Geocoding historical census records in England and Wales

Tian Lan[*1], Guy Lansley[†1], Justin van Dijk[‡1], and Paul A. Longley[§1]

[1] Department of Geography, University College London

January 20, 2019

Summary

This paper describes efforts to geo-reference addresses from the 1901 Census of Population of England and Wales by linking them to the contemporary OS AddressBase. The results indicate that it is feasible to standardise and geocode a large share of unique addresses from the historic database. Roughly 38% of addresses from 1901 could be linked to contemporary address coordinates. A further 25% of records could be allocated to a road. Geographic trends in the proportion of properties that could be matched were then explored to reveal fascinating insights about how housing has changed since 1901.

KEYWORDS: address matching, historical census data, geocoding, geo-demographics

1. Introduction

Historic Censuses are a valuable source of information on the British population and its change over time. Following 100 years since their collection dates, individual level records are made publicly available, and efforts to digitise the decennial censuses from the period 1851 to 1911 have resulted in comprehensive micro databases (Higgs & Schürer, 2014). However, unfortunately, historic geographic datasets are inherently unreliable and imprecise. At best, previous studies have been able to geolocate individual records to towns or parishes (Wall, 1982). This study attempts to apply methodologies that have been successfully employed to build a linked database on individuals from contemporary data to the historic data from 1901 in order to enhance their geographic information.

2. Linking of contemporary data

Over the past years, there has been growing interest in linking address-level data produced by governments and commercial organisations in order to supplement or even replace traditional population datasets. These new forms of data are typically of incomplete coverage and unknown quality (Lansley & Cheshire, 2018). However, residents, businesses and institutions will invariably report some addresses slightly differently on occasions hampering data linkage. One way to validate address data is through linkage to an official address product. The Ordnance Survey (OS) AddressBase provides a consolidated list of geo-referenced addresses. Unfortunately, most consumer and administrative address datasets do not follow a standardised procedure. Instead, address information are often volunteered by members of the public, and often in an unstructured format. Thus, it is not uncommon for inconsistencies to arise. Indeed, our previous research on linking electoral registers and consumer datasets found that across the 20 years of data sources there were twice as many raw unique address strings than there were actual domestic addresses – mainly due to varying level of detail in addresses. A subsequent matching algorithm based on trialling the most probable AddressBase token combinations reduced this list substantially, but even after a succeeding fuzzy matching stage roughly 3 million unique records could not be linked.

[*] tian.t.lan@ucl.ac.uk
[†] g.lansley@ucl.ac.uk
[‡] j.t.vandijk@ucl.ac.uk
[§] p.longley@ucl.ac.uk

3. Linking historic addresses to a contemporary database

The challenge with historic registers is ever greater. Not least because in 1901 there was no postcode system, but also because a large share (about 15%) of records do not include a street number. Furthermore, errors or missing data occurred during transcription. However, we attempt to link as many historic addresses as possible from the 1901 Census to the contemporary AddressBase. Firstly, we attempt to link addresses at the individual level. Where that is not possible, we attempt to link them at the street level instead. Given that the UK has experienced substantive changes since 1901 and that the quality of address information will be poor by today's standard, we do not expect to retain information for every property.

The workflow of address matching between the 1901 Census and OS AddressBase is shown in **Figure 1**. Whilst the contemporary analysis undertook matchings within postcode units, this study considers matches within parishes to account for the recurrence of street names across the UK. Thus, we assign each address in the AddressBase with a parish id, by spatially joining these addresses with a historical parish boundary. We further separate street numbers with street names in the address string, to extract unique street names within each parish. By doing so, we also significantly reduce the numbers of unique cases for subsequent the matching stage. We adopt a Levenshtein-Distance based Python package to match the Census addresses to the AddressBase at two levels. If an address is matched by both street name and street number, we consider it as an address-level match. Likewise, a street level match refers to addresses that can be matched to the thoroughfare only, either because the street number could not be matched or because there was no clear street number at all. In such cases, we assign a closest street number or the first valid street number from the AddressBase accordingly. The matched addresses are then geocoded through the precise coordinates of buildings as provided in AddressBase.

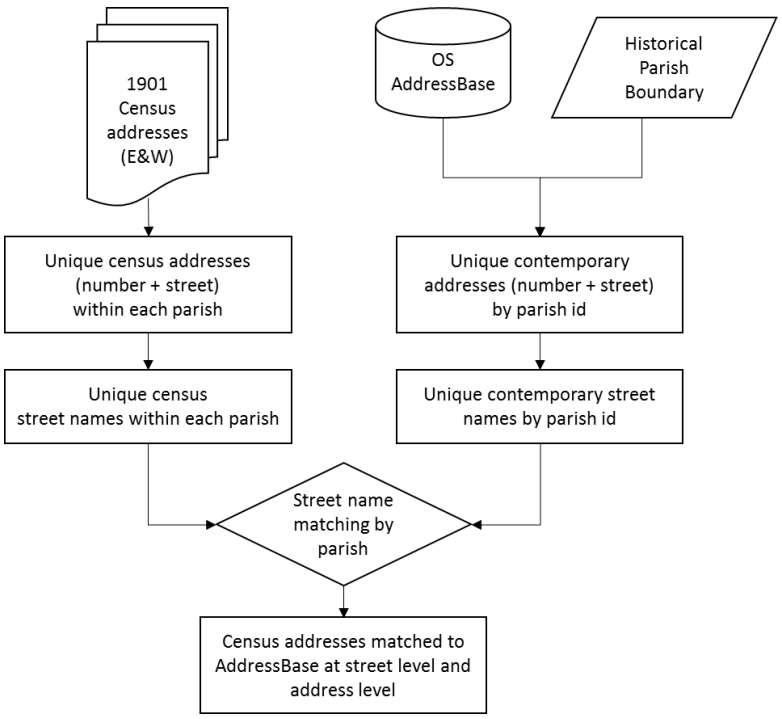

Figure 1 Workflow of the address matching process.

4. Results

Roughly 38% of unique addresses in the 1901 Census were matched to the contemporary AddressBase at the address-level. A further 25% could be matched at the street-level. We visualise the proportion of addresses from 1901 that matched the contemporary records by historical parishes in **Figure 2**. It is apparent that generally the match rates are greatest in and around traditional urban centres. This is probably indicative of better-quality management of address systems which were a necessity to ensure the deliverance of services in sprawling cities in 1901. In contrast, address information is typically scanter in rural areas. Street names and street numbers are less common as local landmarks were effective means of orientation. For instance, sometimes addresses within rural parishes were as vague as "barn" or "house with a carpenter's shop". We also observed that the portion of matches in Wales (2.4%) is lower than that in England (60.9%) largely because a large share of addresses were presented in the Welsh language.

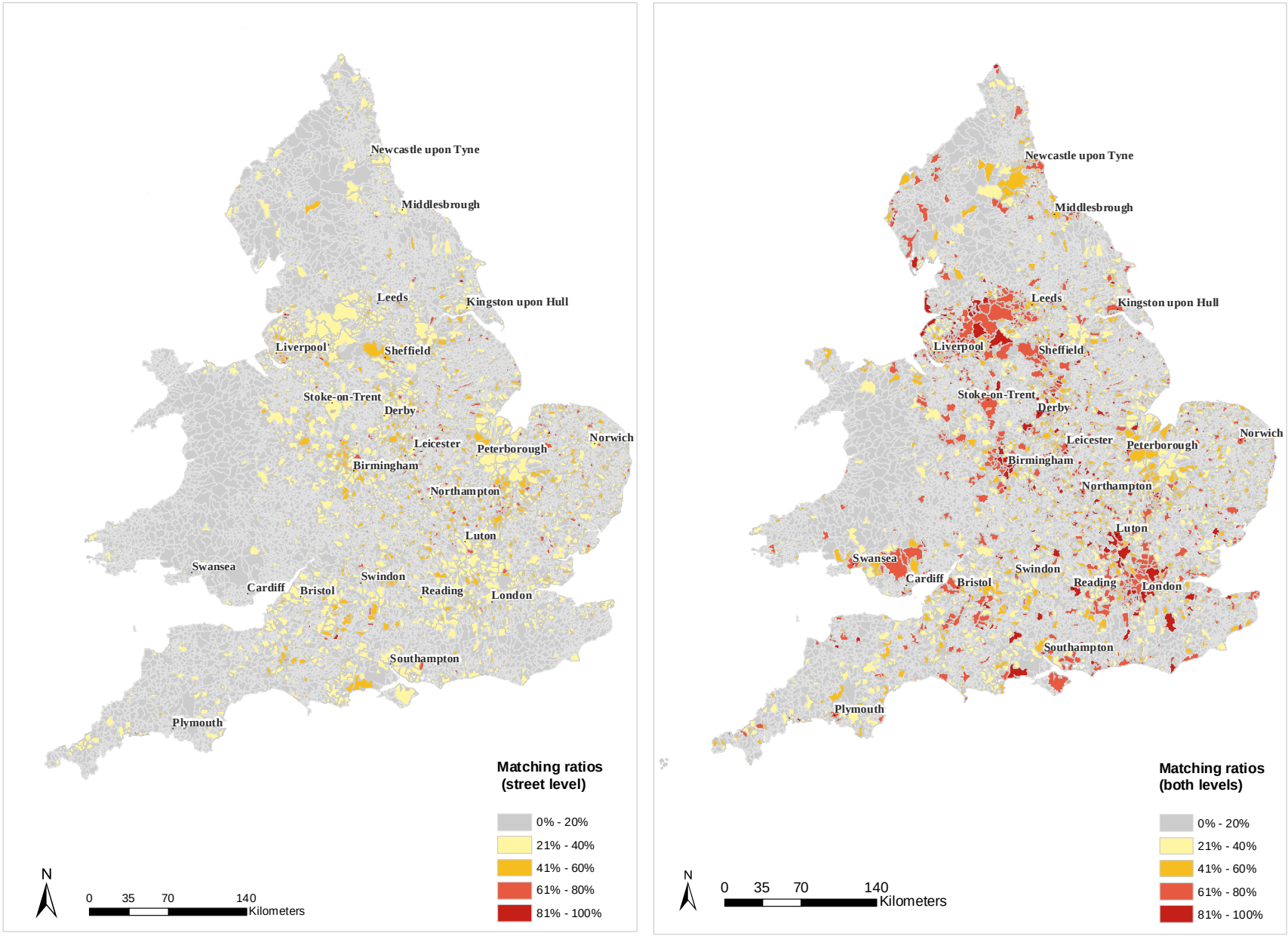

Figure 2 Matching ratios at street level (left) versus at address level plus street level (right) by historical parishes in England and Wales.

It is also apparent that urban changes since 1901 have caused localised geographic concentrations in streets that failed to link to the historic data. To demonstrate this, the street network on London has been coloured to show the streets that have at least one matched address, and those that did not (**Figure 3**). The map is almost indicative of London's suburban expansion following 1901. **Figure 4** compares the matched streets to a historic map of Clapham. Interestingly, the map highlights that large concentrations of unmatched streets occur on what used to be greenfield sites. The roads that cross sect the common in the centre of the image were unmatched presumably because there were no houses on these roads for them to occur in the Census in 1901.

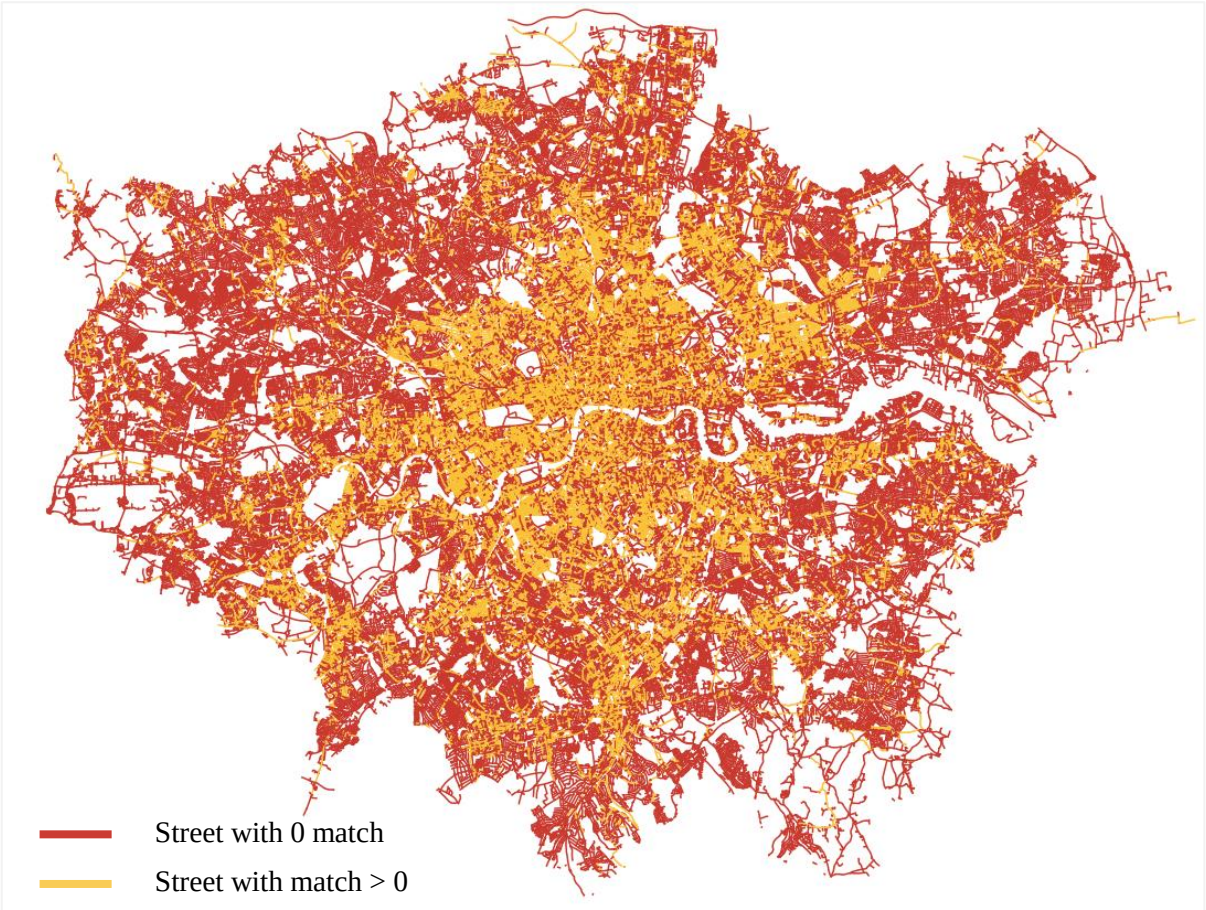

Figure 3 The street network of London in 2016 showing streets that have retained at least one identical address (yellow) and those that have no matches at all (red).
(Source: Street network from the OS Open Road data)

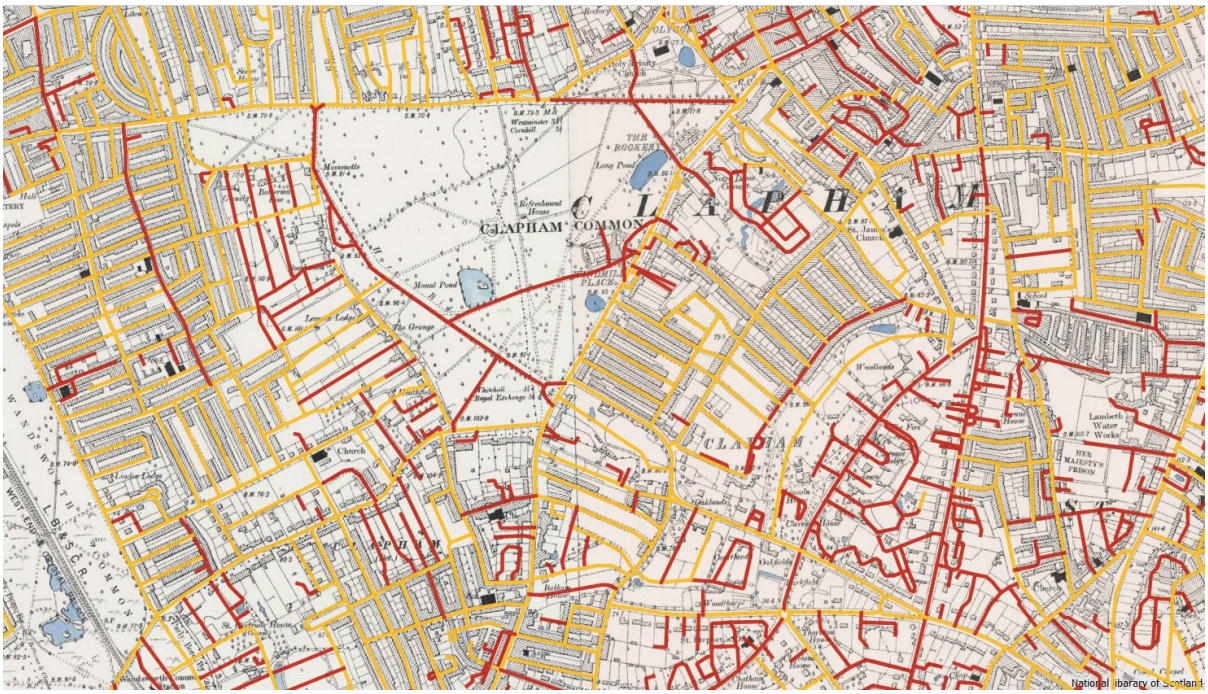

Figure 4 The street network of Clapham in 2016 overlaying a historic map from the early 1900s. (Source: Street network from the OS Open Road data; historic map web service by the National Library of Scotland; historic map data by the OS 1:1million – 1:10K, 1900s.)

5. Conclusions

In this study, we geo-reference addresses from the 1901 Census in England and Wales, by linking them to the contemporary OS AddressBase. The results show that it is feasible to geocode a large share of historic records down to address or street levels. Compared with rural parishes, urban parishes in 1901 exhibited better quality of address data and in many areas a large share of addresses have remained unchanged over a 100 years later. Future research will attempt to allocate the unmatched addresses to their most probable locations. Overall, this research provides an exciting opportunity to enhance geographic studies on the historic population. For the very first time since Charles Booth's pioneering poverty map created end of the Victorian era (see Booth, 1889) it is possible to link historic population databases to streets.

6. Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1 and EPSRC grant EP/M023583/1 ('UK Regions Digital Research Facility').

Biographies

Tian Lan is a Research Associate, working in the Geospatial Analytics and Computing group at the Department of Geography at University College London. He mainly works on contemporary residential segregation using Consumer Registers and the linkage of historical census records.

Guy Lansley is a Research Associate at the UK Consumer Data Research Centre and the Department of Geography at University College London. His research is primarily focused on harnessing geodemographic insight from big consumer datasets of unknown provenance.

Justin van Dijk is a Research Associate affiliated with the Urban Dynamics Lab and the Department of

Geography at University College London. His primary research interests are grouped around the analysis and visualisation of large-scale spatial data, urban mobility, and geographic information systems in general.

Paul Longley is Professor of Geographic Information Science at University College London and director of the UK Consumer Data Research Centre at UCL.

References

Booth, C., 1889. *Life and Labour of the people*. First Series (i) East, Central and South London. Macmillan, London (republished 1969).

Higgs, E. & Schürer, K. (2014) *Integrated Census Microdata (I-CeM), 1851-1911*, [data collection]. UK Data Service SN: 748. DOI: 10.5255/UKDA-SN-7481-1

Lansley, G. & Cheshire, J. (2018) Challenges to representing the population from new forms of consumer data. *Geography Compas*s, p.e12374.

Wall, R. (1982). Regional and temporal variations in the structure of the British household since 1851. In: T. Barker and M. Drake, (Eds.) *Population and society in Britain 1850-1980*, (pp 62-9). London