Georeferencing historical telephone directories to understand innovation diffusion and social change

Nikki Tanu[11], Maurizio Gibin[11] and Paul Longley[11]

[1]Consumer Data Research Centre, Department of Geography, University College London,
Gower Street
London
WC1E 6BT

February 14, 2021

Summary

This paper explores how historical archives of British telephone directories, recently made available, can be used to revisit historical and geographical questions using a modern, quantitative approach. By geolocating telephone subscribers throughout history, we seek to retrace spatial and temporal changes in demography and the spread of the fixed line telephone network in Britain. We develop a prototypical pipeline for capturing and processing these data using archives from 1881 and demonstrate that, if extended and applied to directories from the subsequent 100 years, this research can provide new insights into wide-ranging aspects of British social and economic history.

KEYWORDS: historical GIS, telecommunications, demography, innovation diffusion

1. Introduction

For most of recent British history, censuses that have been conducted since 1801 have been the most official – and often the only viable – way to know who lived and where they lived. The potential of using such data for quantitative analyses has been further enhanced with their linkage to Geographic Information Systems (Batty and Longley, 1996). Nonetheless, while comprehensive in their spatiotemporal coverage, these records provide only decennial snapshots and inevitably partial coverage of personal characteristics, such as income and occupation, that limit insights into major demographic changes in a given locality (Gray et al., 2018).

Often overlooked although they contain some similar information are phone directory archives. Holding information about telephone subscribers and where they were situated, these directories document the history of portions of the British populace as well as that of how the telephone exchange geography in Britain grew. While in 1880 the United Telephone Company offered phone services to around 200 subscribers using just three telephone exchanges, today over 200 exchanges serve London alone (British Telecommunications, 2021). Furthermore, telephone directories have always been updated much more frequently than censuses were undertaken. Nonetheless, the low extent of their digitisation, especially in comparison to census records, has remained an obstacle towards their utilisation in research.

The opportunity that these directories present now, then, is one for us to repurpose the information they hold as a novel source of socioeconomic data about Britain's past. Rather than merely a way to contacts others, they now also represent they hold to answer several longstanding questions. For one, fixed-line telephones were at their onset a luxury that few could afford (Casson 1910) and the inclusion of households or businesses in the earliest directories likely indicate areas where the wealthier resided. Secondly, the gradual transition of the telephone from a luxury to a mass consumer product can also be studied through the expansion in spatial coverage of telephone subscriber locations over time. Finally, the surnames of residential subscribers recorded in these archives could also lend themselves to the growing sub-field of surname analysis which can, among other purposes, indicate migration flows within Britain (Cheshire *et al.*, 2009).

2. Data

The historical phone directories dataset was acquired by the Consumer Data Research Centre (CDRC) through a collaboration with British Telecommunications (BT). In its raw form, the dataset is a collection of 1.6 million (1.2 TB) scanned images of BT phone directories from 1881 to 1984.

Spanning 104 years, the resource contains several issues published every year, missing only the 1893 edition. Naturally, the number of yearly issues increased proportionally with the growth in subscription, following the market penetration and saturation of landline technology. The quality and quantity of the scans varies throughout the year, with the print quality in images of the early editions being often poorer than in later editions.

2.1. Storage and Processing

The raw scanned images originally delivered on a portable drive were transferred onto a UCL secure server. The scans in *.j2k* format were converted to *.tif*, a format easier to manipulate and organise using Optical Character Recognition (OCR) software.

The processing pipeline is at an advanced stage of finalisation and the authors have developed a common set of procedures for all years and editions: image conversion, image preparation, OCR, text cleaning, geocoding and storing. Human intervention was needed to identify pages to be excluded from the OCR process, for instance pages containing only advertisements. The year 1881 was chosen to demonstrate proof of concept, not only because it is both limited in size and contemporaneous with a decennial census, but also because the results of digital encoding could be compared with recent published research on georeferencing historical sources by Lan and Longley (2020) – who develop an approach to chart historical changes in urban morphology and residential differentiation.

Currently, the pipeline is implemented using *command line interface* scripts, the Python and R programming languages and bespoke system libraries for OCR implementation. Currently, most software available on the market incorporates routines to optimize the input image and to specify the orientation of the characters on the page. Following some review and trials of available software, the open source OCR software Tesseract, developed by Google, was chosen for its large user base, availability of documentation, ease of automation, the amount of tuning parameters available and the convenience of licensing arrangements.

We attempted to fine-tune multiple parameters, both for pre-processing images for OCR input and the OCR engine itself. For the former purpose, more notable examples included cropping image borders and binarization (converting colour pixels to black and white pixels). During the trial, 70 multi-colour images were binarized and contrary to expectations based on previous trials (e.g. Gupta *et al.*, 2007; Reul *et al.*, 2019), we found that this worsened text recognition, likely because the 1881 directories frequently contained pencil strikethroughs in grey. With binarization, we supposed that the Tesseract engine was likelier to mistakenly detect these marks as part of the printed text, thereby causing more errors. Table 1

Table 1 reports the original naming convention and the number of issues for the 1881 edition of the phone directory, which covered only subscribers in present-day central London.

Table 1 BT Phone directory folder structure and issues for the year 1881

BT original box	Year edition and issues
bt_900008_box01_1881_jan_001	Jan 1881 - Greater London - London A - Z
bt_900009_box01_1881_apr_001	Apr 1881 - Greater London - London A - Z
bt_900010_box01_1881_jun_001	Jun 1881 - Greater London - London A - Z
bt_900011_box01_1881_jul_001	Jul 1881 - Greater London - London (Professions & Trades)
bt_900012_box01_1881_aug_001	Aug 1881 - Greater London - London A - Z
bt_900013_box01_1881_sep_001	Sep 1881 - Greater London - London A - Z
bt_900014_box01_1881_dec_001	Dec 1881 - Greater London - London A - Z

For the latter purpose, some parameters tuned pertained to the automatic detection of within-page columns, or ‘page segmentation’, and creating an error list of commonly mis-detected characters. Automatic page segmentation through Tesseract proved challenging because unlike many other historical documents which contain solid lines to separate columns, these were absent from most of the telephone directory archives, including the 1881 test images. Instead, we configured Tesseract to read each page as a unitary text block, from which we separated columns for telephone exchange number, name of owner and address fields by querying regular expression (RegEx) patterns emerging from these text strings.

2.2. Geocoding of Addresses

Thereafter, the resultant address field for each telephone directory entry was further segregated into subfields containing their street number, street name and postcode. This was, again, performed using regular expressions. Notably, many frequent misdetections of the characters occurring in London postcodes were accounted for (Table 2) to maximise the yield of address-linked postcodes for ease of geocoding addresses.

Table 2 Misdetections in postcode characters that were accounted for

Actual Character	Common Misdetections
S	‘5’, ‘8’
E	‘B, ‘F’, ‘H’, ‘13’, ‘18’
C	‘G’, ‘O’, ‘0’
.	‘,’, ‘ ’, ‘_’

We used fuzzy string matching to attach locational attributes to the extracted addresses. To achieve this, each entry’s street name was matched against all addresses in London in the digitised Census records of 1881, which had in turn been georeferenced using contemporary addresses in the Ordnance Survey’s AddressBase by Lan and Longley (2019). The AddressBase address that shared the lowest inter-string distance with each telephony directory address was then selected using an algorithm for georeferencing.

A limitation was that, while the Census record addresses are tied to historical administrative units, those in the telephone directories contain only their postcodes. Moreover, the postcode geographies of London had changed considerably in the decades before 1881 (The Postal Museum, 2021), thus complicating the process of matching these postcode areas to administrative units. Therefore, the postcode areas of each entry's address string were used to narrow the matching of these addresses to just the AddressBase addresses located in the registration districts that plausibly corresponded to the postcode area of the extracted address. This change resulted in a threefold decrease in computational time but also a higher proportion of address being matched to AddressBase addresses that were further away in terms of inter-string distance (**Table 3**).

Table 3 Accuracy of String Matching between Telephone Directory and OS AddressBase Records

Variation of Address Matching	**Number of Addresses Matched to**		
	Any address	Address with string distance <10	Address with string distance <5
Entire Dataset from Dec 1881	1,362 (100%)	-	-
Not restricted by registration district and postcode area	1,353 (99.34%)	1,305 (95.81%)	1,163 (85.38%)
Restricted by registration district and postcode area	1,353 (99.34%)	1,296 (95.15%)	1,116 (81.93%)

3. Discussion

At this stage, some preliminary results have been obtained regarding the georeferenced locations of telephone subscribers in the 1881 telephone directories which covered only London. Further, we attempted to classify the telephone subscribers into residential and commercial subscribers based on whether certain words were detected in the subscriber names. This list of words included not only generic markers like "& Sons" and "Limited" but also indicators of specific industries of commerce, such as "Bondholder" and "Dock". Figure 1 below shows the distribution of residential and commercial subscribers in all registration districts of London. As anticipated, a large proportion of subscribers were concentrated in the central districts and Figure 2 shows the same distribution within these districts.

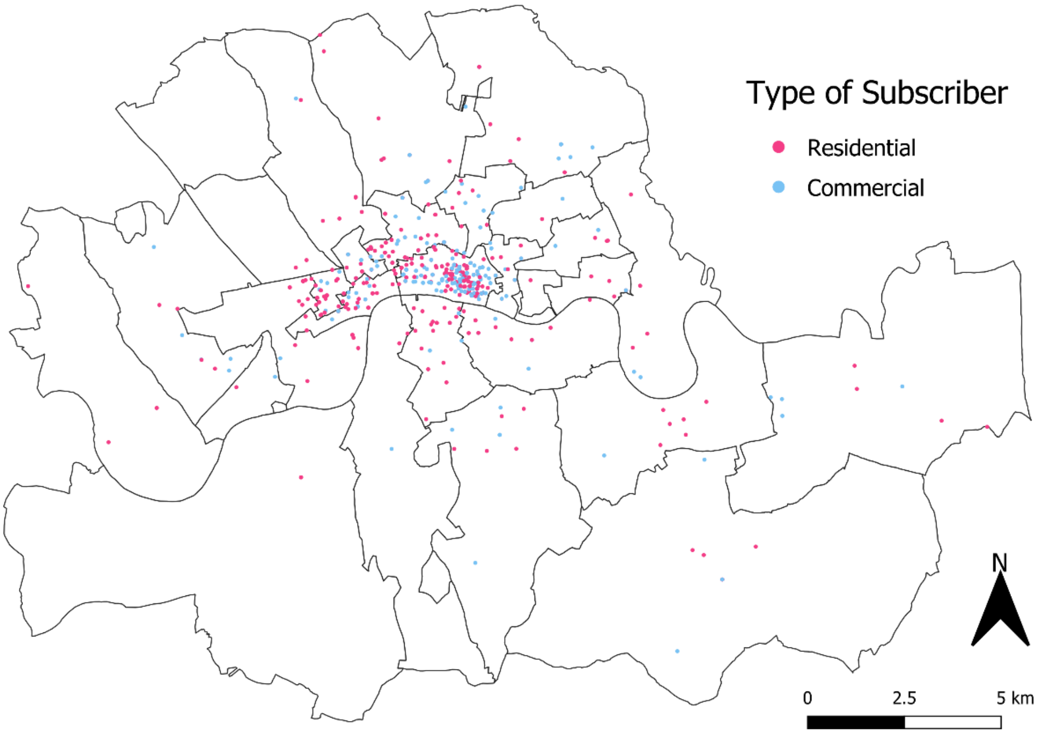

Figure 1 Most probable locations of telephone subscribers in London, 1881

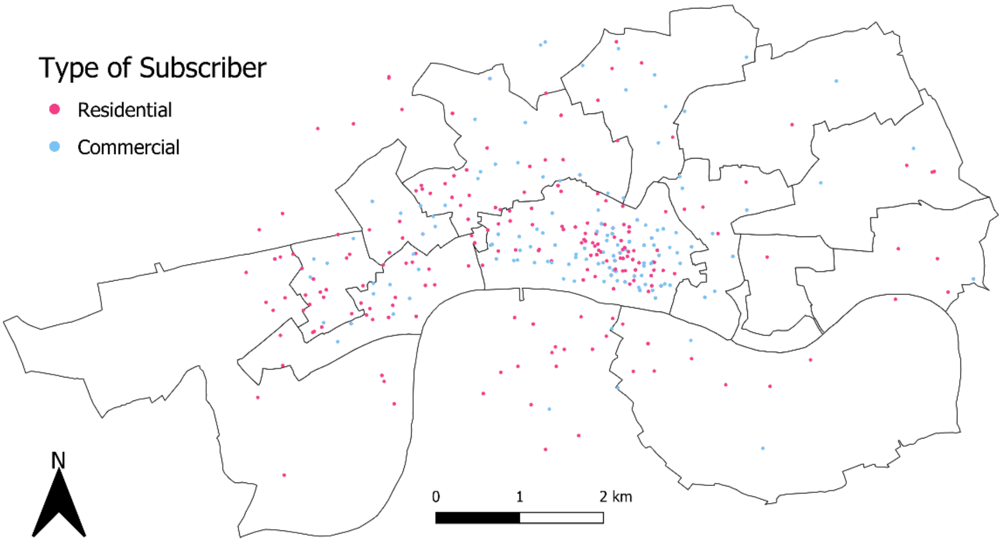

Figure 2 Most probable locations of telephone subscribers in London’s central districts, 1881

4. Future Work

Overall, this research has developed proof of concept that data can be retrieved from historical telephone directories to recreate parts of British history in relation to the development of telecommunications and of the people and places that it served. When used in tandem with the decadal census records, these archives have the potential to provide insights on how population dynamics across Britain changed temporally, especially in the years between censuses.

We also anticipate challenges when having to transfer the application of the data processing pipeline we have developed to telephone directory archives in later years as the page layouts change and the amount of data increases exponentially. Alongside the extent of areal coverage of the telephone in Britain. However, these factors have been considered while developing the current pipeline and we hope that, with appropriate changes, our analysis methods will remain robust for subsequent years of data and yield more insight into residential mobility, population change, industrial development and the history of British telecommunications.

5. Acknowledgements

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1 and an ESRC Ph.D. studentship that is co-funded by BT Archives. Scanned image data were provided by BT Archives under a collaborative agreement.

References

Batty M and Longley P (1996) Analytical GIS: The Future, in Batty M and Longley P (eds). *Spatial Analysis: Modeling in a GIS Environment*, Cambridge: Geoinformation International.

British Telecommunications (2021) 1605 to 1911 - The history of telecommunications - Our history, *www.bt.com*. Available at: https://www.bt.com/about/bt/our-history/history-of-telecommunications/1605-to-1911 (accessed February 2021).

Gray J, Buckner L, and Comber A (2019). Exploring Social Dynamics: Predictive Geodemographics. *AGILE Conference Proceedings 2019*.

Gupta M R, Jacobson N P and Garcia E K (2007). OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40, 389–397.

Lan T and Longley P (2019). Geo-Referencing and Mapping 1901 Census Addresses for England and Wales. *ISPRS International Journal of Geo-Information*, Multidisciplinary Digital Publishing Institute, 8(8), 320.

Lan T and Longley P (2020). Urban Morphology and Residential Differentiation across Great Britain, 1881-1901. *Annals of the Association of American Geographers* (In press)

Biographies

Nikki Tanu is a PhD student at the UCL Department of Geography under the supervision of Professor Paul Longley and Professor James Cheshire.

Maurizio Gibin is a Senior Research Fellow at the UK Consumer Daa Research Centre at UCL. His primary research interests are grouped around the capture, organisation, analysis and visualisation of geographically extensive datasets, including retail footfall, geodemographics, social media and historical archives.

Paul A. Longley is Professor of Geographic Information Science at UCL and Director of the UK Consumer Data Research Centre at UCL. His research focuses on the application of geographic information science, with a strong emphasis on the development and deployment of geo-temporal data infrastructures developed from Big and/or Open Data.