the geo-coding is then evaluated against a manually geo-coded sample of addresses (Rhodes 2024c) before considering its representativeness geographically and on key socio-economic and demographic indicators. The article concludes by exploring some of the new research possibilities that *AddressGB* offers.

Sources and method

Figure 1 provides an overview of the *CensusGeocoder* pipeline. It takes three categories of input data: 1) census addresses to be geo-coded, 2) GIS administrative boundary data, which provide a bridge between the census addresses, and 3) target geometry data containing geo-coded addresses to which the census addresses will be linked. After pre-processing (discussed below), the GIS boundary dataset of historic administrative census units is linked to the census addresses using existing lookup tables. The same GIS boundary dataset is then linked to the target geometry dataset using a spatial join to identify which unit each entity of the target geometry dataset belongs.

Matches are then made between the census addresses and geo-coded addresses in the target geometry dataset in a two-stage process of 1) geo-blocking and 2) fuzzy string matching. The first stage identifies addresses in the census data and the target geometry dataset within the same geo-blocking unit. This differentiates streets with the same name in different places, preventing a High Street in London matching a High Street in Edinburgh. The second stage compares these addresses using fuzzy string matching to determine the most likely match. Census addresses with a match above a specified quality threshold are geo-located to the location

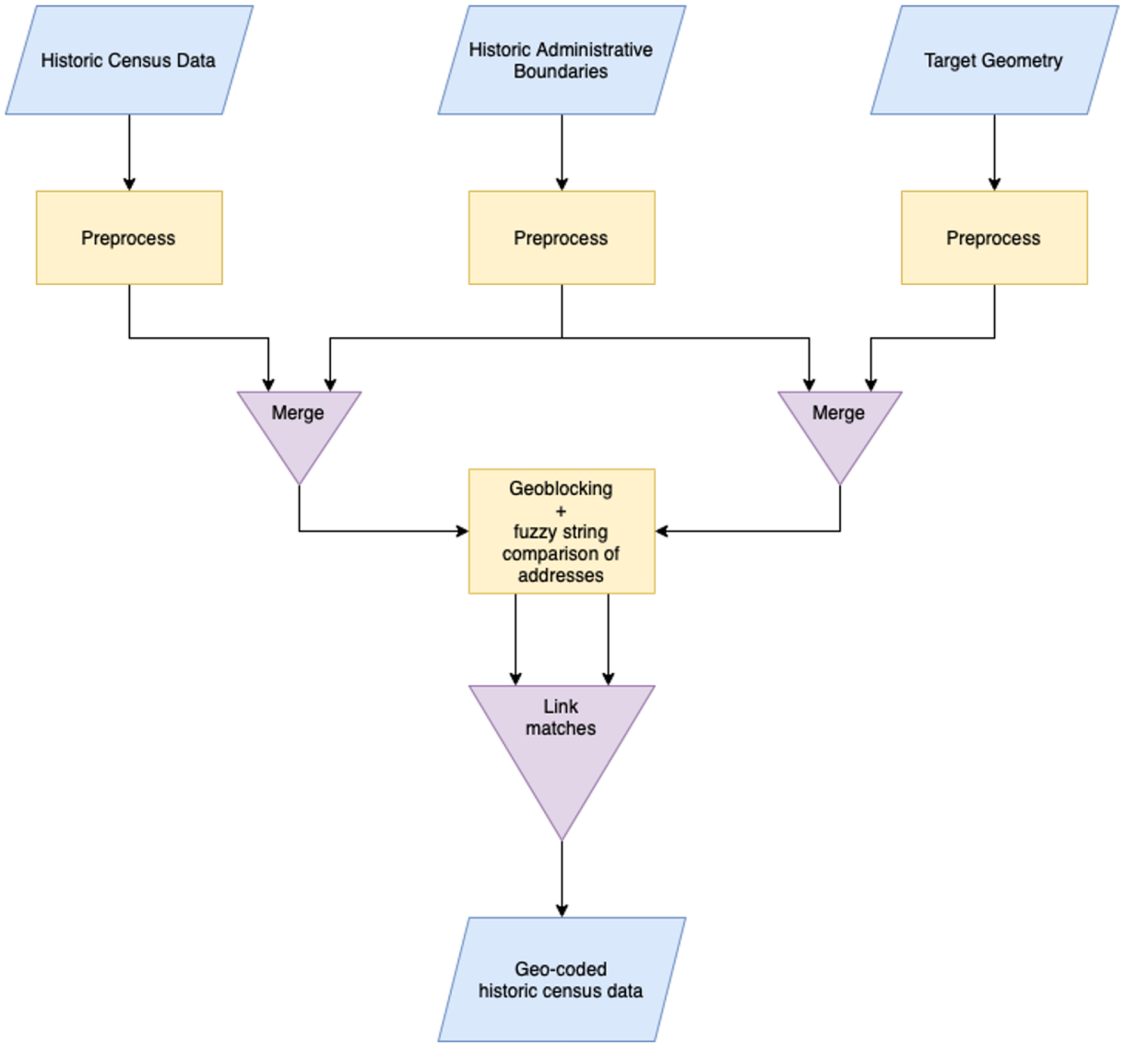

Figure 1. Flowchart of *CensusGeocoder* pipeline.

of their matching entity in the target geometry. The remainder of this section outlines these sources and processes in greater detail.

I-CeM

Table 1 shows a sample from I-CeM of individuals living on Boundary Lane in Manchester, Lancashire in 1901 with variables used in the geo-coding pipeline. At 2 Boundary Lane is Widow Jane Ripley [RecID 22475961] and her lodger Ellen Hooney [RecID 22475962]. At 4 Boundary Lane live part of the Mellor family – husband and wife Charles and Selina Mellor [RecID 22475963 and 22475964] and their eldest son Arthur [RecID 22475965].[5] Each person has a unique identifier – 'RecID' – that links individuals' records between the 'Anonymized' I-CeM and the 'Names and Addresses' supplement.[6] The 'ParID' and 'ConParID' fields link each person to GIS boundary datasets of the census administrative units used for geo-blocking. The Ripley and Mellor households lived in South Manchester, which in 1901 has the unique numeric identifier 'ParID' 10873 within the consistent parish 'ConParID' 108139. Individuals in Scotland are assigned to Scottish GIS boundary datasets using the 'ParID' and 'RecID' fields.

The 'Address' field contains the addresses as transcribed from the original census records to be geo-coded. During pre-processing, addresses are converted to uppercase, cleaned, and standardized using regular expressions to increase the match rate to addresses in the target geometry dataset. Non-alphanumeric characters, such as full stops, commas, and hyphens are also removed. Non-Unicode characters, typically appearing in Welsh addresses are standardized so that, for example, 'Plâs-rhaiadr' becomes 'PLAS RHAIADR'. Common abbreviations like 'St' are expanded to 'STREET' depending on their position in the address, so 'St James' (Saint James) remains unchanged, but 'James St' becomes 'JAMES STREET'. House numbers are also removed from the addresses due to inconsistencies in numbering and recording practices in the original census returns and because there is no straightforward relationship between these and modern house numbering (Higgs 1991). The Ripley and Mellor families at 2 and 4 Boundary Lane are therefore both geo-coded based on their street 'Boundary Lane'.

The global impact of these corrections on the number of unique addresses is shown in Table 2. For most censuses, these corrections substantially reduce the large number of unique addresses to be geo-coded. On average, cleaning and standardizing addresses reduced the number of unique addresses by half. The 1861 Scottish census and 1891 England and Wales census are exceptions due to the limited number of individual building or property addresses recorded in the transcriptions in I-CeM for those years.[7]

Table 1. I-CeM sample data (England and Wales 1901) used in geocoding pipeline.

	Names and addresses I-CeM	Anonymized I-CeM	
RecID	Address	ConParID	ParID
22475961	2 BOUNDARY LANE	108139	10873
22475962	2 BOUNDARY LANE	108139	10873
22475963	4 BOUNDARY LANE	108139	10873
22475964	4 BOUNDARY LANE	108139	10873
22475965	4 BOUNDARY LANE	108139	10873

Historic GIS boundary data

Due to differences in extant GIS boundary datasets, *CensusGeocoder* uses different types of boundaries for the Scottish censuses than it does for England and Wales. A combination of registration sub-districts (RSDs) and consistent parishes are used for England and Wales censuses, whereas historic parishes and urban sub-divisions are used for Scotland.[8]

There are two GIS boundary datasets for each England and Wales census between 1851 and 1911. The first are 'consistent parish' units.[9] Owing to the scale of parish boundary changes between 1851 and 1911, there are two series of 'consistent' units: one for 1851–1891 and another for 1901–1911, which are stored in the ConParID field in I-CeM. The appropriate boundaries for each census year can be generated by linking a GIS of parish boundaries in 1851 (Satchell et al. 2017) to a lookup table with the two series of 'ConParID': 'conparid_51-91' covering 1851

Table 2. Unique and standardized I-CeM addresses, Britain 1851–1911.

		Addresses		
		All	Cleaned and standardized	
Year	Country	*N*	*N*	%
1851	E&W	1,181,171	499,877	42.32
	Scotland	155,143	95,067	61.28
	GB	*1,336,314*	*594,944*	*44.52*
1861	E&W	1,631,816	587,698	36.01
	Scotland	112,778	109,125	96.76
	GB	*1,744,594*	*696,823*	*39.94*
1871	Scotland	236,590	118,995	50.30
1881	E&W	3,198,103	879,459	27.50
	Scotland	296,732	137,888	46.47
	GB	*3,494,835*	*1,017,347*	*29.11*
1891	E&W	900,097	891,104	99.00
	Scotland	325,583	139,048	42.71
	GB	*1,225,680*	*1,030,152*	*84.05*
1901	E&W	4,954,422	978,240	19.74
	Scotland	385,922	157,204	40.73
	GB	*5,340,344*	*1,135,444*	*21.26*
1911	E&W	6,293,900	1,492,063	23.71

to 1891, and 'conparid_01-11' for 1901–1911. Once joined, the 1851 parish geometries can then be dissolved on one of these fields to create a consistent geographic unit for the relevant period. The second GIS boundary dataset comprises geometries of RSD boundaries for 1851–1911. Each RSD has a unique identifier for each census year, e.g. 'CEN_1901' that links to a corresponding field in I-CeM.

For successful geo-blocking, these two GIS datasets must be used together. By accommodating boundary changes over time, the consistent parish geographies become too large in some areas to provide meaningful limitations on possible address comparisons on their own. Figures 2a and 2b show large areas to the south of Manchester and London have respectively been amalgamated into one consistent parish by 1901. There are many streets with the same name in these areas and without further subdivision, addresses could be geo-coded incorrectly to similarly named addresses in other locations. But as Figures 2a and 2b also show, RSDs sub-divide these consistent parishes into smaller, more effective geo-blocking units. Elsewhere (typically rural areas), consistent parishes can be smaller than RSDs, as Figure 3 shows. Therefore, the boundary datasets need to be combined to create the smallest possible geo-blocking units by splitting RSDs on consistent parish boundaries and vice versa.

For Scotland, there are two historic parish boundary datasets publicly available from the National Records of Scotland (NRS). These are pre-1891 parishes (Roughley and Anderson 2019) and post-1891 civil parishes (NRS, Civil Parishes). However, at the time of writing, there were no publicly available lookup tables linking these GIS files to parishes in I-CeM. These have been created manually and are now available openly (Rhodes 2024d). Separate lookup tables for each census year have been produced, which link the unique parish identifier in the GIS files to the 'ParID' field in I-CeM. Scottish censuses between 1851 and 1881 have been linked to the pre-1891 parish boundary file and the 1891 and 1901 censuses to the post-1891 parish boundary file. Key urban centers are subsumed in single parishes which increases the likelihood of geo-coding errors. To address this issue, subdivisions of five Scottish cities (Satchell 2023) have been used to provide small geo-blocking units in these cities.

Target geometry datasets

This section describes the target geometry datasets used to generate *AddressGB*. The first is GB1900, a point geometry dataset of late-nineteenth- and early twentieth-century place names and streets. The second is OS Open Roads, a line geometry dataset of the modern British road network. They have been chosen for their permissive re-use licenses and because used in combination they maximize the number of addresses that can be geo-coded. It should be noted that *CensusGeocoder* works with any other target geometry datasets provided the appropriate parameters are specified.

GB1900

GB1900 is a dataset of 2.6 million map text labels and coordinates transcribed and extracted by a crowdsourcing project from Ordnance Survey's 2nd Edition County Series Six Inch to One Mile maps of Great Britain. The maps were published between 1888 and 1914, making them broadly contemporaneous with nineteenth- and early twentieth-century census data. The labels include street names and a wide variety of other features marked on the maps, including towns and cities (Manchester, Exmouth, etc.), buildings (e.g. gas works, factory), and footpaths (F.P.). Particularly in rural areas, individual properties such as farmsteads and rows of cottages could also be marked on the maps, and therefore feature in GB1900.

The scale of the Six Inch series means that not every street is labeled. Figure 4 shows Hack Street and Lower Trinity Street labeled on the more detailed Twenty-Five Inch to One Mile OS maps but not on the Six Inch series. Usually, the streets not labeled on the 6 inch maps were smaller back roads, lanes, and side streets in dense urban areas. Using GB1900 therefore potentially introduces a bias toward wealthier individuals who tended not to live on these streets in these areas.

GB1900 also captures much less spatial detail than modern road network datasets. The location of each feature in GB1900 are coordinates of the map text label (typically the lower, left-hand corner of the label). Not only do the coordinates refer to part of the label rather than the feature itself, but all features are represented by a single point. This is less problematic for individual buildings but means the spatial footprint of a street is not fully represented. Nevertheless, GB1900 indicates a street's approximate location within a parish or RSD – a major improvement beyond allocating all individuals to one unit without knowing where they lived within it. It is also the only extant source of geo-located British historic placename and street name data contemporaneous with historic censuses.

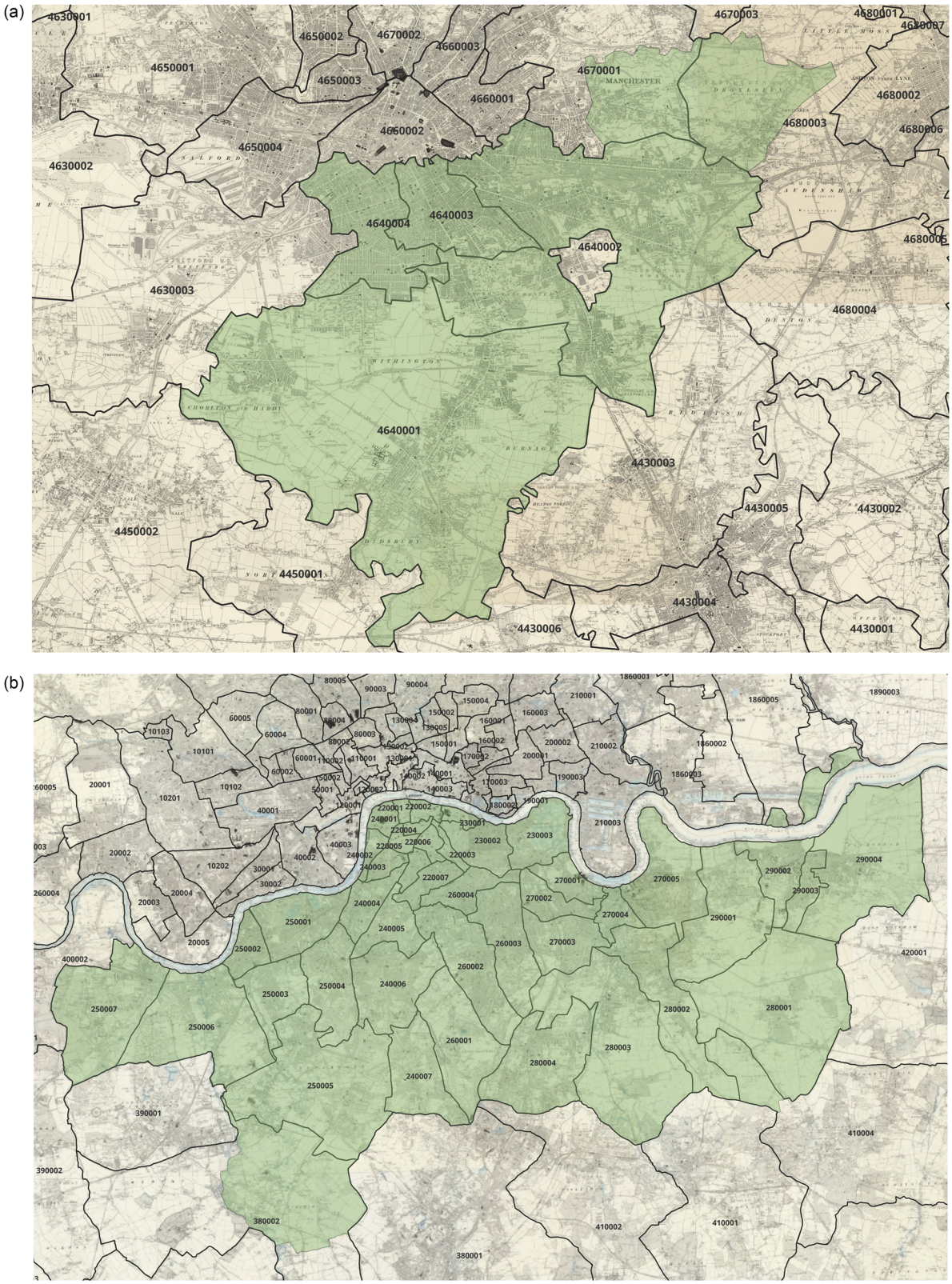

Figure 2. (a) Consistent parish boundaries and RSDs in South Manchester in 1901. [consistent parish boundaries in Green]. (b) Consistent parish boundaries and RSDs in South London in 1901. [consistent parish boundaries in Green]. Reproduced with the permission of the National Library of Scotland.

To prepare GB1900 for linking to census addresses, the historic administrative unit that each point resides within needs to be identified. A spatial join between GB1900 and the geo-blocking unit identifies intersections

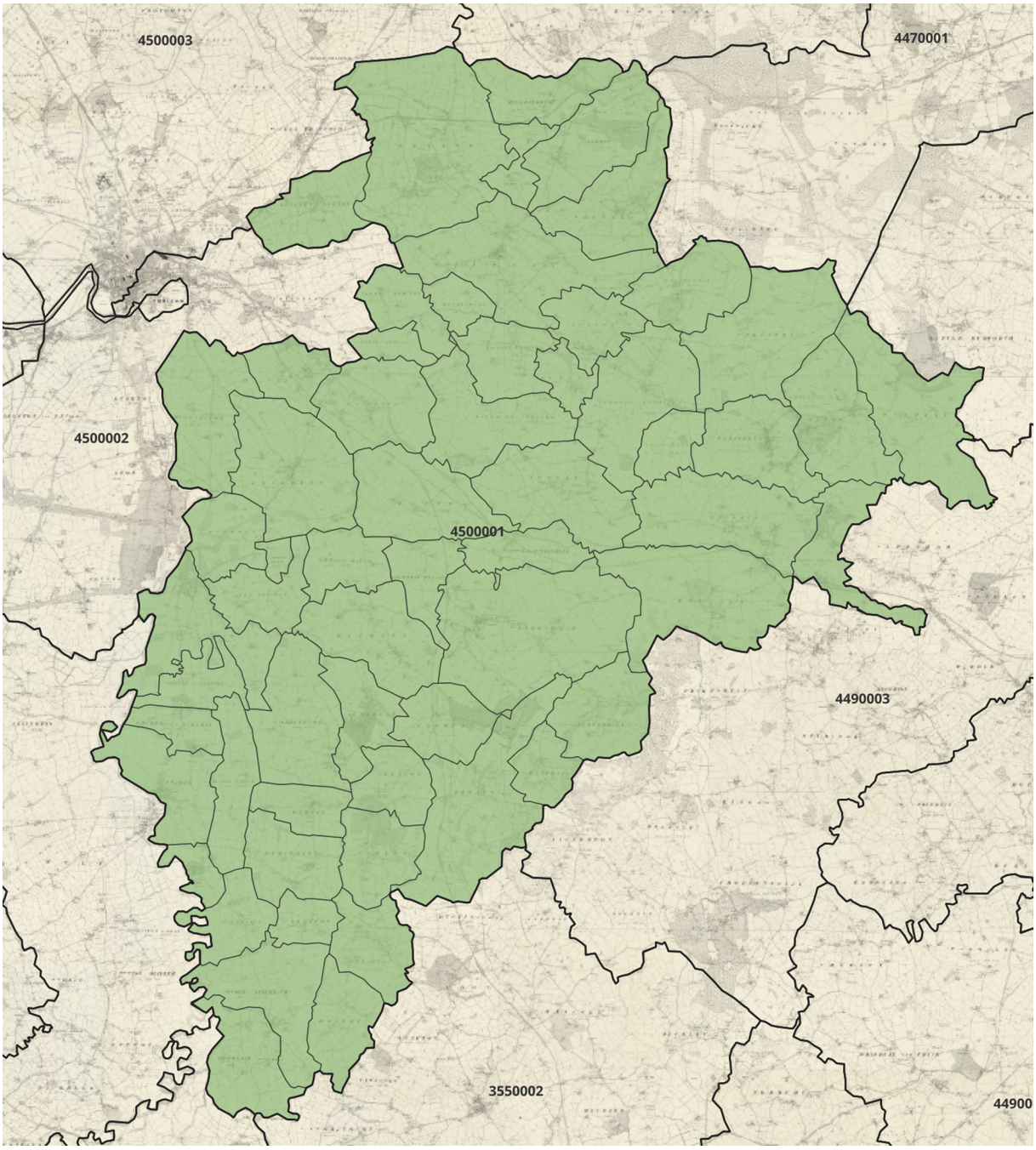

Figure 3. Consistent parish boundaries and RSDs near chester, England in 1901. [consistent parish boundaries in Green]. Reproduced with the permission of the National Library of Scotland.

between the two datasets, and each point in GB1900 is associated with the unique identifier for that historic administrative unit. A series of pre-processing steps also improve the quality of final matches to census addresses. GB1900 labels are converted to uppercase to match census addresses in I-CeM and abbreviations for street and road, such as 'ST.' and 'RD' are expanded to 'STREET' and 'ROAD'. To prevent census addresses matching non-street or place features, it is also important to remove as many of these type of labels from GB1900 as possible prior to linking. These include labels marking common features, such as 'F.P.' (Footpath), 'P.H.' (Public House), and 'Chap.' (Chapel). These kinds of labels can be removed by identifying repeated labels in each geo-blocking unit.[10]

OS Open Roads

OS Open Roads is Ordnance Survey's line vector dataset of the modern British road network. It provides

Figure 4. Comparison of street names present on OS Six Inch (left) and Twenty Five Inch (right) maps. Reproduced with the permission of the National Library of Scotland.

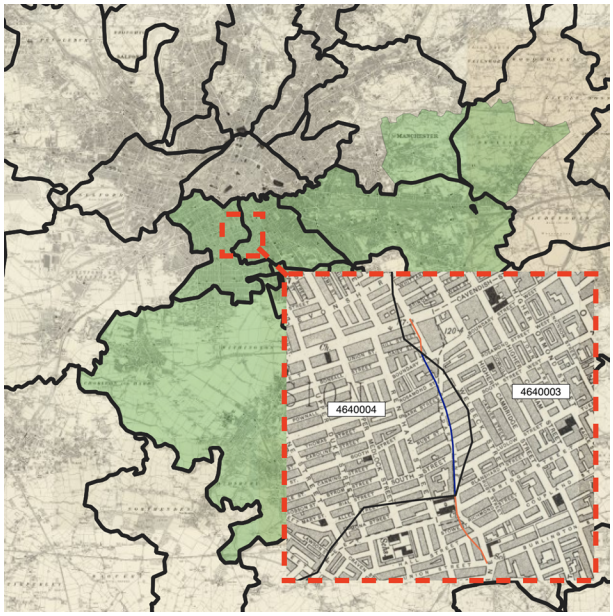

Figure 5. Street line vector allocation. Reproduced with the permission of the National Library of Scotland.

comprehensive coverage of the location and name of roads and is available on an Open Government License, which permits use 'in any way and for any purpose' (OS Open Roads). Its main limitation for use in historical research is that it captures the modern road network. Urban redevelopment and slum clearances have resulted in substantial changes in layout and road names since the nineteenth- and early twentieth-century censuses were taken. But there is often striking similarity between historic streets on nineteenth- and twentieth-century Ordnance Survey maps and the modern road network (Lan and Longley 2019, 2021). Supplementing OS Open Roads with historic street and place names from GB1900 plugs many of these gaps.

CensusGeocoder assigns streets in OS Open Roads (and other line vector datasets) to historic administrative units in a slightly different way to point data such as GB1900. Where line representations of streets fall wholly within one unit, they are assigned to that unit in the same way as point data. But many streets in OS Open Roads span multiple geo-blocking units. As Figure 5 shows, streets are segmented at each point they cross a boundary and each new street segment assigned to the unit that it sits within. In Manchester in 1901, Boundary Lane ran close to boundaries of two RSDs (CEN_1901: 4640003 and 4640004), which themselves reside within the large consistent parish area (ConParID 108139) shaded green. When assigning Boundary Lane to a geo-blocking unit, it is split into three segments because it crosses the boundary twice. The segment in blue is classified as residing in 4640004 and the two red segments are joined and treated as one segment classified as being in 4640003. Individuals returned in the census as living on Boundary Lane and residing in RSD 4640003 will then only be linked to the red segment.

Geo-blocking and fuzzy string matching

Once every census address and entity in the target geometry dataset (GB1900 or OS Open Roads) has been assigned to one of the geo-blocking units, *CensusGeocoder* isolates entries in the census and target geometry datasets from the same unit. To continue the example from South Manchester in 1901, there were 92,195 individuals living at 932 addresses within the unit [ConParID 108139 and CEN_1901 4640004] in I-CeM. Within this unit, there are 320 streets in

OS Open Roads and 278 entities in GB1900. When matching addresses in the South Manchester area, the search is limited to 298,240 (932 census addresses x 320 OS Open Road streets) pairs to compare between I-CeM and OS Open Roads and 259,096 (932 census addresses x 278 GB1900 entities) pairs between I-CeM and GB1900.

Geo-blocking limits the comparison of addresses between I-CeM and a target geometry dataset to bounded areas. This is an efficient and accurate way of geo-coding. It reduces the number of candidates passed to the computationally intensive fuzzy string matching algorithm. Between c.250,000 and c.300,000 address pairs to compare for South Manchester in 1901 may seem a large number but comparing the 320 South Manchester census addresses to every entry street in OS Open Roads would require making over 250 million comparisons.

Once geo-blocked, the similarity of each pair of addresses is computed using fuzzy string-matching algorithms. It is important to allow some degree of difference or 'fuzziness' in the comparison of addresses to account for spelling differences due to mis-transcription of the original census records or minor changes between historic and modern spellings. There are many different types of string-matching algorithm, and it is important to select the right one, or right combination of multiple algorithms, to achieve robust matches while maximizing the number of addresses that are linked.

Table 3 sets out the string comparison results for a series of census addresses and GB1900 entities calculated using different algorithms. These examples have been chosen to illustrate the types of matches that should be allowed and those that should not, as well as how the algorithm used for 1911 differs from other census years. From the 1901 England and Wales census, YORKE PLACE and THE GREYHOUND YORK PLACE should both link to the GB1900 entry YORK PLACE, and census address LONSDALE VILLAS UPPER LLOYD STREET should link to UPPER LLOYD STREET but not LLOYD STREET or NEW STREET. Lastly, from the 1911 census, GARLANDS FARM STEEPLE BUMPSTEAD should link to the GB1900 entity GARLANDS FARM but not the placename STEEPLE BUMPSTEAD (a village in Essex). For each possible address combination, Table 3 shows the normalized similarity scores for three algorithms – Levenshtein, Aligned, and Weighted Composite - expressing similarity as a value between 0 (very little or no similarity) and 1 (exact match).

Levenshtein calculates the minimum number of deletions, insertions, or substitutions needed to transform one string to another string. It works very well at identifying addresses that closely match but that might have some typographical differences. A comparison of YORKE PLACE and YORK PLACE achieves a high score of 0.91 because there is only one letter difference. But often it is necessary to link addresses when there are many alternations needed to transform one string to another. Census addresses often include additional details that do not appear in target geometry datasets. For example, 'THE GREYHOUND YORK PLACE' appears in the census but only 'YORK PLACE' appears in OS Open Roads and GB1900. We want to be able to link this census address to York Place even if we cannot specify the Greyhound Inn's precise location in York Place. But because of the number of insertions required by 'THE GREYHOUND', a comparison of these two addresses receives a very low Levenshtein score of 0.42. The same is true for the comparisons with LONSDALE VILLAS UPPER LLOYD STREET. We therefore need an algorithm that can locate a shorter string within a longer string to identify YORK PLACE within THE GREYHOUND YORK PLACE. The Aligned algorithm achieves this, indicating an exact match between THE GREYHOUND YORK PLACE and YORK PLACE, and between LONSDALE VILLAS UPPER LLOYD STREET and UPPER LLOYD STREET. But it also returns an exact match for LLOYD STREET and a high match for NEW STREET.

CensusGeocoder therefore uses a Weighted Composite algorithm that combines the scores from a range of algorithms including Levenshtein and Aligned to reflect similarity overall and between sub-strings. The addresses for which Aligned returned

Table 3. Comparison of string comparison algorithms.

I-CeM address	GB1900	Levenshtein	Aligned	Weighted composite	Alignment length	Final score	Alignment rank	Final score (1911)
YORKE PLACE	YORK PLACE	0.91	0.90	0.95	10	9.52	–	–
THE GREYHOUND YORK PLACE		0.42	1.00	0.90	10	9.00	–	–
LONSDALE VILLAS UPPER LLOYD STREET	UPPER LLOYD STREET	0.53	1.00	0.90	18	16.20	–	–
	LLOYD STREET	0.35	1.00	0.90	12	10.80	–	–
	NEW STREET	0.26	0.82	0.86	10	8.55	–	–
GARLANDS FARM STEEPLE BUMPSTEAD	GARLANDS FARM	0.42	1.00	0.9	13	11.70	2	23.4
	STEEPLE BUMPSTEAD	0.55	1.00	0.9	17	15.30	1	15.3

an exact match are now 0.9, and the single edit to transform YORKE PLACE and YORK PLACE is now reflected in a higher weighted composite score of 0.95 than the Aligned score of 0.9. NEW STREET appears to score highly (0.86) but many address comparisons will result in scores between 0.8 and 0.89 using the Weighted Composite algorithm. In this context, only scores of 0.9 or over are high quality matches (in contrast to Levenshtein, where a score of 0.8 may indicate high similarity). Address pairs with similarity scores less than 0.9 are considered too dissimilar for them to be true matches and are discarded. Address pairs with a similarity score that equals or exceeds this threshold are considered sufficiently similar to be potential matches and are kept for further processing.

Matching accuracy is improved further by using the alignment length to differentiate between comparisons with the same Weighted Composite score. LLOYD STREET and UPPER LLOYD STREET match equally well (0.9) to LONSDALE VILLAS UPPER LLOYD STREET but the lengths of the aligned sub-strings are different. In the case of LLOYD STREET, this is 10 characters, whereas UPPER LLOYD STREET is 18 characters. Multiplying the Weighted Composite similarity score by the alignment length produces a final score, by which measure UPPER LLOYD STREET is now the highest scoring comparator. This method improves matches to streets with common geographic modifiers, such as 'upper', 'lower', 'east' and 'west'.

A modified version of this algorithm is used for the 1911 census because the information recorded in the address field contains parishes and places as well as buildings and streets. Prior to 1911, householders filled out schedules that were then copied into books by census enumerators. The original household schedules were destroyed, so the census enumerator books are the documents on which the digitized census data in I-CeM is based. For 1911, the census process changed, and the census returns were filled out by householders themselves. This introduced more variation in the responses recorded in the address field, with self-reported addresses in 1911 around 50% longer on average than those supplied by enumerators in 1901.[11] Extra details often included parish, town, or city of residence, as the inclusion of STEEPLE BUMPSTEAD in Table 3 illustrates. This introduces two new possible issues. Firstly, these longer addresses are less similar (in terms of string comparison) to the street names in the target geometry datasets (and therefore less likely to reach the 0.9 threshold). This problem is particularly acute in London, where addresses might include the street name, parish (e.g. Paddington), and 'London'. Secondly, the inclusion of place names (parishes, towns, and cities) in the address field introduces an additional source of error when linking to GB1900, which itself contains place-names as well as street names. GB1900 has entries for Garlands Farm and Steeple Bumpstead because they were both labeled on the Six Inch Maps. This issue can largely be dealt with by ranking matches by the order they appear in census addresses on the assumption that specific information (buildings, streets) is listed before more general location information (parishes, towns, and cities). The final score for 1911 is therefore calculated by multiplying the weighted composite score by the alignment length and alignment ranking.

Having identified the highest scoring target geometry entity/entities for each census address, a final match is determined when there is only one address pair with the highest similarity score for that census address. Multiple address pairs with the same highest score are discounted because it is currently not possible to differentiate between these matches of equal quality.[12]

Results

In all, *CensusGeocoder* geo-codes the addresses of 121 million people (67%) across the 1851–1911 censuses. Matches increased over time – around 50 per cent of addresses and people were geo-coded in 1851 rising to 80 per cent in 1911 (see Figures 6 and 7). Broadly, this reflects the closer approximation of the modern road network in OS Open Roads and the creation of the OS maps underpinning GB1900 to censuses taken

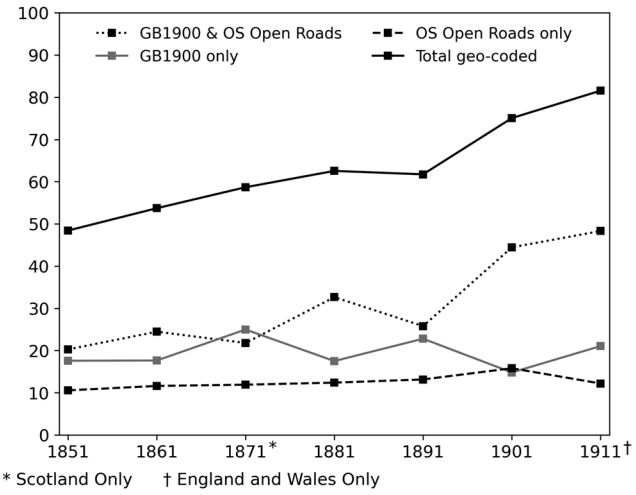

Figure 6. Percentage of British addresses geo-coded, 1851–1911.
Note: 'Total geo-coded' includes addresses linked to at least one of the target geometry datasets and is the sum of 'GB1900 only', 'OS Open Roads only', and 'GB1900 and OS Open Roads'.

around the turn of the twentieth century. The balance between GB1900 and OS Open Roads as sources of links varies though, largely reflecting differences in urbanization over time and place. Individuals/addresses matched to only GB1900 or only OS Open Roads are classified 'GB1900 only' or 'OS Open Roads only' as appropriate. Those matched to an entity in GB1900 and an entity in OS Open Roads are classified as 'GB1900 and OS Open Roads'. The total geo-coded population is the sum of these three categories.

People living in urban areas tended to report their place of residence on a street, which can be linked to either GB1900 or OS Open Roads. In contrast, those living in rural locations were much more likely to give their address as a particular farm or cottage, the name of the village (to indicate that they lived in the village proper and not the surrounding countryside), or a certain area within the parish. These locations are not found in OS Open Roads but were often (but by no means always) marked on Ordnance Survey maps and are therefore in the GB1900 dataset. The higher proportions of addresses in Scotland and in the earlier census years where only a match to GB1900 can be made reflect differing levels of urbanization.

Proportions of the population geo-coded in 1881 (61%), 1891 (69%), and 1901 (72%) (see Figure 7) are broadly similar but marginally lower than Lan and Longley's (66%, 73%, and 77%). The slightly lower linking rate of *AddressGB* may be attributed to differences in string-matching methods and matching thresholds. But the biggest factor is most likely to be the smaller geo-blocking units used by *CensusGeocoder*, which – crucially – results in *fewer but higher quality matches* since larger units increase

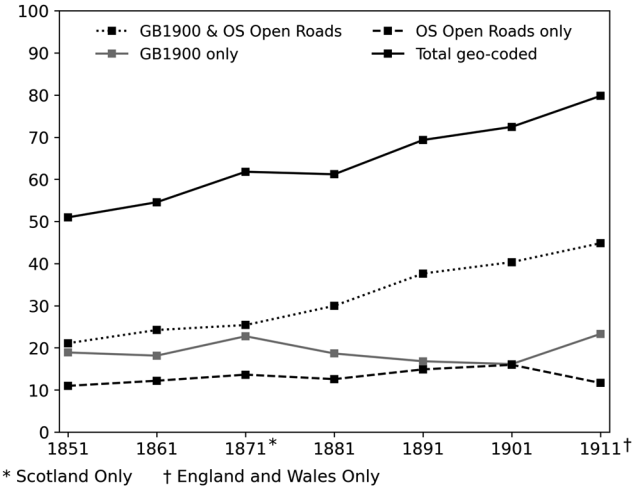

Figure 7. Percentage of British population geo-coded, 1851–1911.
Note: 'Total geo-coded' includes addresses linked to at least one of the target geometry datasets and is the sum of 'GB1900 only', 'OS Open Roads only', and 'GB1900 and OS Open Roads'.

the chances of matching incorrectly to streets or places with the same or similar names. With no comparable validation metrics, it is not possible to comment definitively on differences in accuracy between the two datasets. Releasing *AddressGB* and the manually geo-coded evaluation dataset openly ensures others will be able to benchmark their results against those presented here.

Evaluation

To validate the automated matches produced by *CensusGeocoder* a random sample of 7,200 addresses (1000 from each England and Wales (E&W) census, 200 from each Scottish census) have been independently, manually geo-coded (Rhodes 2024c). The sample reflects population densities across Great Britain: Lancashire, Yorkshire, and London account for about a third of the E&W sample each year, and Glasgow (Lanarkshire), Aberdeenshire, and Edinburgh make up around 50% of the Scottish sample each year. This sample was drawn from raw I-CeM addresses, with no pre-filtering or pre-processing of the address strings. The validation sample contains addresses of a similar length (16.4 characters) to those in the full I-CeM dataset (16.3 characters). About 5 per cent of the sample addresses were blank, truncated, or garbled, or were generic terms, such as 'cottage' or 'house'. Including these in the manually linked sample was important to ensure that *CensusGeocoder* appropriately handles these types of addresses.

The manual and automated outputs are compared using a standard typology for measuring the performance of automated algorithms. The result of the geo-coding process for each address is either 'positive' (geo-coded) or 'negative' (not geo-coded). Comparing them with the manual results enables us to classify these as 'True' (same as manual match) or 'False' (differs from manual match). A 'True Positive' (TP) indicates that *CensusGeocoder* has successfully geo-coded an address to the correct entity in GB1900 or OS Open Roads. A 'False Positive' (FP) indicates that *CensusGeocoder* has geo-coded an address but to the wrong entity. These types of errors are the most serious for a geo-coding application to make since they result in addresses being located in the wrong place, rendering the dataset unreliable for spatial analysis. Instances where both *CensusGeocoder* and a manual search have not geo-coded an address are 'True Negatives' (TN). Cases when *CensusGeocoder* fails to geo-code an address which has been successfully manually geo-coded are labeled 'False Negatives' (FN),

since *CensusGeocoder* should have returned a match in these instances. These errors are far less problematic than false positives but may affect how representative the geo-coded subset is of the whole population.

Tables 4 and 5 show the proportion of TP, FP, TN, and FN links to GB1900 and OS Open Roads for each census year. These show a high rate of true positives and negatives on average across all censuses for GB1900 (around 82%) and OS Open Roads (around 85%). Importantly, the number of false positive results is very low – less than 5% of addresses were geo-coded to the wrong GB1900 or OS Open Roads entity. Overall, the median distance of false positives to the correct address was 842 meters for GB1900 and 473 meters for OS Open Roads. The false positive rate for England and Wales 1911 addresses linked to GB1900 is slightly higher (7.4%). This higher false positive rate is due to matching some addresses to place labels in GB1900 and not building or street labels. Without these errors, the false positive rate would be 2.9 per cent. The algorithm employed for 1911 minimizes but does not prevent this issue completely. Geo-coding the 1911 E&W census using OS Open Roads remains highly accurate (with a false positive rate of just 1.8%) because it only contains streets and is therefore not subject to the errors found in some matches to placenames when matching to GB1900. The recommended use of *AddressGB* – to use OS Open Roads where possible, infilling gaps with additional matches to GB1900 – largely mitigates this issue.

The most common errors were false negatives - addresses that were successfully manually geo-coded but not geo-coded by *CensusGeocoder*. Around 11% of addresses were missing a possible link to OS Open Roads and around 14% to GB1900. Reasons for

Table 4. Manual evaluation statistics for geo-coding of GB1900.

Year	Country	TP (%)	TN (%)	FP (%)	FN (%)	Precision	Recall	f1	FP distance (m)
1851	E&W	344 (34.4)	460 (46)	38 (3.8)	158 (15.8)	0.90	0.69	0.78	794
	Scotland	81 (40.5)	77 (38.5)	7 (3.5)	35 (17.5)	0.94	0.69	0.80	235
	GB	*425 (35.4)*	*537 (44.8)*	*45 (3.8)*	*193 (16.1)*	*0.91*	*0.69*	*0.78*	
1861	E&W	400 (40)	437 (43.7)	39 (3.9)	124 (12.4)	0.92	0.76	0.83	751
	Scotland	82 (41)	80 (40)	3 (1.5)	35 (17.5)	0.96	0.70	0.81	
	GB	*482 (40.2)*	*517 (43.1)*	*42 (3.5)*	*159 (13.25)*	*0.92*	*0.75*	*0.83*	
1871	Scotland	98 (49)	77 (38.5)	4 (2)	21 (10.5)	0.95	0.82	0.88	513
1881	E&W	461 (46.1)	373 (37.3)	38 (3.8)	128 (12.8)	0.93	0.78	0.85	1048
	Scotland	90 (45)	68 (34)	7 (3.5)	35 (17.5)	0.93	0.72	0.81	258
	GB	*551 (45.9)*	*441 (36.8)*	*45 (3.8)*	*163 (13.6)*	*0.93*	*0.77*	*0.84*	
1891	E&W	496 (49.6)	340 (34)	36 (3.6)	128 (12.8)	0.93	0.79	0.86	568
	Scotland	92 (46)	65 (32.5)	9 (4.5)	34 (17)	0.93	0.73	0.82	316
	GB	*588 (49)*	*405 (33.8)*	*45 (3.8)*	*162 (13.5)*	*0.93*	*0.78*	*0.85*	
1901	E&W	533 (53.3)	299 (29.9)	37 (3.7)	131 (13.1)	0.94	0.80	0.87	883
	Scotland	106 (53)	60 (30)	5 (2.5)	29 (14.5)	0.95	0.78	0.86	
	GB	*639 (53.3)*	*359 (29.9)*	*42 (3.5)*	*160 (13.3)*	*0.94*	*0.80*	*0.87*	
1911	E&W	601 (60.1)	200 (20)	74 (7.4)	125 (12.5)	0.89	0.83	0.85	966

Note: FP distance is the median distance in meters between the correct geo-coded address and the false positive. Distances are calculated only for false positives where a manual match to GB1900 has been made (rather than a false positive being identified because a GB1900 entry has been linked to but manually the address has been found on the 25 inch maps).

Table 5. Manual evaluation statistics for geo-coding of OS Open Roads.

Year	Country	TP (%)	TN (%)	FP (%)	FN (%)	Precision	Recall	f1	FP distance (m)
1851	E&W	244 (24.4)	601 (58)	58 (5.8)	97 (9.7)	0.81	0.72	0.76	265
	Scotland	56 (28)	125 (62.5)	6 (3)	13 (6.5)	0.90	0.81	0.85	
	GB	*300 (25)*	*726 (60.5)*	*64 (5.3)*	*110 (9.2)*	*0.82*	*0.73*	*0.78*	
1861	E&W	305 (30.5)	550 (55)	51 (5.1)	94 (9.4)	0.86	0.76	0.81	1140
	Scotland	65 (32.5)	110 (55)	5 (2.5)	20 (10)	0.93	0.76	0.84	
	GB	*370 (30.8)*	*660 (55)*	*56 (4.7)*	*114 (9.5)*	*0.87*	*0.76*	*0.81*	
1871	Scotland	75 (37.5)	108 (54)	2 (1)	15 (7.5)	0.97	0.83	0.90	
1881	E&W	396 (39.6)	463 (46.3)	36 (3.6)	105 (10.5)	0.92	0.79	0.85	734
	Scotland	72 (36)	99 (49.5)	8 (4)	21 (10.5)	0.90	0.77	0.83	850
	GB	*468 (39)*	*562 (46.8)*	*44 (3.7)*	*126 (10.5)*	*0.91*	*0.79*	*0.85*	
1891	E&W	487 (48.7)	367 (36.7)	38 (3.8)	108 (10.8)	0.93	0.82	0.87	666
	Scotland	68 (34)	97 (48.5)	9 (4.5)	26 (13)	0.88	0.72	0.80	75
	GB	*555 (46.3)*	*464 (38.7)*	*47 (3.9)*	*134 (11.2)*	*0.92*	*0.81*	*0.86*	
1901	E&W	524 (52.4)	316 (31.6)	47 (4.7)	113 (11.3)	0.92	0.82	0.87	659
	Scotland	101 (50.5)	83 (41.5)	3 (1.5)	13 (6.5)	0.97	0.89	0.93	
	GB	*625 (52)*	*399 (33.3)*	*50 (4.2)*	*126 (10.5)*	*0.93*	*0.83*	*0.88*	
1911	E&W	527 (52.7)	311 (31.1)	18 (1.8)	144 (14.3)	0.97	0.79	0.87	424

Note: see Table 4.

addresses not being geo-coded were varied. In some cases, streets changed names, and only reference to multiple historic map series enabled the correct street to be located. In others, possible matches could be confirmed by locating adjacent addresses in the census records on the map. Another cause of false negatives was entities in the target geometry datasets or I-CeM addresses not being allocated to the correct geo-blocking units. Some GB1900 points for roads were assigned to the adjacent geo-blocking unit to which the majority of the road belonged, because the map label lay outside the boundary. False negatives might also be the result of inaccuracies in the boundary GIS datasets and I-CeM. Boundaries could be in the wrong location and cause a street to be assigned to the incorrect geo-blocking unit. In other cases, individuals have been allocated to the wrong parish in I-CeM and therefore are not assigned to the correct geo-blocking unit.

The overall performance of *CensusGeocoder* can be evaluated using three metrics calculated from the TP, FP, TN, and FN. These assess *precision* (the ratio of correctly geo-coded addresses (TP) to the total number of addresses geo-coded by *CensusGeocoder* including addresses that were incorrectly geo-coded (TP + FP). *Recall* is the ratio of correctly geo-coded addresses (TP) to the total number of addresses that should have been geo-coded (TP + FN). *Precision* reflects the accuracy of the geo-coded addresses returned by *CensusGeocoder*, while *recall* measures how good *CensusGeocoder* is at geo-coding addresses that were geo-codable. These two metrics capture the tradeoff in an automated process such as *CensusGeocoder* – which aims to maximize the number of addresses that are geo-coded while minimizing the number of these that are geo-coded incorrectly. A third metric, known as an *F1 score*, combines both *precision* and *recall* to express *CensusGeocoder*'s overall performance on a scale of 0 to 1 (Sammut and Webb 2010, 397).

CensusGeocoder achieves an average *F1 score* of 0.84 across all census years when linking to GB1900 and 0.85 when linking to OS Open Roads. The low rate of false positives is reflected in a high average precision score of 0.93 for GB1900. The average precision rate for OS Open Roads was 0.91 but slightly more varied, ranging from a low of 0.81 for the 1851 E&W census to 0.97 for the 1871 and 1901 Scottish censuses and the 1911 E&W census. Recall rates were lower across the board (0.78 on average for GB1900, 0.79 for OS OpenRoads), reflecting the higher number of addresses which *CensusGeocoder* should have geo-coded but did not. The average *F1 score* was similar for GB1900 and OS Open Roads (0.84 and 0.85 respectively). The lowest *F1 scores* were for the 1851 E&W census (0.78 for GB1900, 0.76 for OS Open Roads). The best overall performance for linking to GB1900 was achieved for the 1871 Scottish census (0.88) and the best for OS Open Roads was the 1901 Scottish census (0.93).

Tables 6 and 7 outline overall TP, TN, FP, and FN rates by gender, age, select occupational groups and locations for GB1900 and OS Open Roads. They give an indication of which people's addresses were more likely to be linked correctly or not. There were no substantial gendered differences in the types of links. Nor were there large differences across different age groups, though for both GB1900 and OS Open Roads, there were higher rates of False Negatives among older people, perhaps reflecting poorer linking in

Table 6. Manual evaluation statistics by age, gender, and select occupations and locations for GB1900.

		TP (%)	TN (%)	FP (%)	FN (%)
Gender	Female	1,700 (46.8)	1,258 (34.6)	146 (4)	528 (14.5)
	Male	1,678 (47.8)	1,232 (35.1)	150 (4.3)	452 (12.9)
Age	Under 20	1,467 (47.3)	1,088 (35.1)	132 (4.3)	415 (13.4)
	20 - 40	1,039 (46.7)	784 (35.2)	89 (4)	315 (14.1)
	40 - 60	652 (49.4)	438 (33.2)	59 (4.5)	170 (12.9)
	Over 60	226 (44.4)	186 (36.5)	17 (3.3)	80 (15.7)
Occupations	Domestic service	195 (48.6)	122 (30.4)	16 (4)	68 (17)
	Farming	201 (41.4)	162 (33.4)	20 (4.1)	102 (21)
	Manufacturing	370 (51.5)	224 (31.2)	40 (5.6)	84 (11.7)
	General labourers	68 (47.2)	56 (38.9)	11 (7.6)	9 (6.3)
E&W	Cornwall	35 (41.7)	31 (36.9)	2 (2.4)	16 (19)
	Norfolk	38 (36.2)	47 (44.8)	3 (2.9)	17 (16.2)
	Lancashire	439 (59.4)	158 (21.4)	43 (5.8)	99 (13.4)
	London	365 (47.4)	300 (39)	21 (2.7)	84 (10.9)
	West Riding Yorkshire	266 (54.8)	120 (24.7)	26 (5.4)	73 (15.1)
	Pembrokeshire	9 (36)	9 (36)	1 (4)	6 (24)
	Glamorganshire	39 (36.8)	43 (40.6)	5 (4.7)	19 (17.9)
Scotland	Argyll	7 (30.4)	7 (30.4)	1 (4.3)	8 (34.8)
	Edinburgh	49 (37.1)	73 (55.3)	2 (1.5)	8 (6.1)
	Lanark	153 (60.2)	61 (24)	6 (2.4)	34 (13.4)

Note: For source of occupational groupings, see footnote 14.

Table 7. Manual evaluation statistics by age, gender, and select occupations and locations for OS open roads.

		TP (%)	TN (%)	FP (%)	FN (%)
Gender	Female	1,497 (41.2)	1,590 (43.8)	141 (3.9)	404 (11.1)
	Male	1,418 (40.4)	1,589 (45.2)	140 (4)	365 (10.4)
Age	Under 20	1,222 (39.4)	1,442 (46.5)	122 (3.9)	316 (10.2)
	20 - 40	933 (41.9)	967 (43.4)	86 (3.9)	241 (10.8)
	40 - 60	571 (43.3)	534 (40.5)	61 (4.6)	153 (11.6)
	Over 60	194 (38.1)	244 (47.9)	12 (2.4)	59 (11.6)
Occupations	Domestic service	172 (42.9)	166 (41.4)	19 (4.7)	44 (11)
	Farming	77 (15.9)	368 (75.9)	18 (3.7)	22 (4.5)
	Manufacturing	342 (47.6)	247 (34.4)	39 (5.4)	90 (12.5)
	General labourers	52 (36.1)	69 (47.9)	10 (6.9)	13 (9)
E&W	Cornwall	8 (9.5)	66 (78.6)	5 (6)	5 (6)
	Norfolk	39 (37.1)	52 (49.5)	3 (2.9)	11 (10.5)
	Lancashire	351 (47.5)	256 (34.6)	45 (6.1)	87 (11.8)
	London	375 (48.7)	242 (31.4)	13 (1.7)	140 (18.2)
	West Riding Yorkshire	203 (41.9)	217 (44.7)	20 (4.1)	45 (9.3)
	Pembrokeshire	6 (24)	16 (64)	1 (4)	2 (8)
	Glamorganshire	44 (41.5)	38 (35.8)	1 (0.9)	23 (21.7)
Scotland	Argyll	3 (13)	20 (87)	- (0)	- (0)
	Edinburgh	64 (48.5)	59 (44.7)	2 (1.5)	7 (5.3)
	Lanark	121 (47.6)	87 (34.3)	5 (2)	41 (16.1)

rural areas with older populations. There was greater variability in types of links by occupation and location. Those working in farming had the lowest combined rate of TP and TN for GB1900 (74.7%) but the highest rate for OS Open Roads (91.8%). Accurate linking was achieved for poorer occupational groups, such as general labourers, who had the highest proportion of correct links (TP and TN) for GB1900. Domestic servants, who were living and working in more affluent households (and therefore linked to their employers' addresses), were linked at lower or the same accuracy to general labourers for both GB1900 and OS Open Roads. In terms of geography, major urban centers such as Edinburgh and London were linked very accurately, with FP making up less than 3% of links to both GB1900 and OS Open Roads. Yet, there is scope to improve overall linking rates in London in particular, since 18.2% of London addresses for OS Open Roads were FN. Elsewhere, the picture was varied, and there was no clear division between distinctly rural counties (Cornwall, Norfolk, Pembrokeshire, and Argyll) and counties with industrial, urban centers (Glamorganshire, Lancashire, and West Riding of Yorkshire). Overall, most differences in FP and FN rates by age, gender, occupation, and geography varied for GB1900 and OS Open Roads. Poorer quality linking to one was often mitigated by higher quality linking in the other, suggesting that using both geometry datasets in conjunction would achieve the most balanced sample of geo-coded addresses.

Finally, the process of validating *CensusGeocoder*'s matches also identified errors in the manual results. Re-checking discrepancies between the manual and automated outputs identified instances where *CensusGeocoder* found addresses which had not been identified manually (100 for GB1900, 55 for OS Open Roads). Its strict application of geo-blocking also meant that it correctly geo-coded 90 addresses to GB1900 points and 45 addresses to OS Open Roads that had been initially manually linked to the wrong entities.[13] These mistakes were made because a manual approach is less suited to geo-coding scattered addresses across Britain, and better applied to localized areas where addresses can be worked through as they appear on the map. In this way, you can build up local knowledge, use the context of adjacent addresses, and cross-reference with multiple map series to view different types of addresses, streets, places, and buildings at different scales. Moving around the country necessitates starting from scratch in each locality. Furthermore, manual geo-coding is subjective: it can be difficult to consistently define or justify when a possible match or probable match becomes a definite match. Time is a key factor, too – how long should you attempt to find an address before determining that it cannot be geo-coded? Here, despite its limitations, the automated method is, by comparison, highly systematic in its application of rules.

Representativeness

AddressGB is highly accurate in the addresses it geo-codes, though of course it does not geo-code all addresses. Some of these, as noted above, are false negatives, and others could be geo-coded if gazetteers of other historic maps collections (particularly the 25 inch maps) were available in place of GB1900. Individuals with a geo-coded address are therefore a

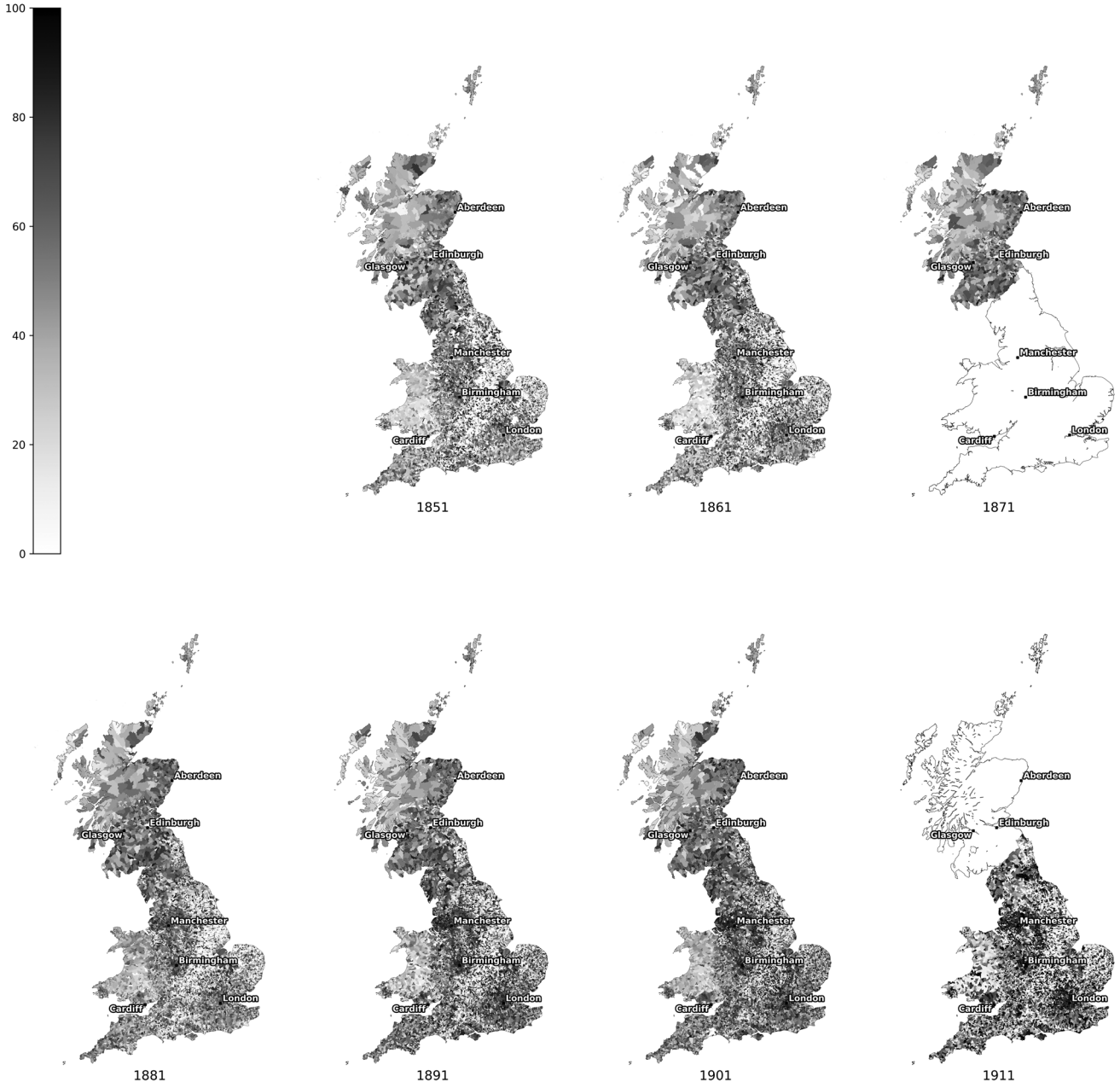

Figure 8. (a) Percentage of population geo-coded by administrative unit, Great Britain 1851–1911.

subset of the population and we need to assess how far they reflect (or do not reflect) the wider population. This section examines the geographical distribution of *AddressGB* and its representativeness on key socio-demographic indicators of age, gender, and occupation.

Figure 8a shows the geographical distribution of linking rates across Great Britain between 1851 and 1911. It shows the total percentage of people geo-coded (to GB1900 and/or OS Open Roads) in each geo-blocking unit. Overall, linkage rates were higher in urban than rural areas because rural addresses were less structured than their urban counterparts, listed as individual properties on unnamed roads, rather than listed in order on named streets. Clusters of high rates can be seen around key urban centers such as London, Birmingham, Manchester, Cardiff, Glasgow, and Edinburgh, and low rates in rural areas of Wales and Scotland in particular.

Figures 8b and 8c show the linking rates for individuals matched to only OS Open Roads or only GB1900. Addresses linked to only OS Open Roads (with no corresponding match made to GB1900) were predominantly in urban areas which had developed street networks (Figure 8b). This was particularly

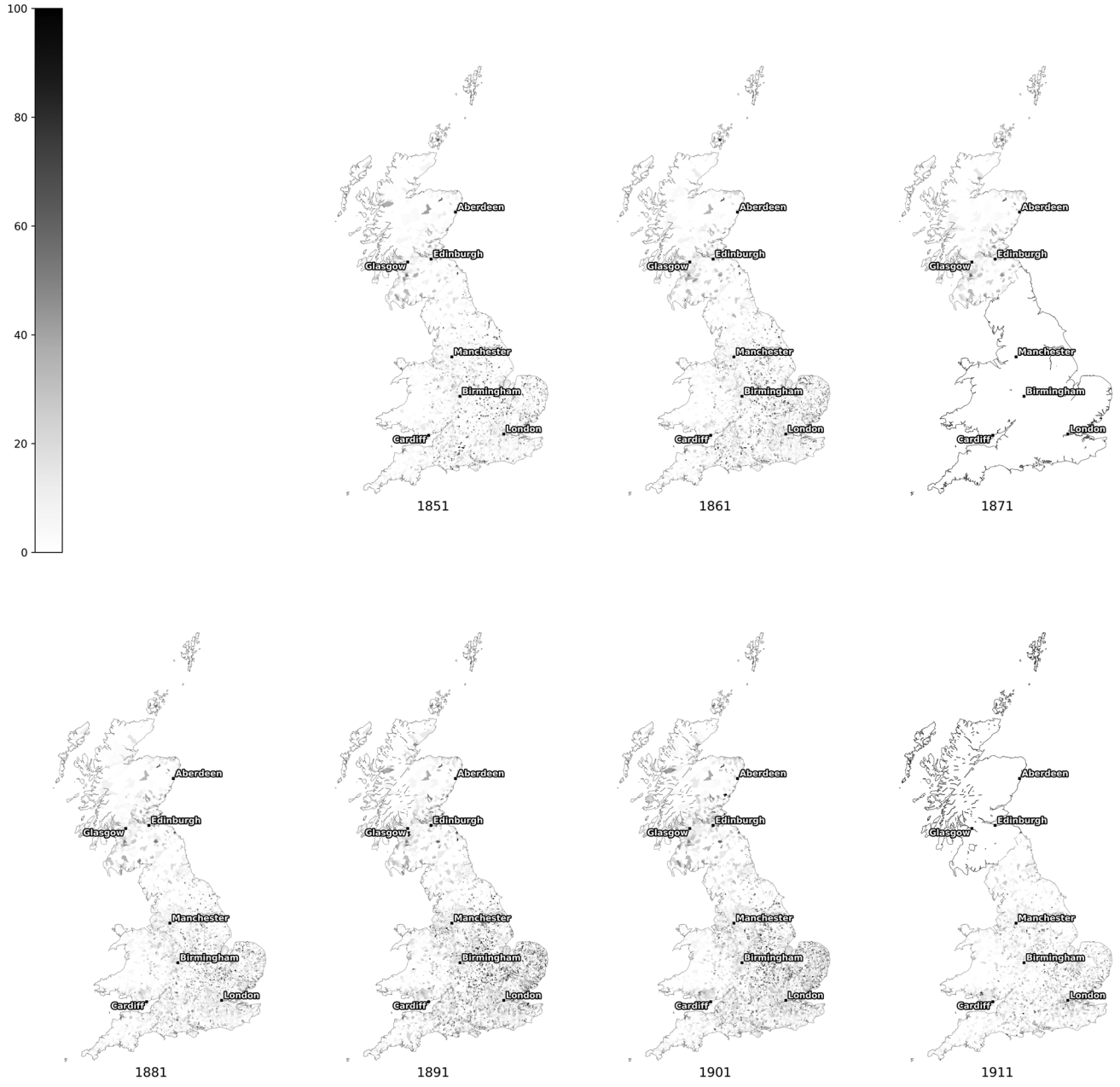

Figure 8. (b) Percentage of population geo-coded to OS open roads only by administrative unit, Great Britain 1851–1911.

pronounced in 1911 with the area of major urban centers of London, Birmingham, Manchester, and Cardiff visibly delineated from the surrounding rural areas by their higher linking rates. In comparison, Figure 8c shows GB1900 was an important source of links in Cornwall, Wales, northern parts of England, and Scotland. GB1900 provides more matches for rural addresses because some of these individual properties appeared on the historic Ordnance Survey maps. Using *AddressGB* combines the respective strengths in urban and rural coverage of OS Open Roads and GB1900.

Figures 9, 10, and 11 compare differences in gender, age, and occupational distributions of individuals according to their geo-code status. Figures 9 and 10 show minimal differences in gender and age distributions. Women were slightly over-represented among the OS Open Roads only and GB1900 & OS Open Roads categories. The age distributions were broadly similar for each census year and linking source too. People under the age of 40 (but not children under 10) were slightly over-represented among those geo-coded, while individuals over 40 were under-represented. The over-representation of women and younger people in *AddressGB* reflects urban demographics due to the higher matching rates in urban areas.

Figure 11 shows the occupational structure of Britain by geo-code status, using an occupational classification developed by Bennett et al. (2017, 2018).[14] The figure

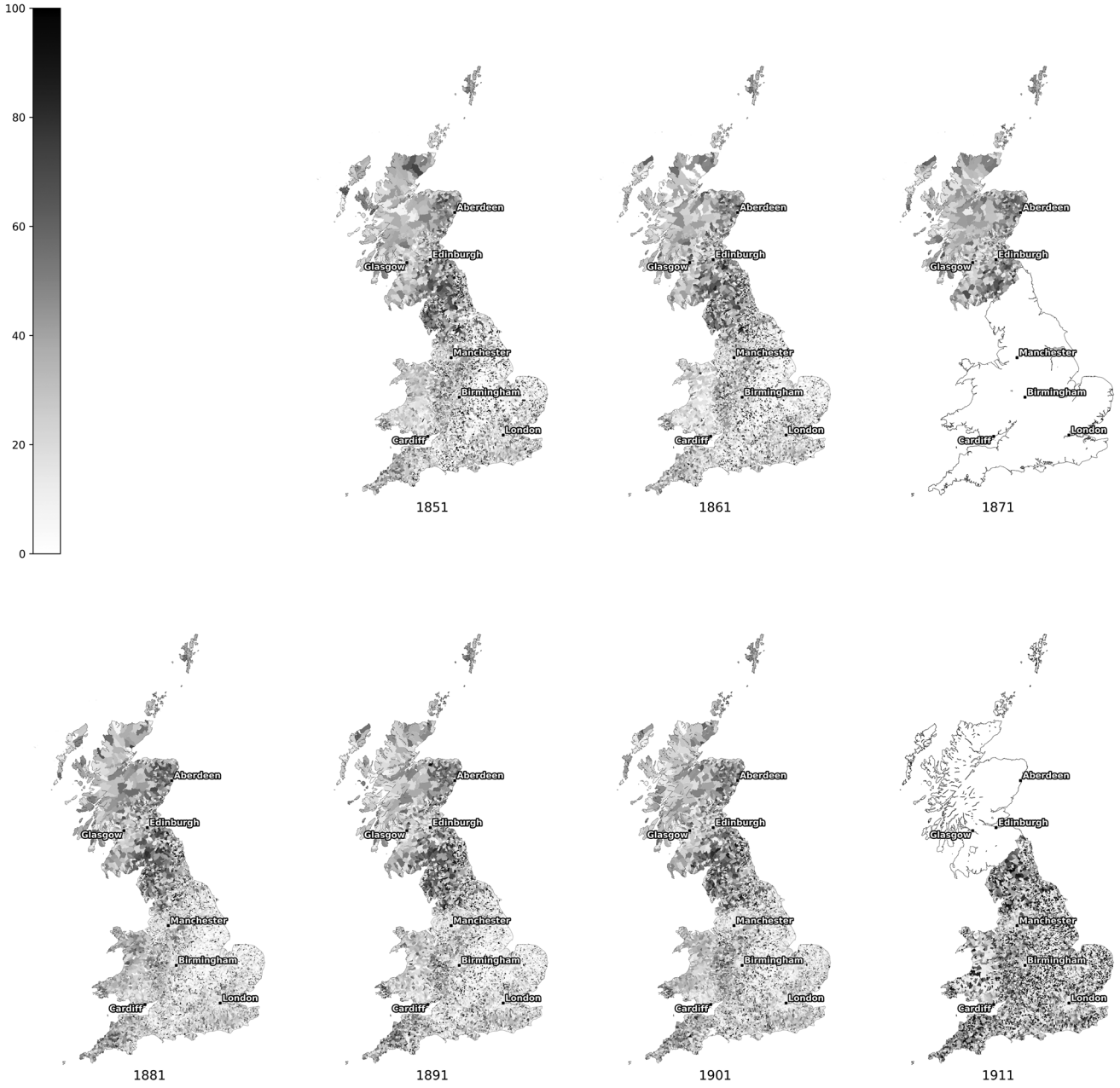

Figure 8. (c) Percentage of population geo-coded to GB1900 only by administrative unit, Great Britain 1851–1911.

shows the proportion of individuals in each occupational category. For example, in the 1851 not geo-coded subset, 20.7% of the working population worked in manufacturing compared to 22.2% in the total geo-coded subset. For many occupational groups there was minimal difference between the geo-coded and not geo-coded individuals. Figure 11 only shows occupational categories with the largest differences. The largest differences were often between who was linked to GB1900 compared to OS Open Roads, which essentially reflect the dominance of rural addresses linked to GB1900 and urban addresses linked to OS Open Roads. But when combined in the total geo-coded subset these differences are more muted.

The categories of farming and manufacturing illustrate this point most clearly. In the 'GB1900 Only' subset for 1851, for example, those working in farming made up over 30% of the workforce. By contrast, among subsets with individuals linked to OS Open Roads ('OS Open Roads only' or 'GB1900 and OS Open Roads'), workers in farming constituted only around 10% of the working population. Combining the subsets ('total geo-coded') the proportion of workers in farming was around 19% compared to 24% among those not geo-coded. The disparity in the 1871 Scottish data is particularly pronounced between OS Open Roads only and GB1900 only but much closer when comparing the

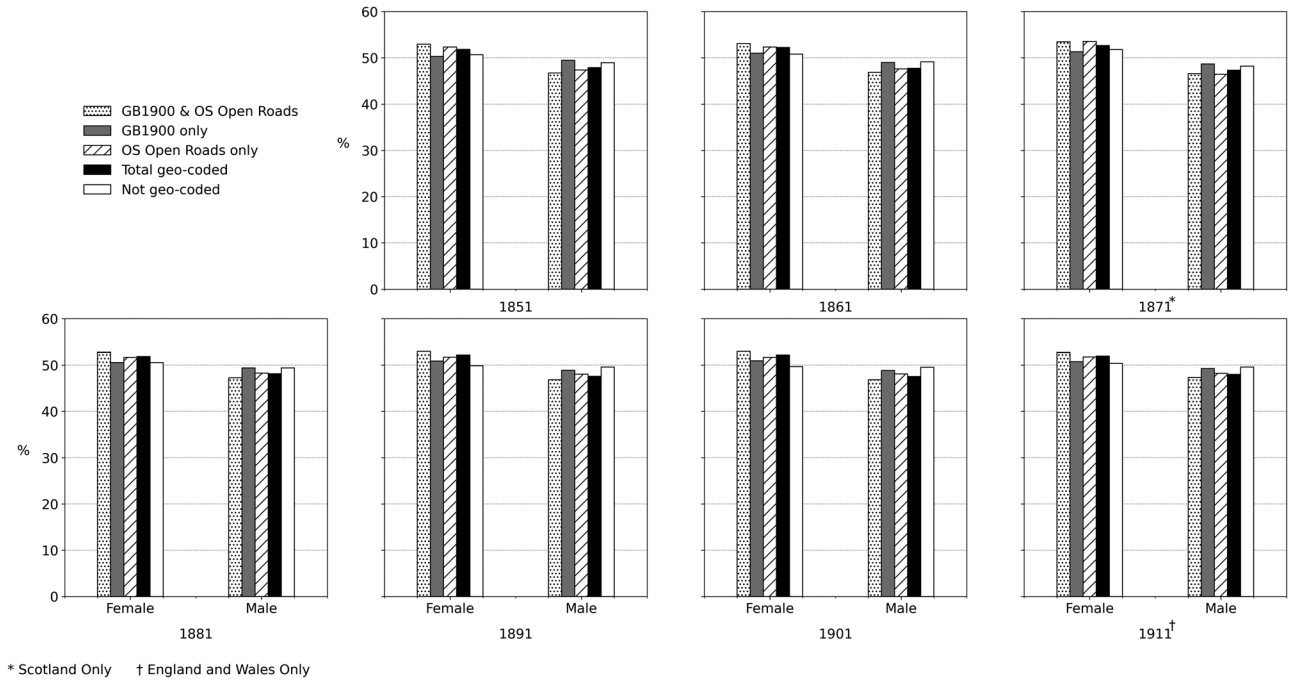

Figure 9. Gender distribution by geo-code status, Britain 1851–1911.

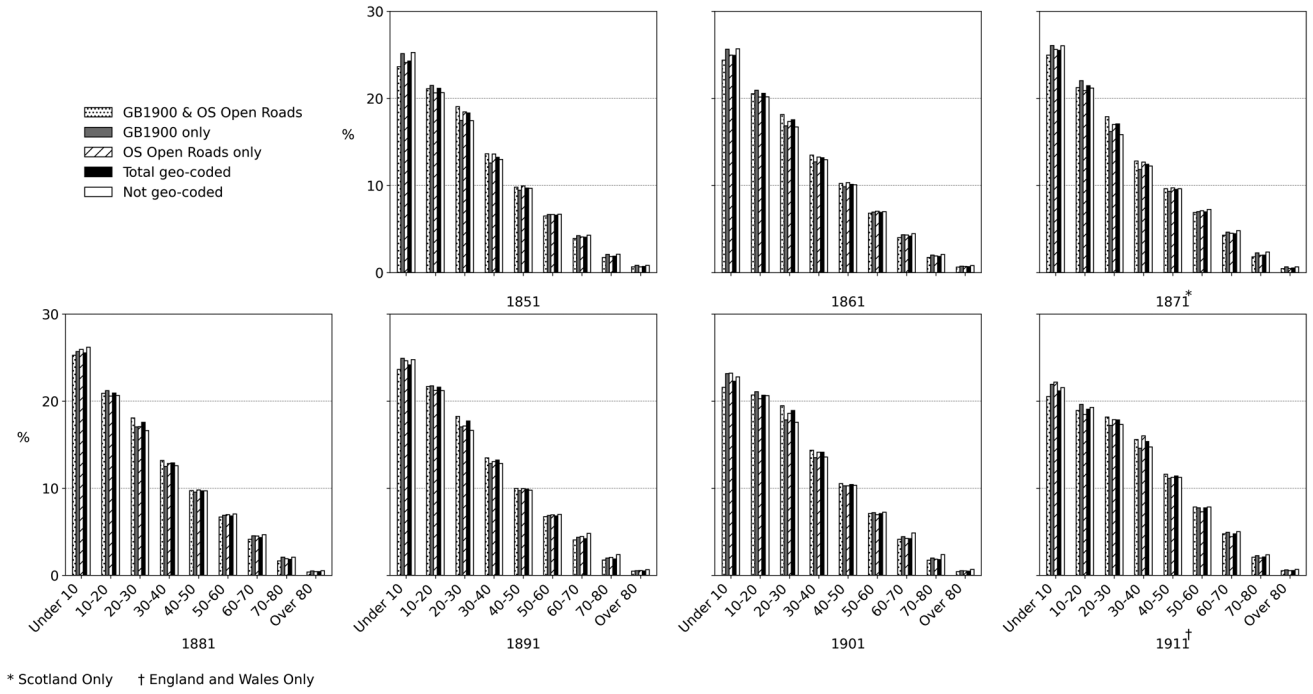

Figure 10. Age distribution by geo-code status, Britain 1851–1911.

not geo-coded and total geo-coded subsets. In each year, the proportion of farm workers is underrepresented in the total geo-coded group. The opposite is the case for manufacturing, whose workers were situated in and around urban centers (where linking rates were higher). Using GB1900 and OS Open Roads together compensates for these urban-rural disparities to some extent. Overall, we can be confident that the gender, age, occupational, and geographical distribution of *AddressGB* is representative of the full census. Crucially, any differences are consistent over time despite substantial changes in the proportion of individuals geo-coded in each census.

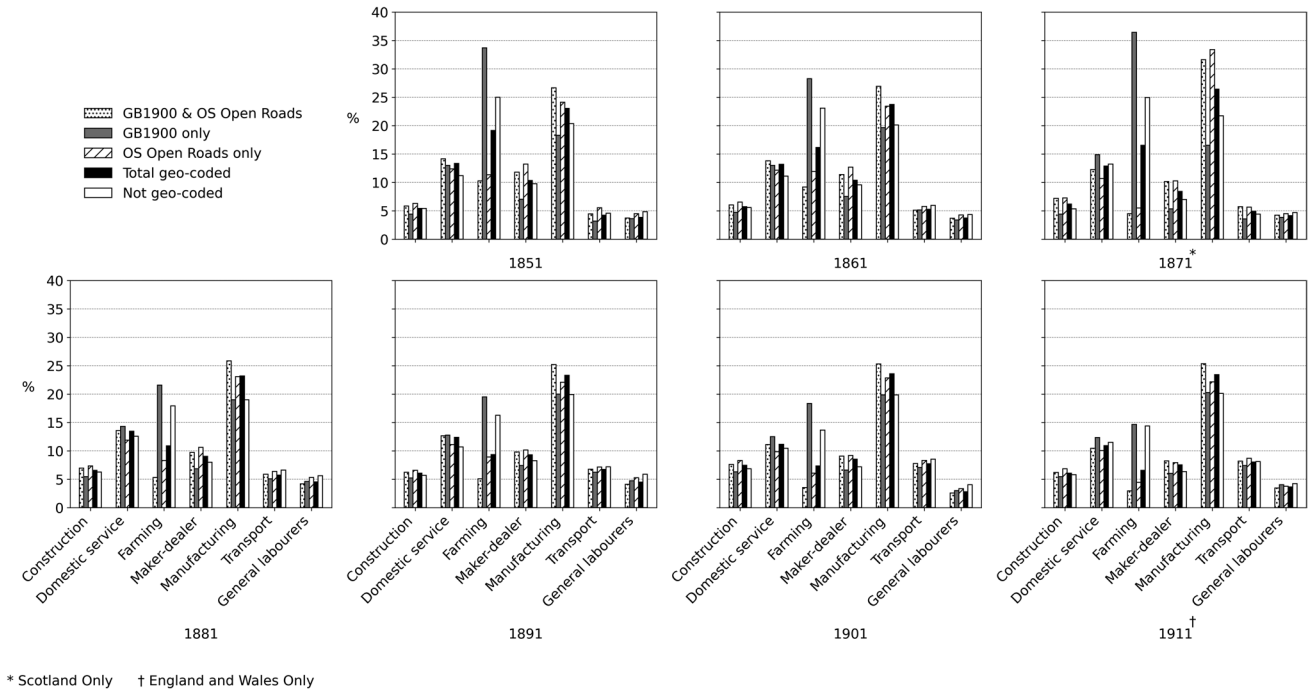

Figure 11. Occupational structure by geo-code status, Britain 1851–1911.

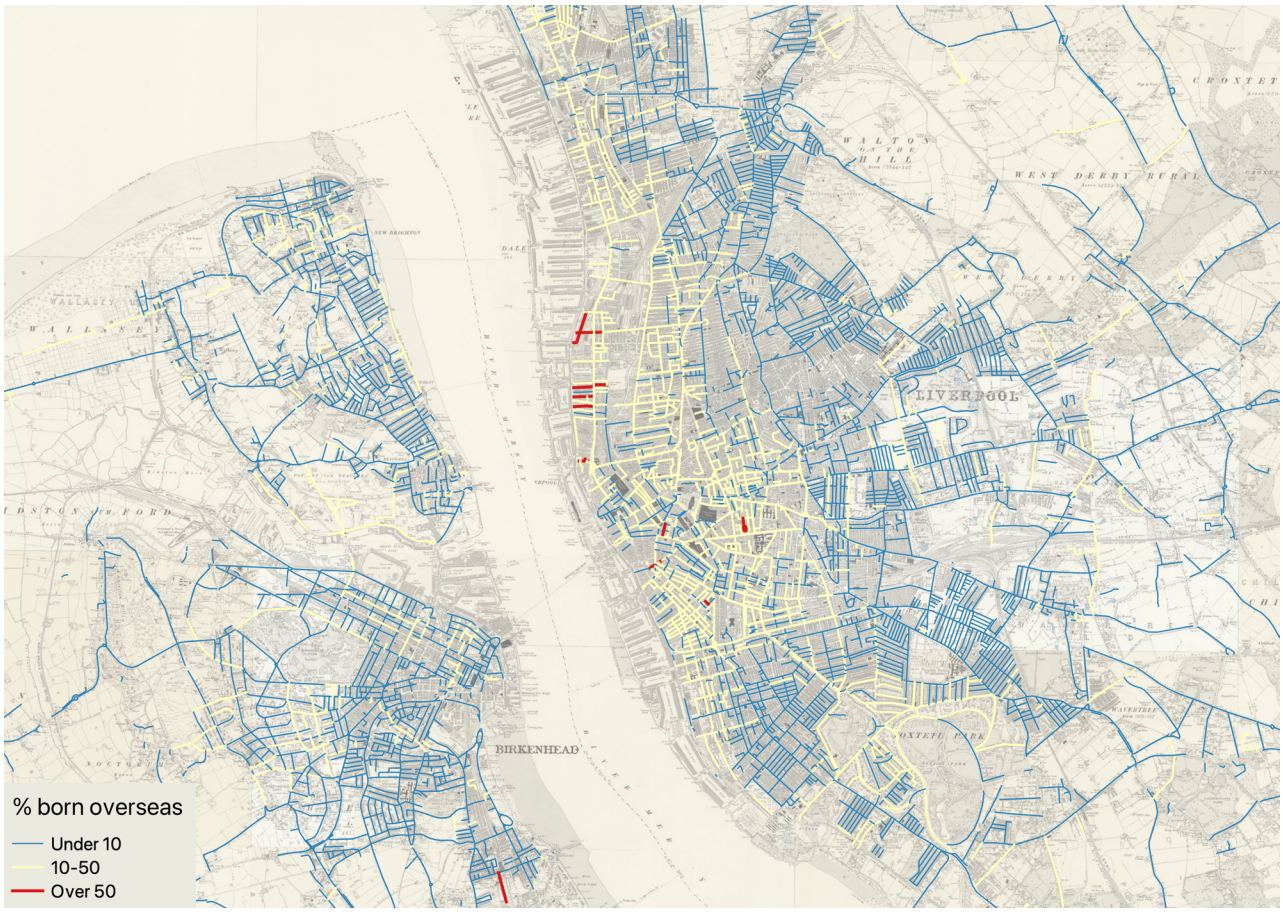

Figure 12. Percentage of individuals born overseas by street, Liverpool 1901. Reproduced with the permission of the National Library of Scotland.

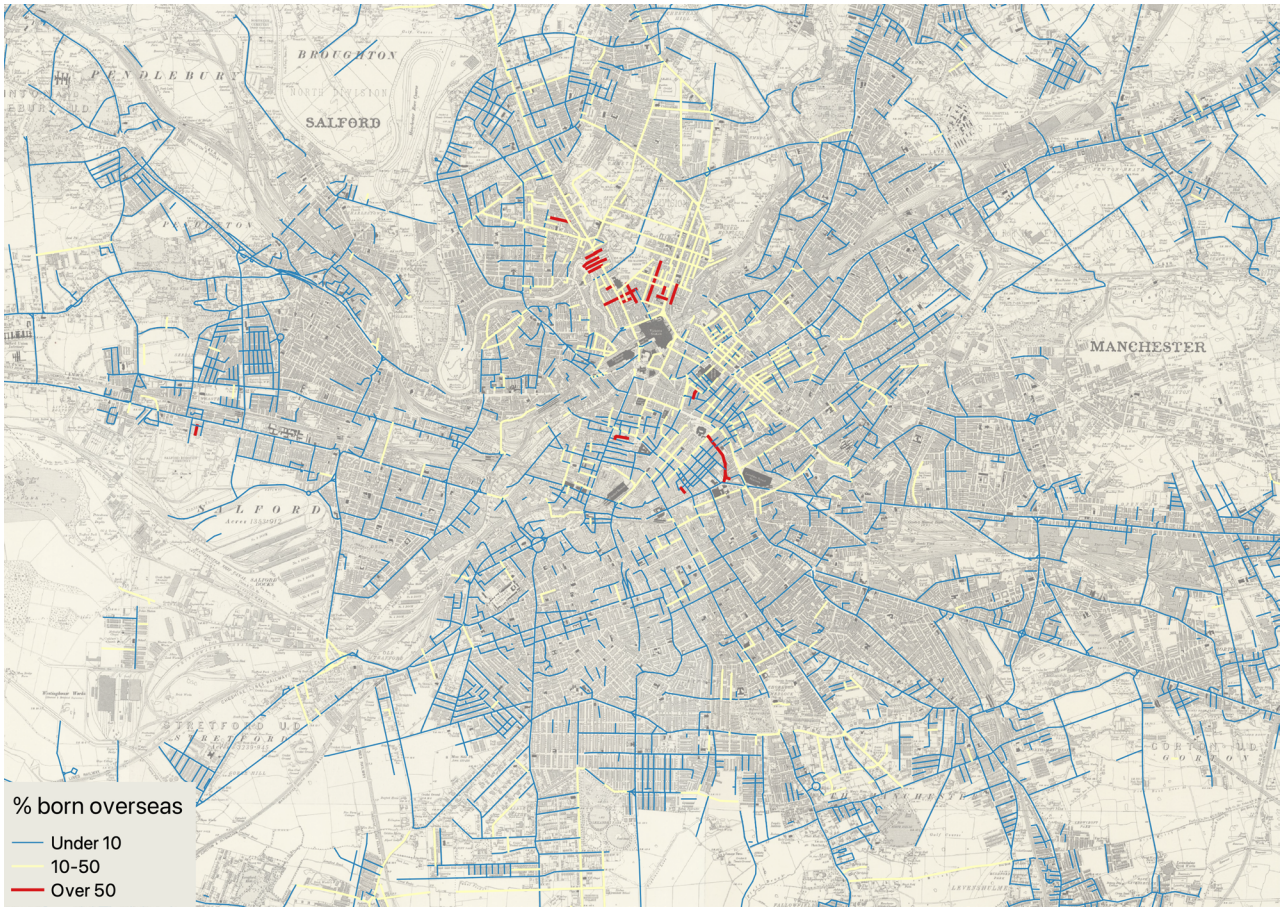

Figure 13. Percentage of individuals born overseas by street, Manchester 1901. Reproduced with the permission of the National Library of Scotland.

Conclusion and future research possibilities

AddressGB offers exciting new research possibilities. Primarily, it enables census variables to be mapped at a higher spatial resolution than has been previously possible to identify patterns and clusters of streets with similar characteristics. Figures 12-14 demonstrate *StreetGB*'s potential for identifying streets with high proportions of individuals born overseas in Liverpool, Manchester, and London in 1901.[15] These cities are well-known for having clusters of foreign-born communities (for example in London's West and East End), but *AddressGB* delineates the extent of these communities street by street. *AddressGB* allows researchers to isolate these streets and analyze their demographic and socio-economic make-up using other census variables.

More significantly, *AddressGB* enables researchers to spatially link I-CeM at street level to other high-resolution geo-referenced datasets, opening up important and exciting new research directions. Individuals in I-CeM can then be categorized according to any variables in the secondary linked dataset. *AddressGB* is currently being used in this way to re-assess the impact of the development of the nineteenth- and early twentieth-century rail network (Rhodes et al. forthcoming) by linking it to geo-located station data (Coll Ardanuy et al. 2021) and track and building data (Hosseini et al. 2022).

Applications of this type are wide-ranging and far-reaching, especially in the context of the increasing quantity of digitized nineteenth-century material made available online. For example, Charles Booth's late nineteenth-century poverty maps of London have recently been digitized and geo-referenced (Charles Booth's London). Booth's classifications of social status and deprivation would substantially enrich the existing demographic and occupational data in I-CeM. *AddressGB* allows census data to be linked to these detailed maps for the first time. Alternatively, *AddressGB* could be used to explore long-run localized socio-economic inequalities by linking historic censuses to high-resolution geographies of 2021 census data. By releasing *AddressGB* and *CensusGeocoder* openly, I invite other researchers to explore these possibilities and more.

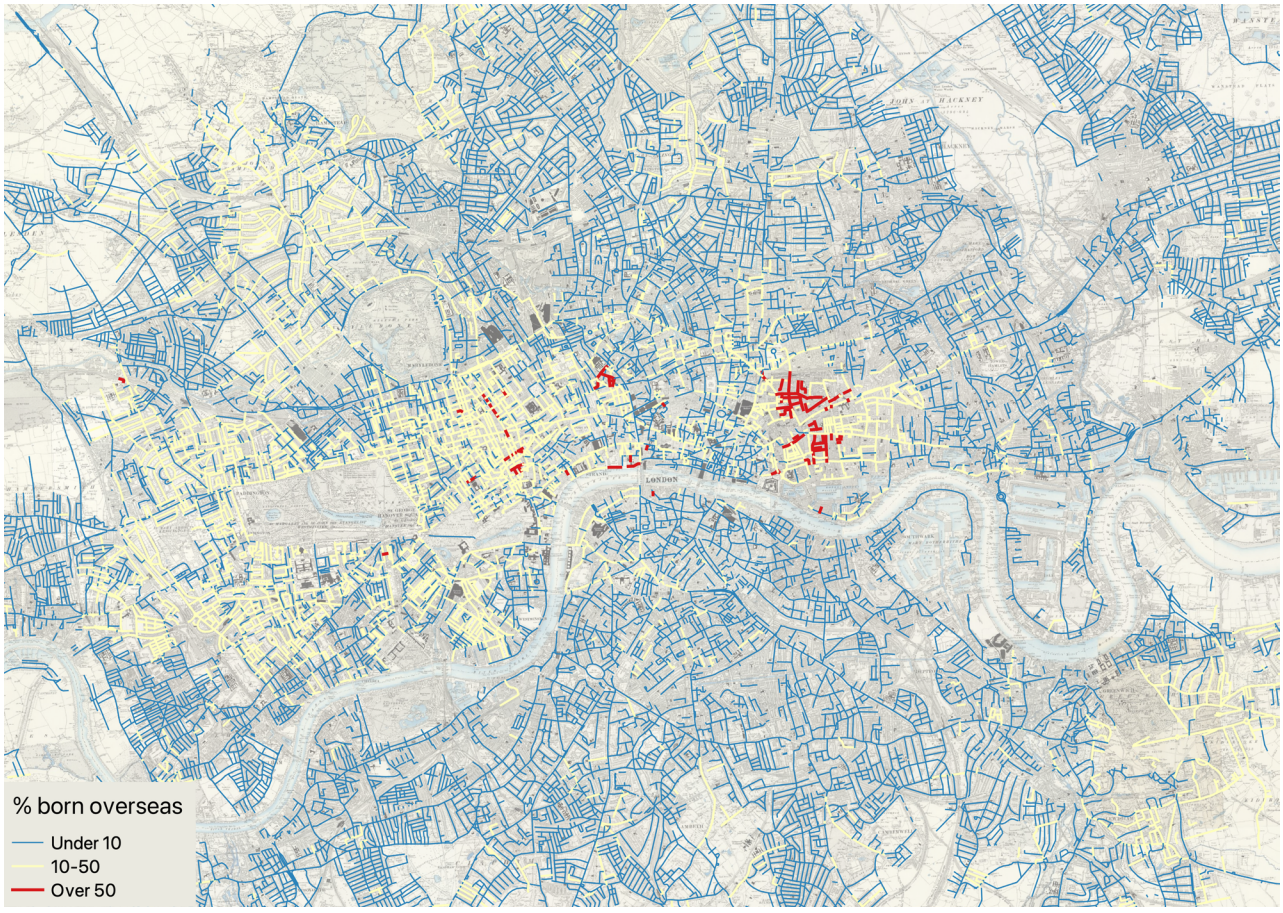

Figure 14. Percentage of individuals born overseas by street, London 1901. Reproduced with the permission of the National Library of Scotland.

Notes

1. Day et al. (2016) contains high resolution registration sub-district boundaries not yet publicly available. I am very grateful to Joe Day for supplying these boundaries in advance of their deposit with the UK Data Service (UKDS). Simplified versions can be downloaded from https://populationspast.org.
2. Calculated from Day et al. (2016).
3. The 2024 release of I-CeM includes the latitude and longitude of 1921 census addresses.
4. For full details of the variables, see the I-CeM website documentation at https://www.campop.geog.cam.ac.uk/research/projects/icem/.
5. The rest of their household is not shown in this sample.
6. 'RecID' also provides a way to link secondary or derived datasets, such as *AddressGB* or BBCE (Bennett et al. 2020), back to I-CeM.
7. Lan and Longley (2021) also note the drop in the number of unique addresses recorded in 1891 I-CeM data.
8. There are extant GIS datasets for Scottish parish boundaries for 1851-1911 but only parish boundaries in 1851 for England and Wales. In contrast, RSDs are available for England and Wales 1851-1911 but only consistent Registration Districts are available for Scotland (Satchell 2023).
9. This section describes the ConParID system as it was in I-CeM (Schürer and Higgs 2014). This system has been replaced in the new version of I-CeM (Schürer, Higgs, and FINDMYPAST LIMITED 2024; Schürer, Wakelam, and FINDMYPAST 2024a), which corrects some of the errors in the old consistent parishes and creates one boundary series for the whole 1851-1921 period. However, due to the large size of the new consistent parishes and the quality of the boundary data (suitable for aggregate visualisation but not for precise address geo-blocking), I have used the old consistent parish system. For more details on the new consistent parish system, see I-CeM documentation and website.
10. If a label appears three or more times within a single unit, we can be confident that it refers to a common map feature rather than a place or street entity. These labels are removed. But the same label appearing twice within a unit could be a street, since long streets might have their names written at different points along them. Alternatively, they could be two roads with the same name in the same administrative unit. To address this issue, if two labels were further than 1km apart, they are assumed to be different roads and are removed because it is not possible to distinguish between them. But if the labels were less than a 1km apart, it is deemed reasonably likely that they referred to the same street, and so one of the labels is kept and the other removed.
11. The mean number of characters in addresses in the

England and Wales 1901 census was 20.8 (median 20), compared to 30.3 (median 30) in 1911.

12. They are output in separate files for further inspection, with the potential for identifying final matches among these at a later stage. See *CensusGeocoder* Github Repository for further information.
13. The errors identified in the manual sample were corrected before calculating the proportion of TP, FP, TN, and FN in Tables 4 and 5.
14. A modified version of the 17cat scheme is used, which removes the 'Persons of property' category. An error in the original scheme categorises all scholars and others with no occupational information to this category. Consequently, 'Persons of property' makes up c. 30-40% of individuals in each census, which is a misleadingly high proportion for such an elite social group. Only categories with the largest differences between the geo-coded and not geo-coded subsets have been included.
15. The examples given here use only OS Open Roads geometries to visualise street layouts. Overseas refers to those not born in England, Wales, or Scotland.

Acknowledgements

I would like to acknowledge Living with Machines colleagues involved in the wider work strand on railspace for which *CensusGeocoder* was developed: Kaspar Beelen, Mariona Coll Ardanuy, Jon Lawrence, Katie McDonough, Guy Solomon, and Daniel Wilson. I am also very grateful to Charmian Mansell for reading and commenting on the full article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been supported by The Living with Machines Project (Grant Reference AH/S01179X/1). This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library and Cambridge, King's College London, East Anglia, Exeter, and Queen Mary University of London.

ORCID

Joshua Rhodes http://orcid.org/0000-0002-4017-2777

References

Code

Rhodes, J. CensusGeocoder. 2024a. https://github.com/Living-with-machines/CensusGeocoder

Data

Bennett, R., H. Smith, C. Van Lieshout, and G. Newton. 2018. Sector and occupation classification data download for "business sectors, occupations and aggregations of census data 1851–1911". doi: 10.17863/CAM.26433.

Bennett, R., Smith, H., van Lieshout, C., Montebruno, P., Newton, G. 2020. British Business Census of Entrepreneurs, 1851–1911. UK Data Service SN: 8600. http://doi.org/10.5255/UKDA-SN-8600-2

Coll Ardanuy, M., Beelen, K., Lawrence, J., McDonough, K., Nanni, F., Rhodes, J, Tolfo, G, Wilson, D. 2021. StopsGB: Structured Timeline of Passenger Stations in Great Britain. https://doi.org/10.23636/wvva-3d67

Day, J. 2016. Registration sub-district boundaries for England and Wales 1851–1911. Simplified, Public Versions Available as Reid et al. 2018.

GB1900 GAZETTEER, (COMPLETE). https://www.visionofbritain.org.uk/data/#tabgb1900

Hosseini, K., D. C. S. Wilson, K. Beelen, and K. McDonough. 2022. MapReader_Data_SIGSPATIAL_2022 (v0.3.3). *Zenodo*. doi: 10.5281/zenodo.7147906.

Jaadla, H., and K. Schurer. 2023. Address dictionary for subdivisions of consistent Scottish registration district boundaries in larger towns in ICeM, 1851–1901. *Apollo - University of Cambridge Repository*. doi: 10.17863/CAM.95058.

Reid, A. M., S. J. Arulanantham, J. D. Day, E. M. Garrett, H. Jaadla, and M. Lucas-Smith. 2018. Populations past: Atlas of Victorian and Edwardian population. https://www.populationspast.org/

Rhodes, J. 2024b. *AddressGB*: Geo-coded British census addresses, 1851–1911. doi: 10.5281/zenodo.10473597.

Rhodes, J. 2024c. *AddressGB* manual evaluation sample. doi: 10.5281/zenodo.13770048.

Rhodes, J. 2024d. I-CeM Scotland Parish lookup tables, 1851–1901. doi: 10.5281/zenodo.10473644.

Roughley, C., and M. Anderson. 2019. Scotland's parish populations: Parish boundaries, 1755–1891. https://www.nrscotland.gov.uk/files//geography/products/scotlands-parish-populations-1755-1891.pdf

Satchell, A. 2023. Consistent Scottish registration district boundaries with subdivisions in large towns, 1851–1901. *Apollo - University of Cambridge Repository*. doi: 10.17863/CAM.94398.

Satchell, A., P. Kitson, G. Newton, L. Shaw-Taylor, and E. Wrigley. 2017. 1851 England and Wales census parishes, townships and places. *UK Data Service SN: 852816*. doi: 10.5255/UKDA-SN-852232.

Satchell, M., L. Shaw-Taylor, E. Wrigley, P. Kitson, and G. Newton. 2018. 1851 England and Wales Census registration counties. *UK Data Service SN:852949*. doi: 10.5255/UKDA-SN-852949.

Schürer, K., and E. Higgs. 2014. Integrated Census Microdata (I-CeM), 1851–1911. *UK Data Service SN:7481*. doi: 10.5255/UKDA-SN-7481-1.

Schürer, K., and E. Higgs. 2015. Integrated Census Microdata (I-CeM) names and addresses, 1851–1911: Special Licence Access. *UK Data Service SN:7856*. doi: 10.5255/UKDA-SN-7856-1.

Schürer, K., and E. Higgs. 2020. Integrated Census Microdata (I-CeM) names and addresses, 1851-1911: Special Licence

Access, 2nd edition. *UK Data Service SN:7856*. doi: 10.5255/UKDA-SN-7856-2.
Schürer, K., and E. Higgs, FINDMYPAST LIMITED. 2024. Integrated Census Microdata (I-CeM), 1851–1911, 2nd edition. *UK Data Service SN:7481*. doi: 10.5255/UKDA-SN-7481-3.
Schürer, K., and A. Wakelam, FINDMYPAST LIMITED. 2024a. Integrated Census Microdata (I-CeM), England and Wales, 1921. *UK Data Service SN:9280*. doi: 10.5255/UKDA-SN-9280-1.
Schürer, K., and A. Wakelam, FINDMYPAST LIMITED. 2024b. Integrated Census Microdata (I-CeM) names and addresses, England and Wales, 1921: Special licence access. *UK Data Service SN:9281*. doi: 10.5255/UKDA-SN-9281-1.
Southall, H. R., I. Gregory, N. Burton, and P. Aucott. 2022. Great Britain Historical Database: Digital boundaries for registration counties of England and Wales, 1851–1911. *UK Data Service SN:9033*. doi: 10.5255/UKDA-SN-9033-1.
OS Open Roads, Ordnance Survey. https://www.ordnancesurvey.co.uk/products/os-open-roads

Websites

Charles Booth's London https://booth.lse.ac.uk/
Baics, G., W. Kennedy, R. Kobrin, L. Kurgan, L. Meisterlin, D. Miller, and M. Ngai. 2021. *Mapping historical New York: A digital atlas*. New York, NY: Columbia University. https://mappinghny.com.

Secondary works

Bennett, R., H. Smith, C. Van Lieshout, and G. Newton. 2017. *Business sectors, occupations and aggregations of census data 1851–1911*. doi: 10.17863/CAM.9874.
Bennett, R. J., H. Smith, C. van Lieshout, P. Montebruno, and G. Newton. 2019. *The age of entrepreneurship: Business proprietors, self-employment and corporations since 1851*. London and New York: Routledge International Studies in Business History. doi: 10.4324/9781315160375.
Bogart, D., X. You, E. J. Alvarez-Palau, M. Satchell, and L. Shaw-Taylor. 2022. Railways, divergence, and structural change in 19th century England and Wales. *Journal of Urban Economics* 128:103390. doi: 10.1016/j.jue.2021.103390.
Day, J. 2020. The process of internal migration in England and Wales, 1851–1911: Updating Ravenstein and the step-migration hypothesis. *Comparative Population Studies* 44:447–96. doi: 10.12765/CPoS-2020-13.
Higgs, E. 1991. *Making sense of the census*. London: HMSO.
Jaadla, H., A. Reid, E. Garrett, K. Schürer, and J. Day. 2020. Revisiting the fertility transition in England and Wales: The role of social class and migration. *Demography* 57 (4):1543–69. doi: 10.1007/s13524-020-00895-3.
Jiao, C., M. Heitzler, and L. Hurni. 2021. A survey of road feature extraction methods from raster maps. *Transactions in GIS* 25 (6):2734–63. doi: 10.1111/tgis.12812.
Kim, J., Z. Li, Y. Lin, M. Namgung, L. Jang, and Y. Chiang. 2023. The mapKurator system: A complete pipeline for extracting and linking text from historical maps. In SIGSPATIAL '23: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, 35, 1–4, doi: 10.1145/3589132.3625579.
Lan, T., and P. Longley. 2019. Geo-referencing and mapping 1901 census addresses for England and Wales. *ISPRS International Journal of Geo-Information* 8 (8):320. doi: 10.3390/ijgi8080320.
Lan, T., and P. Longley. 2021. Urban morphology and residential differentiation across Great Britain, 1881–1901. *Annals of the American Association of Geographers* 111 (6):1–20. doi: 10.1080/24694452.2020.1859982.
Lawton, R. 1978. *Census and social structure*. London: Routledge.
Li, Z., Y. Y. Chiang, S. Tavakkol, B. Shbita, J. H. Uhl, S. Leyk, and C. A. Knoblock. 2020. An automatic approach for generating rich, linked geo-metadata from historical map images. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3290–3298. doi: 10.1145/3394486.3403381.
Marsh, D. 1965. *The changing social structure of England and Wales*. London: Routledge.
Notter, I. R., and J. R. Logan. 2022. Residential segregation under Jim Crow: Whites, Blacks, and Mulattoes in Southern Cities, 1880–1920. *City & Community* 21 (1):42–61. doi: 10.1177/15356841211052534.
Philips, R. C. M., M. Calabrese, R. Keenan, and B. van Leeuwen. 2022. The regional occupational structure in interwar England and Wales. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55 (2):78–97. doi: 10.1080/01615440.2022.2027303.
Rhodes, J., J. Lawrence, K. Beelen, K. McDonough, and D. C. S. Wilson. Forthcoming. Beyond the tracks: re-connecting people, places and stations in the history of late-Victorian railways. In *Ruth Ahnert, Emma Griffin, Jon Lawrence and the Living with Machines Project, Living with Machines: Computational Histories of the Age of Industry, University of London Press, London*. https://read.uolpress.co.uk/projects/living-with-machines
Shertzer, A., R. P. Walsh, and J. R. Logan. 2016. Segregation and neighborhood change in northern cities: New historical GIS data from 1900–1930. *Historical Methods* 49 (4):187–97. doi: 10.1080/01615440.2016.1151393.
Smith, H., R. J. Bennett, and C. van Lieshout. 2022. Industrial districts, entrepreneurship and the economic geography of Great Britain, 1851–1911. In *Industrial clusters: Knowledge, innovation systems and sustainability in the UK*, eds. J.F. Wilson, C. Corker and J. Lane, 10–31. New York: Routledge. doi: 10.4324/9781003036357.
Sammut, C, and Webb, G. I., eds. 2010. *Encyclopedia of machine learning*. New York: Springer. https://link.springer.com/referencework/10.1007/978-0-387-30164-8.
Walford, N. S. 2019. Bringing historical British Population Census records into the 21st century: A method for geocoding households and individuals at their early-20th-century addresses. *Population, Space and Place* 25 (4):e2227. doi: 10.1002/psp.2227.