DIGITISING HISTORICAL TELEPHONE DIRECTORIES TO UNDERSTAND INNOVATION DIFFUSION AND HISTORICAL SOCIAL CHANGE IN BRITAIN, 1880-1951

Nikki Tanu

A thesis submitted in conformity with the requirements of Master of Philosophy (M.Phil.)

Department of Geography, University College London

DECLARATION

I, Nikki Tanu, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the work.

Signed

CONTENTS

List of Figures

LIST OF TABLES

ABSTRACT

Recent years have ushered in a trend of the increasing use of quantitative methods to study research questions across social science disciplines. However, their applicability has remained largely confined to most contemporary datasets, while excluding more dated data sources that were originally produced as hardcopies and, in many instances, not well preserved. This thesis thus presents an inquiry into whether newer quantitative methods could feasibly be applied on a large historical source of population data: the archives of telephone directories held by British Telecommunications. It details the development of a robust pipeline that transforms the raw data of image scans into structured tables of information, a format more amenable to newer techniques of quantitative analyses, alongside a documentation of quality control checks implemented and a discussion of key assumptions made. This is done with the help of various open-source software and accompanied by the creation of teldiR, a package of functions that would aid end-users of the dataset to both extend the work of and make improvements to data capture that has been completed. This thesis then explores some potential applications of this digitally encoded dataset for re-examining historical social science questions. Through quantification of subscription volumes in major settlements, comparisons of telephone adoption patterns can be made across space and time in urban Britain. Linking these data to information from the Censuses helps to further differentiate telephone subscription profiles in different settlements and augments an understanding of the provenance of the telephone subscription dataset as a source of population data. On the whole, the digitised dataset is able to provide coverage, albeit incomplete, of the historical population in Britain in a period where its granularity in combination with its breadth of coverage is unparalleled.

Impact Statement

The main impact of research undertaken by this thesis is represented by both the successful digitisation of large historical dataset for social science applications, as well as the rigorous methodology that has been developed for said purpose. First and foremost, the digitisation of the telephone directories data every 10 years between 1881 and 1951 is envisioned to add to the pool of available resources of historical social data. For academic research, it is hoped to facilitate more quantitative and broad-based studies of, among other topics, demography and urban growth in British history, as a complement to the existing spread of historical qualitative studies on these same topics. Moreover, the manner in which data encoding was done also allows for users with narrower interests to select subsets of the data to study based on spatial and temporal attributes. Beyond academic research, the potential userbase of these digitised data may encompass individuals or local historians who are interested in using smaller samples of the telephone directory records as a means to trace their personal or local genealogies.

The comprehensive pipeline that has been developed to digitise the data described above serves to further augment the possible use-cases of the dataset. This pipeline is comprised of numerous modularised functions, each serving a distinct purpose in organising the outputs of initial data capture into structured tables that are more easily analysed. However, the everchanging layouts of telephone directories and the limitations of time in this research meant that the pipeline had to be tailored to some directories, resulting in less accurate outputs for other directories. Made available online as 'teldiR', the library of functions used in the pipeline will enable end-users to make tailored improvements on the data capture already done on particular subsets of the records they are interested in. In this manner, it is also possible for them to extend data capture to directories that had not already been digitised, by adapting code written for another directory with a similar layout and making appropriate changes. Altogether, the pipeline extends the applicability of the telephone directories dataset by equipping users with the ability to refine existing data capture or extend its coverage, as they see fit for their purpose. On a broader scale, the successful creation of a semi-automated workflow to digitally encode such an expansive historical dataset might also signal the possibility for the employment of similar methods on other kinds of massive historical datasets that remain yet to be digitised.

ACKNOWLEDGMENTS

I would like to thank my supervisors, Professor Paul Longley and Professor James Cheshire, for their guidance and the expertise they have shared. I would also like to thank Maurizio Gibin and Richard Arnold for the extensive help they have rendered throughout this research. I am grateful to the friends and colleagues from the Consumer Data Research Centre, with whom the last few years have been made memorable, as well as the others with whom I have shared the conversations that were formative to this research. Finally, I would finally like to acknowledge the Economic Social and Research Council for providing the funding for this work.

1 INTRODUCTION

Traces of the collective histories of communities and, on a larger scale, societies can be found in a plethora of artefacts, records and other non-tangible remnants of the past which are orally transmitted, such as tradition and stories. While some of these media were produced with the very intention of preserving certain information for future generations, others were not purposed as such originally, but nevertheless came to serve as useful clues of what aspects of life were like in a previous era. Some examples of the latter sort of historical sources that have become commonplace both within and beyond academic research include photographs, speech transcripts, and written documents of all types. In fact, many of these sources continue to form the basis for how social, economic and political phenomena are currently studied, whether the period of interest lies in the past or the present. These means of understanding history have, until recently, been made possible using a variety of quantitative and qualitative research methods that mostly required intensive manual labour. In the case of the former, perusing large and sometimes long textual sources can yield detailed analysis of given themes as they relate to small localities. Because of the context-specificity of the conclusions reached, such time-consuming analyses may have to be repeated on similar sources from other localities before findings can be generalised to a wider context. Undertaking quantitative historical studies, on the other hand, required researchers to transcribe and digitise source material by hand, such as from hardcopies stored in archives or from the scans thereof. Studying the archives in this manner, research typically had to be organised such that either its scope was limited enough to become feasible timewise, or that it could draw on more manpower resources to cover a larger scope of digitisation.

In stark opposition to the conventions detailed above are more recent advances that aid quantitative research techniques. Made possible by sustained improvements in the capabilities of computer machines, such methods are able to handle large amounts of digital data at once, and run calculations on, summarise key patterns in and even visualise results in a time-efficient manner, considerably reducing the manual effort that has to be invested to help researchers arrive at a conclusion. In the current age of data, these methods synergise excellently with the large-scale datasets, or 'Big Data', that are prevalent and contain large amounts of granular information about large swathes of the population of a bounded space such as a city or country. Patterns of human activity, interactions with one another, and the influence of the environment can thus be studied extensively without the need for huge time investments from the researchers. As a longstanding contributor the subfield, Dennis (1991) similarly noted that historical geographers were largely excluded from advances that were had during the quantitative revolution of the 1960s. A plausible explanation is that these more recent methods have remained incompatible with historical data due to the lack of digitisation of the latter, most of which remain preserved in hardcopy formats. However, recent technology has begun to bridge this gap by making it possible to not just convert hardcopy documents into digital images but now also to extract the texts that are therein embedded. This shift heralds new opportunities for the application of new computer-assisted methods to datasets which are dated but which hold an immense wealth of information about people who lived in the past, with potential that can be unlocked if they can be converted into usable formats.

The telephone directory archive of Britain is one such resource, which also forms the basis for this enquiry. Following the inauguration of the public telephone service in the United Kingdom in 1879, the country's first telephone directories came into circulation in 1880. Thereafter, the expanding sequence of directories marks not simply the ever-increasing complexity and geographical coverage of the telephone network all through 1984, but also hints at the prevailing socioeconomic conditions which served as the backdrop for its service adoption. While fixed-line telephones gradually turned into a ubiquitous household possession during the course of the 20th century in much of the world, the U.K. included, increases in its adoption would vary greatly, favouring some people and places over others. A mix of both demand- and supply-related factors contributed to this variation, not least its affordability, the perceived need for its service by consumers, and the selective network expansion of telephone companies. Yet, in spite of the uneven representation of population sub-groups among its subscribers, the telephone would nonetheless achieve penetration rates among U.K. that likely rank among the highest of services and goods provided in the 20th century. Crucially, its services were offered in a format – where the information of every user was recorded and organised – which lent itself to conveniently doubling as a snapshot of increasingly large samples of the population at many points in history.

This thesis presents a study into the feasibility of harnessing the masses of information – contained in the archives of telephone directories – for revisiting historical questions in the social sciences. Access to the telephone directory archives, managed by BT Archives, was brokered through the Consumer Data Research Centre (CDRC), a big data initiative funded by the Economic and Social Research Council (ESRC). While these data have been licensed out to commercial organisations that deal with ancestry information, they have not been made available for purposes of academic reasons before, nor have they been digitised to an extent that they can be used for in-depth analysis. This data agreement thereby creates an unprecedented opportunity for exploring how such non-official 'consumer' data can be used to study historical populations, in contrast or supplementary to the official government sources – namely the national censuses – which have long maintained pre-eminence in producing population data with near-full geographical coverage at regular intervals in time. When successfully digitised, the telephone subscription data will enable quantitative research methods now common in social science to be applied to historical data in the way that they are often used to examine contemporary sources of big data. Successful digitisation may also prompt inquiry as to whether other non-official datasets with large population coverage could be similarly processed to enlarge the pool of data available for quantitative historical analyses. Ultimately, this is an enormous undertaking that comes with its unique methodological challenges, making it also an aim of the thesis to outline and justify the assumptions made, while also qualifying the conclusions that can be reached with the application of this new dataset.

1.1 Aims and Objectives

The overarching goal of this thesis is to explore the potential of the telephone directories dataset to yield new insights about the origins and growth of telephone usage in the UK, and this is split into two objectives that are sequential. The first objective is methodological and is a pre-requisite for any efforts put into the second, more substantive and exploratory objective to be at all effective.

The first of these objectives is to devise a common set of procedures that is able to transform the raw data into a format that lends itself to ease of analysis as conducted through computer-aided methods. As things stand, information cannot be obtained in a systematic manner from scanned images of telephone directory pages; texts in the original hardcopy directories must be extracted and given structure so that that they can be useful for researchers and other end-users with potential interest in the data. In turn, two end-products relating to this objective were envisioned. Firstly, this thesis aimed to develop a sophisticated processing pipeline that extracts textual information from the raw data and separates them into distinct fields based on their content. This pipeline must be simple enough to troubleshoot, such that this research and end-users alike may make modifications to it as they deem necessary, so as to ensure that it is robust to the ever-changing layouts of telephone directories through different editions. With such a pipeline developed, the thesis also strived to create and refine a dataset containing chosen samples of all available data that will be used to achieve the second aim. Implicit to the creation of these end-products is the need to detail assumptions made, and to outline the key sources of uncertainty involved in this process, so that the quality of the pipeline outputs may be better understood for future applications. The creation and refinement of the pipeline and the resultant digitised dataset was also iterative, whereby the former was incrementally updated to deal with ever emerging flaws that were detected in new samples of the data, as they were created.

The second objective was to demonstrate some potential avenues for further research using the digitised dataset created earlier. Broadly, the thesis pursued this goal by looking at the utility of the dataset through two different lenses: as data on the spread of the fixed-line telephone as a technology, and more implicitly as data about the population sub-groups that were at all represented in them. Seen as data about the spread of a technology, this research summarised macro-level variations in the patterns of telephone adoption across the major urban settlements of Britain and across time. Seen as data about the users of this technology, this research drew from existing research and other datasets to conduct exploratory analysis into the geodemographic provenance of the dataset, in order to answer the question: what kinds of people tended to be included in the list of subscribers, and where did they reside? Looking at the usefulness of the newly created data in this way can also open up the possibility for the quantification of demographic patterns in historical Britain when no other dataset that has both a comparable spatial coverage and such granularity exists. It is hoped that these initial inquiries will pave the way for future research to delve deeper into the possible applications of this data, whether independently or in conjunction with related datasets.

1.2 Outline

The thesis begins with this introductory chapter that, shortly hereafter, provides an overview of a history of the diffusion of the fixed-line telephone, both in its development into a universal technology and in its regional history more specifically to the British case. It discusses how the dataset of telephone directory archives, when functionally digitised, is envisioned to enable an in-depth, quantitative study of the telephone. This is then followed by a description of the raw dataset of telephone directory archives obtained from BT, upon which the bulk of this research is based, in Chapter 2. It proceeds to detail various steps of a comprehensive processing pipeline developed to transform this dataset into a digital format which lends itself to greater ease of subsequent usage and analysis. Even with the utilisation of modern technologies of text capture and recognition, among others, the quality of the outputs still suffers due to limitations innate to the raw data owing to their age, and to the wear and tear that comes with years of use of the source, hardcopy telephone directories. Major sources of uncertainty herein are discussed alongside strategies employed to mitigate their impact on the credibility of the pipeline outputs. By developing a proof of concept working on the telephone directories of London and thereafter broadening the pipeline's application to directories from elsewhere in Britain, this chapter concludes by demonstrating the possibility of digital extraction and encoding of the archival information. The results are structured records of historical telephone subscribers that, this research argues, has been digitised to a reasonable degree of accuracy, and which most definitely improves on the present dearth of granular information about Britain in the same time period.

Having produced preliminary data that are analysable, Chapter 3 explores how they can be used to study historical patterns of the spread of fixed-line telephony across Britain. Through a combination of manual annotation and probabilistic algorithms, more properties are added to the data, namely the attribution of subscriber records to a major urban settlement and the nature of their use of telephones – commercial or residential. These outcomes are then used to track how the telephone adoption rates varied from 1881 to 1951, between the 53 largest urban settlements in Britain as of 1901 as well as outline differences that exist on a broader level between the regions of England, Wales and Scotland. The rates of growth in different settlements are also discussed in tandem with the time at which they first appeared in the telephone directory archives. Altogether, this chapter showcases a straightforward application of the digitised dataset: to retrospectively quantify the diffusion of an important innovation in telecommunications in its earliest days of operation in Britain, and to visualise the variations across space that become apparent.

Chapter 4 seeks to extend the utility of the telephone directories dataset through a series of exploratory analyses that seek to exemplify how the newly digitised dataset could be used in future research, both in silo and in tandem with other sources of historical data. This is done, firstly, by examining the telephone subscriber records at an individual level and testing the feasibility of making linkages between these granular records, both within the dataset (across time) and between the datasets and that of the national Censuses. This was done to understand if traces can be established of a person who remains a telephone subscriber and resident at the same address over time and could broadly give a sense of how residentially stable or mobile the profile of telephone adopters in different locations are. It discusses difficulties in using the data at such a granular level, before then proceeding to exploratory analysis at the more aggregated level. By doing so, it looks to enhance interpretation of the telephone adoption patterns by providing local geodemographic context to subscribers, that hints at the different use-cases subscribers had for the telephone. This chapter therefore represents an initial foray into the possibilities that the newly digitised telephone directories bring for the analysis of historical Britain through the varied lenses of urban geography and demography, among others.

1.3 A Brief History of the Telephone

The fixed-line telephone was a transformative technology of telecommunications which was invented and patented towards the end of the 19th Century (The Economist, 1904). In the short term, it could be argued that telephones enabled people merely to do what they had done before – with the electronic telegraph system – more rapidly, albeit at a greater cost initially (Hamill, 2010). In fact, its invention by Alexander Graham Bell was a direct consequence of him trying to improve the telegraph; an understanding of the coming about of the fixed-line telephone is thus invariably linked to developments of the telegraph and its usage (Library of Congress, n.d.), something that will be discussed later in this section. Both modes of telecommunication shared a fundamental similarity in their principle of operation, with the telegraph system transmitting messages through electricity in the form of codes, while the telephone sent messages as sounds (ibid.). However, the improvements of the telephone over the telegram were obvious: where sending a telegram necessitated time and some deliberation, a telephone call could be made instantaneously and with much less difficulty (Kay, 2014). Moreover, the telephone, in its lifespan as a product, would far exceed the telegraph in terms of its penetration of populations in many countries, with just more than 1 in 3 British households having access to the telephone in 1970, nearly a decade after its invention (Post Office Telecommunications, 1970). For this reason, data on the usage of and subscription to fixed-line telephones have the potential to be much more revealing of demographic structures than similar data on the telegraph would ever be.

If the telephone subscriber data is to be used as a proxy for understanding aspects of the historical demography, the factors that influenced the uptake of the telephone among populations, and thus the factors that determine the biases of representation in the dataset, must be discussed. The first factor influencing the uptake of telephones by individual households is the urbanity of their place of residence. In this regard, economic studies on the relationship between urbanity and innovation have indicated that, the larger the population size of a settlement, the greater the degree of diffusion of the telephone tends to be (Calvo, 2006). As shown by the digitised data, the most populous settlements in Britain were indeed inclined to become served by telephone provision first; they moreover tended to have higher rates of telephone adoption than smaller population centres. This relationship also makes sense because telecommunication networks, including that of the telephone, were largely built on the perceived demand for those services. This perceived demand, in turn, was determined in no small part by the need to connect centres of population, much like how the transport links that preceded telecommunication links had grown out (Hamill, 2010). Furthermore, the telephone was also an innately social technology – where people subscribed to its services to contact others in their circles who had already possessed a telephone. Social links between people tended to be concentrated in the largest settlements, where on the whole there were more people, and so it was for good reason that the diffusion of telephone happened the most quickly in urban centres.

Another important factor that strongly correlated with telephone adoption and usage in its first decades of operation was social class. One obvious reason for this was that, monetarily, adopting a personal residential telephone would have been way out of reach for the working-class masses everywhere (Perry, 1977). This is not to say that only upper-class persons came into contact with telephones, as it was prevalent in Western European countries, including Britain, for them not to want to use the telephone directly. Instead, servants would be the ones using the telephone and passing messages on to their employers (Connected Earth, 2006); this implies that the telephone still received exposure outside the circles of the most privileged in society, although this exposure would still have been considerably less for those of lower classes. Another influence on telephone adoption, that was again interwoven with social class, was that of privacy concerns. From the onset, anxiety with regard to the potential of the telephone to intrude into personal lives had been rife, although these concerns tended to be the highest among those who were already on the fringes of society: the poor (Dutton, 1992). Conversely, the well-to-do could afford not only basic telephone services, but also supplementary services which made privacy significantly less of an issue, such as identification services and answering machine devices (ibid.). Together, these two reasons strongly suggest that early representation in the telephone directories would hint at a telephone subscriber belonging to the upper echelons of society.

A look at the historical telephone directories, as this research has gained access to, would thus be an exploration of how this unique corpus of names and addresses facilitates a quantitative examination of the diffusion of the telephone technology, through the late 19th and 20th Centuries. Networks of transport and communication – the latter of which the telephone falls under – emerge as a result of historical social processes (Sawyer, 2005) and therefore, a study of how telephones developed would also provide clues as to how society had been organised in the past. To this end, the usefulness of telephone subscription data as a source of population data has also largely been overlooked. Pool (1977), an academic whose research interests lay at the intersection of social sciences and information technology, asserted that the study of the telephone had been relatively neglected by social scientists as a whole, despite its importance as a mode of communication and despite its connections and similarities to other technologies. In the case of Great Britain, Perry (1977) went further to say that social and economic historians have 'minimised or ignored' (p. 69) the history of the telephone in Great Britain, citing an example of a study of the period 1870-1939 in which the telephone was mentioned just four times, while multiple pages were dedicated to a discussion of the railways, despite the fact that these technologies were both transformative and significant in their own rights. Dutton (1992) also wrote about the possibility of businesses using telephone numbers as data, on their own or in tandem with other demographic data, to profile their consumer bases and more effectively target their telemarketing campaigns, although noting that this strategy had not really been employed in any meaningful way. What these authors point to is the untapped potential of historical telephone data as an opening into the study of historical populations; this potential is precisely something that this thesis attempts to leverage on.

1.3.1 The Telephone and the Telegraph in Britain

Although generalisations can be made of some commonalities of the histories of the telephone around the world, as had been discussed above, the British case also presented peculiarities in the spread of the telephone. For much of the period from its introduction until at least 1912, when the British telephone industry was nationalised, the growth of telephony in Britain had trailed behind that in the US, the most prolific adopter nation of the telephone in the Anglosphere (McGuire, 2019). Much more so than in the American context, where telephones had been viewed as fit for 'all purposes' (Stein, 1996: 106), telephones in Britain in their early years were strongly skewed towards use by the aristocratic classes and businesses (ibid.). Concerns about the intrusion of the telephone into the domestic sphere also extended beyond those of privacy, as in Britain it was seen as a technology that was additionally socially impolite, even rude (Perry, 1977). However, perhaps the most considerable particularity of the growth of telephone networks in Britain was its commercial history of operations and the relationship between private companies in the industry with state regulators. For a meaningful discussion to be had of the history of the telephone in Britain, its links with its predecessor industry, that of the electronic telegraph, must be first understood.

The first electrical telegraph became functional in 1839, operating between Paddington in central London to West Drayton, with plans to lay undersea cables to connect Britain to France being formulated as early as 1840 (Hamill, 2010). By 1857, nearly every large town in Britain was connected to the electronic telegraph network (Marsden and Smith, 2008) and finally, in 1868, this industry in Britain was nationalised, with the government reasoning that it would enhance service provision by extending the coverage of the telegraph network to 'important Districts' that individual companies had yet to serve (Kieve, 1973). Quantitative data illustrating the use of the telegraph before its nationalisation are elusive (ibid.), and this pattern is repeated with the telephones, as data becomes more consistently accessible with its nationalisation in 1912, something that will be discussed in more depth later. Based on what data are available, following nationalisation of the industry, the volume of telegraphic messages sent in Britain increased steadily, rising from just shy of 10 million in 1870 to slightly over 90 million in 1900, at which point this figure peaked began to plateau (see Figure 3 in Mitchell, 2011). These golden years for the telegraph services in Britain were not to last, with its demand quickly falling as the telephone became more accessible, especially in the period after 1911 (Kieve, 1973).

In Britain, telephones were initially seen as a technological advancement that was not indisputably beneficial for the country, since it posed a competitive threat to the economic viability of electronic telegraphs – a state-controlled industry administered by the Post Office (Kingsbury, 1915). Because of this rivalrous relationship between the two modes of telecommunications, it has been widely argued that the Post Office and the Treasury actively strived to retard the expansion of provision of telephony in Britain (Perry, 1977), thus also providing an explanation for why rates of expansion of the telephone in other countries, especially the US, exceeded that in Britain. Evidence for this can be seen through actions taken and policies set by the Post Office in the 1880s. Within a year of the two pioneering telephone companies operating in Britain, the then-Postmaster-General had even filed information against these companies, receiving a favourable ruling from the High Court of Justice that the transmission of messages via telephone went against the telegraphic monopoly held by the Post Office (Hemmeon, 1912; Hamill, 2010). The United Telephone Company (UTC), having been formed in 1880 through the emerger of the two aforementioned companies, agreed to uphold the judgment of the High Court and pay a royalty of 10% on its earnings to the Post Office, in exchange for a license to expire in 1911, granted by the Postmaster-General to operate telephone services within a five-mile radius of central London (Hemmeon, 1912). As the Post Office also began to open telephone exchanges of its own, in other parts of the country, requests to establish an exchange that directly rivalled the offerings of the Post Office were generally also denied (Perry, 1977), thereby further slowing the early expansion of the British telephone network. The result of these policies was that, even as late as 1890, the telephone grid in England and Wales remained a series of largely disconnected regional networks, each having been built by companies that had licenses only to operate regionally (Robson, 2006).

As depicted above, the transition from electrical telegraphs to the telephone as the dominant mode of telecommunications in Britain was ridden with obstacles, not least due to the vested interests of the Post Office in slowing the advances of the telephone. However, the history of the development of the telegraph system in some ways mirrors the changes that would occur in the early years of the telephone services being provisioned, most notably in that, as one might say in hindsight, that the nationalisation of the telephone industry was inevitable, much like the fate of its predecessor. In the process of nationalisation, for both the cases of the telegraph (in 1868) and the telephone (in 1912), the Post Office had inherited service provision from different companies, each with separate regional networks; a key upside for interested researchers was that quantitative data for the respective systems also became much more reliably accessible as they underwent centralisation (Kieve, 1973; Hamill, 2010). Having touched upon the intertwined history of the two industries, this section will proceed onto a closer look at the history of British telephone companies, and how the services they provided were received by the public.

1.3.2 British Telephone Companies and their Reception

Historical records of general telephony and the services thereof, provided by the Post Office, are scarce (Crutchley, 1938) and the accounts of service provision prior to the industry's nationalisation in 1912 are even harder to come by, since there were many smaller telephone companies in existence (Hamill, 2010). Perry (1977), in his examination of the development of the telephone from 1876 to 1912, even labels this development a 'failure' and the period as 'the Years of Delay', referring to the assertion introduced earlier that the Post Office and the Treasury intentionally sought to delay the growth of telephone provision. This sub-section will seek to detail what information is available about the operations of telephone companies in Britain from the creation of the industry up till their takeover by the Post Office.

As two initially separate companies, the shareholders of The Telephone Company (Bell Patents) Ltd. And Edison Telephone Company of London Ltd. approved a merger of their companies to form the United Telephone Company Ltd. (UTC) in June 1880 (Freshwater, 2024). The UTC held the patents for important patents relating to the telephone, including those of inventors Alexander Graham Bell and Thomas Edison, and thus would go on to play an instrumental role in enabling the expansion of telephone networks in Britain, albeit within the restrictions set by the Post Office. The policy of the UTC was for their operations to be bound to London, but grant permissions for apparatus it held patents for to be used elsewhere in the country by subsidiary companies, thereby leaving to its subsidiaries the obligation to negotiate with the Post Office for provincial licences (Hemmeon, 1912); this stood in line with the telephone policy of the Post Office from 1880 to 1884, whereby the licenses it gave out to companies restricted their operations to small areas, so that their potential negative impact on earnings from the telegraph may be minimised (ibid.). By April 1885, the UTC reported having about 3,500 subscribers (in London) and made claims that among the ranks of its customers were the

Metropolitan Police, the London Fire Brigade in addition to a number of railway companies, newspapers and other smaller companies (Hamill, 2010). By the end of 1887, the UTC had lain 5,750 telephone lines (including both public exchange and private lines), accounting for 21.9% of all telephone lines in Britain, while comparatively the Post Office had lain only 1,370 telephone lines, or 5.2% of all British telephone lines (see Table on p. 269 in Kingsbury, 1915). In this period the UTC also claimed that other companies had successfully linked many Northern Towns including Liverpool, Manchester, Edinburgh and Glasgow (Hamill, 2010).

During this period of its introduction, control of the telephone service was also gradually handed over to the Post Office, in accordance to a High Court judgment that the telephone infringed on the telegraphic monopoly held by the Post Office (Hemmeon, 1912; McGuire, 2019). As a long-drawn process beginning 1880, this transition resulted in the full nationalisation and control by the Post Office of telephone services in 1912, except for the localities of Portsmouth, Hull and Guernsey, which remained independent for longer (British Telecommunications, 2007). However, the fragmented British telephone industry, full of companies operating at provincial levels, had already begun a process of consolidation long before 1912. While the Post Office refused to greenlight a proposed merger of the UTC and all their associated companies, it nonetheless continued on this trajectory at a smaller scale, merging the Lancashire and Cheshire Telephone Companies with the UTC in 1889 (Freshwater, 2024). In doing so, the UTC hoped not only to improve the attractiveness of its service offerings by connecting its subscribers to more parts of the country on its grid, but also reduce operational costs through economies of scale; even the telephone exchanges in less populated places could take advantage of their new connection to exchanges located in more bustling settlements (Robson, 2006).

While the costs of subscription to telephony services and their network coverage were important considerations for the public thinking of adopting a personal line – and certainly what telephone companies focused on improving –, the diffusion of telephones in Britain had also been limited by other pertinent factors. By 1912, the potential of the telephone to become a useful tool to the common person had not yet been reached, although much of the public was aware of both its existence and of the practicality it was said to have (Perry, 1977). Yet, prevailing social attitudes construed the telephone as something luxurious, rather than indispensable in daily life, and many people could not justify having one, even if they had the financial means to (ibid.). There was also disgruntlement with the quality of services provided among those who had already subscribed, with common complaints of delays in connections and long waiting times to be connected to the system when making a call – particularly so for those in suburban or outlying districts, whose calls had to be routed through exchanges in nearby, more major settlements – , among other issues

(Stein, 1996). Because the national telephone industry had been so fragmented prior to its nationalisation, problems also arose from telephone subscribers of one company wishing to contact subscribers under a different provider: in some cases this was more expensive and, in other cases, it was not possible at all (Hemmeon, 1912).

Obstacles to a more accelerated pace of growth of the telephone industry stemmed not just from the concerns of existing subscribers, as discussed above, but also from those of potential subscribers. An example of a less pervasive, but nonetheless pertinent issue, was that the telephone services were perceived to be insensitive to the needs of those who suffered from some degree of hearing loss (McGuire, 2019). In addition, those who used payphones (or 'public call offices') took issue with the difficulty of getting reimbursement for calls that did not go through, or broke off mid-way for unexpected reasons; this led to some potential adopters, commercial and residential alike, to doubt if having a telephone subscription would surely bring positive changes to their lives (Perry, 1977). These are but some of the demand-side issues that telephone faced in its early days, with concerns from a range of different groups in society, including existing residential telephone subscribers, those who used the public payphones and those who were yet to use the telephone. Nonetheless, they showcase how differences in the uptake of the telephone may have had been underlain by factors beyond simply the ability to afford the subscription.

In spite of the plethora of issues that faced telephone subscribers and potential subscribers in the first two decades, other evidence suggests that demand for the telephone had still been considerable, so much so that it was supply-side factors, and not those of demand, that was limiting its growth. At the 1904 annual meeting of the National Telephone Company – a separate entity from the UTC and its affiliates, which held the provincial licenses to operate in Scotland, the English Midlands and North of Ireland –, its general manager reported that in the waiting list for a telephone were approximately 3,300 people in London and another 7,700 elsewhere in the country (The Economist, 1904). This lack of ability of local telephone companies to cater to growing demand was also what led to a claim, made in 1907 by Glasgow and some smaller Scottish and English towns, that municipalities should be granted permissions to install their own telephone systems (Hemmeon, 1912). Even as late as 1914, there were still considerable delays between purchasing a subscription and having one's home connected to the telephone grid and this was as long as 51 days in Birmingham (Perry, 1977). In comparison to the American case, where the diffusion of the telephone was greatest in rural areas, telephones in Britain were strongly concentrated within cities, which may have contributed to this bottleneck situation in the largest urban settlements (Stein, 1996); the delays faced by potential customers in the suburban or outlying districts of major cities would often also be worse, since exchanges were built beginning in the central areas (Connected Earth, 2006).

Disparate accounts of the telephone industry in Britain have been pieced together above from what literature and other material could be located, covering a range of aspects of its development, including the histories of the British telephone companies up until their consolidation under the Post Office, attitudes of consumers to their services and operational issues which limited the supply of the telephone. While having a personal telephone line at home was still out of reach of most of the British population as of the nationalisation of the industry in 1912, the telephone had already very much become a part of public consciousness. 'Public Call Offices', or payphones, were established beginning 1884 and proliferated shops, hotels, post offices and railway stations, among other public locations, granting telephone access to a wider group in society who could afford it (ibid.). What more, the gradual increase in the numbers of overhanging telephone cables and poles – something private telephone companies were forced into doing due to their lack of statutory rights of wayleave – generated visibility for the telephone to the public, even if they had not had direct contact with its apparatus (Stein, 1996). The telephone would go on to expand its reach to many parts of society untouched by the telegraph - which was used overwhelmingly for business, and not private, matters – (Kieve, 1973), especially as costs for adopting one fell over time. The versatility of the telephone for a range of social applications would ultimately help the telephone turn into, over the course of a century, an important tool for business as well as an everyday object in British Households. What is offered by this research through an analysis of the archives of historical telephone directory, therefore, is a deep, quantitative dive into understanding the diffusion of this immensely transformative technology and its relationship with the people and places of old Britain.

1.4 Impacts and Applications

The potential of creating a historical dataset of the early telephone adopters in Britain as produced by this research is immense. For one thing, it charts the rise among British households of a then-novel innovation which fundamentally changed common modes of communication. Such data, although purposed in their original form primarily for users of the telephone to find other users, have in this research been repurposed to understand how service provision and adoption discriminated between people based on the characteristics of their locations of residence.

It also helps to fill the gap of available granular information about British individuals in the period after 1921, where national Census records at the individual level have not yet been made available for research, at the time of its undertaking. As time passes, greater proportions of the country's population become represented on the telephone directories as new subscribers; although this sample will scarcely be representative of the population as a whole even at its conclusion in 1981, it is argued that these data can nonetheless serve as a starting point for the quantification of demographic patterns in a time where no other dataset that features both national coverage and such granularity is available.

The field of potential for its application becomes even wider when linked with other datasets already used in research in related disciplines. These include those of economic history, local history and migration studies, which would similarly stand to benefit from the option to conduct quantitative and computer-aided research on pertinent samples of the telephone directories data. Local historians could utilise these data in conjunction with other sources of information about places of their interest to produce analyses that combines the advantages of using quantitative and qualitative analyses. This research has largely focused on residential subscribers, but by looking specifically at subscription records that were identified as commercial in nature, an economic historian could study a set of wholly different questions about industry and their changing relationships with this mode of telecommunication. Furthermore, having devised such a detailed pipeline for digitising the raw information and made available online a large proportion of the data processing tools, the work already done in this research can be easily extended to the years of data and for the locations for which records have not been digitised; researchers with interests in particular subsets, geographical or otherwise of the telephone subscription data, could also use the tools to further refine the results of text capture for these subsets.

2 The Telephone Directories Dataset

The first key contribution that this research makes to the study of historical geographies is a digitised dataset of unprecedented temporal granularity, detailing where and when considerable proportions of the British population lived. By establishing and iteratively refining a semi-automated data processing pipeline, the raw inputs – individual, scanned pages from telephone directories held by BT – are transformed into formatted tables of spatiotemporal information which lend themselves to analyses by computer-aided quantitative techniques. Because of imperfections in the inputs and the fallibility of algorithms employed in the pipeline, the outputs attain a respectable, albeit imperfect, accuracy in transposing their source material, which, this research argues, represents a massive improvement in the availability of data for the time period concerned. Moreover, these outputs are also important in facilitating the synthesis of specific research objectives that are written about in later chapters, seeing that the availability of data is what strongly determines the kinds of insights that can be derived in such data-driven approaches to the social sciences (Miller and Goodchild, 2015).

This chapter therefore outlines the decisions made and the challenges faced which are associated with this process and is structured more specifically. Firstly, the raw dataset and its structure will be outlined. Following this step, the various procedures for extraction of text from the image scans are elaborated, along with the reorganisation of these data into a table with fields containing distinct information. Interspersed within this approach will also be a discussion of the key choices and assumptions made, so as to ensure transparency and increase ease of replication or modification where a user so desires (King, 2011).

2.1 Data

Given the expansiveness of data contained in the telephone directory archives, this project has chosen to undertake an incremental plan of work for their digitisation. Attempting to strike a balance between breadth and depth, focus is placed on processing data contemporaneous to the annual censuses (on a ten-yearly basis, beginning 1881), with the assumption that the routines for data from these years would then be easily adaptable for processing those of neighbouring intercensal years. This process was subdivided into two phases.

2.1.1 Data Description

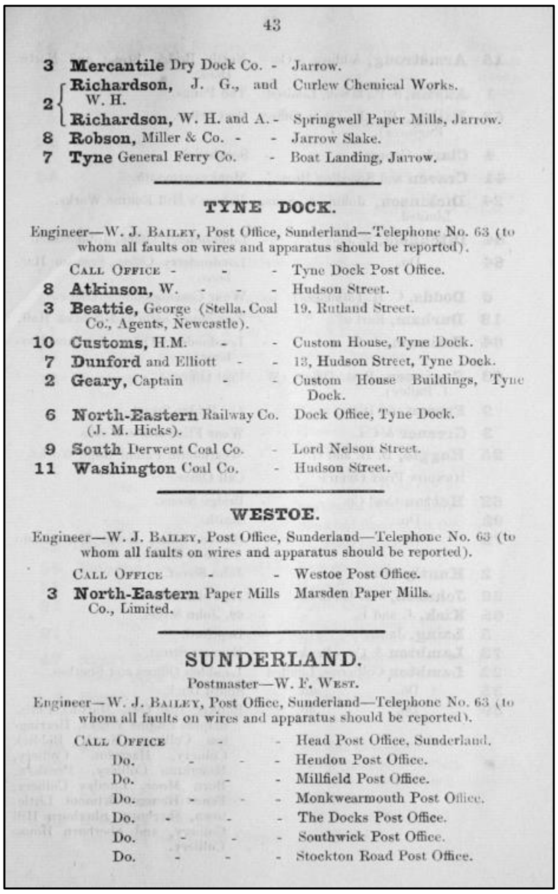
Access to the telephone directories dataset was procured by the Consumer Data Research Centre (CDRC) through partnership with British Telecommunications (BT) and this resource spans over a century, from 1880 to 1984, containing multiple issues published in each year within the timeframe apart from the missing 1893 editions. Comprising the raw dataset are 1.6 million scans of single pages of historical, hardcopy telephone directories that collectively occupy over 1.2 TB of storage space. Their geographical coverage spans most of England, Wales, Scotland and some parts of Ireland, although Ireland lies outside the frame of interest of this project and therefore. Of these, the focus for digitisation in this project will be the selected years when decadal Census took place in England, Wales and Scotland, meaning that, proportionally, an estimated 160,000 scanned directory pages were candidates for digitisation. Within this timeframe, both the number of yearly issues and number of records contained per issue increased proportionally with the growth in subscription, following the market penetration and saturation of landline technology (Hall and Das, 2017). The quality of the scans also varies across editions, with the print quality of images of the early editions being often poorer than in later editions.

Figure 1: Examples of the scanned directory images from 1901 for London (left) and for the North East of England (right).


When Calling give Name of Exchange as well as Number. 363 (L'dn.)

LONDON.

NAME OF EXCHANGE.			
Sydenham ..	84	**PEARCE** John ......" Haddon Lodge,"	W Dulwich
Avenue	1901	**PEARCE,** Pelly&Co.,IsinglassManufrs, Glucose Merchants	4 Colonial av
Avenue	1495	**PEARCE** R. F. & Co...Timber Merchts,	85 Gracechurch st
King's Cross	316	**PEARCE** Sarah E.	37 Hampstead rd
Avenue	2685	**PEARCE,** Sharpe & Co.,MetalMerchts, Brokers	150 Leadenhall st
Paddington	1381	**PEARCE** W. H. ....Builder,	1a Morshead rd W
East	221	**PEARCE** William & Sons Ltd.,Manufg Chemists	1 Bow common la
Holborn......	1617	**PEARCE** & Aldridge ....Solicitors,	1 Clement's inn
Hop	260	**PEARCE** & Conjuite.,Brewers'Merchts,	107 Southwark st
Avenue	895	**PEARCE** & Rowse ....Solicitors,	31 Liverpool st
Gerrard	1595	**PEARCE** & Sons ....Tailors,	19 Clifford st W
Holborn......	364	**PEARCE-JONES** & Co..................	33 John st W.C
London Wall	239	**PEARCY** E. D., Builder & Decorator,	Bayer stGolden la
Bank	282	**PEARD** & Son ...Solicitors & Comsnrs,	23 Budge row
Croydon ...	175	**PEARD** & Son ....Solicitors,	Union bank chas
Holborn......	1104	**PEARKS,** Gunston & Tee Ltd., Provn Merchants	6 Bayer st Golden la
Holborn......	1104a	**PEARKS,** Gunston & Tee Ltd., Grcrs, Provision Merchts (Warehouse)	Bayer st
Holborn......	1105	Ditto Provision Merchts,	6 Bayer st Golden la
London Wall	622	**PEARKS,** Gunston & Tee Ltd.	6 Bayer st E.C
Hop	373	Ditto Cheesemgrs & Grocers,	266 Brixton rd
Hop	26	Ditto (Stables),	2a Tranton rd
Hop	26a	Ditto Provision Merchts, Grocers,	254 S'hwark pk rd
Hop	1257	Ditto ditto	12 Peckham pk rd
Paddington	354	Ditto Provision Merchts,	306 Edgware rd
Kensington	637	Ditto Provision Merchts, Grocers,	477 King's rd Chelsea
Westminster	401a	Ditto ditto	4 Lr Belgrave st
Gerrard ...	5299	**PEARKS,** Gunston & Tee Ltd...........	65 Haymarket
London Wall	706	**PEARKS,** Gunston & Tee Ltd. (Dairy),	6 Pitfield st Hoxtn
London Wall	706a	**PEARKS,** Gunston & Tee Ltd.	22 Pitfield st
North.........	264	**PEARKS,** Gunston & Tee Ltd.	27 Dartmth pk av
King's Cross	577	**PEARKS,** Gunston & Tee Ltd.	44 Durham rd Finsbury pk
London Wall	1451	**EX-DIRECTORY**	
Westminster	401	**PEARKS** Henry....... Provision Dealer,	8 Lr Belgrave st
Streatham...	25	**PEARKS'** Stores, Grcrs, Provsn Mhts,	140 High rd
Kensington	664	**PEARKS** & Co....... Provision Dealers,	109 Gloucester rd
Gerrard	2585	**PEARS** A. & F. Ltd.	71 New Oxford st
Ealing	23	**PEARS** A. & F. Ltd.	The Soap wks Isleworth
Ealing	23a	**PEARS** A. & F. Ltd.	Ditto
Gerrard	5231	**PEARS,** Ellis, Pears & Brandreth, Solrs,	17 Albemarle st
Holborn ...	1639	**PEARSALL** Jas. & Co ....Silkmen,	71 Little Britain
Bank	5787	**PEARSE** Arthur L ......Mining Engr,	Worcester ho Walbrook
Gerrard......	3216	**PEARSE** J. & B. & Co., Army Cntctrs,	28 Floral st W.C
London Wall	1669	**EX-DIRECTORY**	
Avenue	2973	**PEARSE** L. F. & Co. ...Tea Brokers,	St Dunstan's ho
Gerrard	4065	**PEARSON** C. A. Ltd........Publishers,	17 Henrietta st
Kensington	956	**PEARSON** C. Arthur	32 Prince's gate
Holborn......	1709	**PEARSON** E. J., Harness Manu-facturer, Contractor	45 St John st rd
Bank	5887	**PEARSON** FireAlarmSystemLtd.,The,	32 Watling st
London Wall	1028	**PEARSON** Fire Alarm System Ltd., (Call Station)	55a Whitecross st
Avenue	641	**PEARSON** J. A. ...Potato Salesman,	59 Brushfield st
Ealing	78	**PEARSON** Jas. C., (Heston Stud Fm),	Kirbyldge Heston
Hop	1538	**PEARSON** H. T. & Co. ....Leather Factors, Importers	72 St Thomas' st
London Wall	92	**PEARSON,** Huggins & Co. Ltd., Uniform Clothing Contractors	51 Scrutton st


43

3	**Mercantile** Dry Dock Co. -	Jarrow.
2	**Richardson,** J. G., and W. H.	Curlew Chemical Works.
	Richardson, W. H. and A. -	Springwell Paper Mills, Jarrow.
8	**Robson,** Miller & Co. - -	Jarrow Slake.
7	**Tyne** General Ferry Co. -	Boat Landing, Jarrow.

TYNE DOCK.

Engineer—W. J. Bailey, Post Office, Sunderland—Telephone No. 63 (to whom all faults on wires and apparatus should be reported).

	Call Office - - -	Tyne Dock Post Office.
8	**Atkinson,** W. - - -	Hudson Street.
3	**Beattie,** George (Stella Coal Co., Agents, Newcastle).	19, Rutland Street.
10	**Customs,** H.M. - -	Custom House, Tyne Dock.
7	**Dunford** and Elliott - -	13, Hudson Street, Tyne Dock.
2	**Geary,** Captain - -	Custom House Buildings, Tyne Dock.
6	**North-Eastern** Railway Co. (J. M. Hicks).	Dock Office, Tyne Dock.
9	**South** Derwent Coal Co. -	Lord Nelson Street.
11	**Washington** Coal Co. -	Hudson Street.

WESTOE.

Engineer—W. J. Bailey, Post Office, Sunderland—Telephone No. 63 (to whom all faults on wires and apparatus should be reported).

	Call Office - -	Westoe Post Office.
3	**North-Eastern** Paper Mills Co., Limited.	Marsden Paper Mills.

SUNDERLAND.

Postmaster—W. F. West.

Engineer—W. J. Bailey, Post Office, Sunderland—Telephone No. 63 (to whom all faults on wires and apparatus should be reported).

Call Office - -	Head Post Office, Sunderland.
Do. - - -	Hendon Post Office.
Do. - -	Millfield Post Office.
Do. - - -	Monkwearmouth Post Office.
Do. - - -	The Docks Post Office.
Do. - - -	Southwick Post Office.
Do. - - -	Stockton Road Post Office.

Figure 1 shows two examples, among the many possibilities, of how the scanned directory pages could look like. The differences between them provide clues to some of the challenges faced and considerations that had to be made in this research; they will first be briefly highlighted here before discussed in more depth as they become relevant later in this thesis. Firstly, the two sample pages vary considerably in terms of the density of information they contain, with the page on London *(Figure 1, left)* being just one of numerous pages listing the city's subscribers, while some settlements in the North East of England barely have ten subscribers in 1901 *(Figure 1, right)*. Such geographical disparities characterise patterns of telephone adoption throughout the timespan of concern in this research. Secondly, the images differ typographically, in terms of conventions of what information is capitalised and/or bolded, and which punctuations are mainly used to fill empty spaces on a line. As such, when trying to capture information from different directories, this research had to create slightly customised routines for each directory/ year of directories, which incorporate such oddities. Finally, the information they contain and the order in which this information is presented – notably with the inclusion of telephone exchange names for the London page, but not for that for North East of England – constantly change with time and across regions. Therefore, as aforementioned, data processing routines also have to take into account the particularities of how every directory presents its constituent information.

The raw scanned images, originally delivered on a portable drive, were transferred firstly onto a secure server managed by the CDRC at UCL and, thereafter, duplicated and stored on UCL's secure online data storage platform, DataSafeHaven (DSH). Samples of files are hosted locally on the researcher's personal computer to facilitate the writing of programming to code to process and analyse the data. Because this dataset contains digitised versions of resources which, in history, had been openly distributed to the public, the data are not considered to be sensitive. Nonetheless, appropriate measures have been undertaken to prevent the access of data obtained by anyone unaffiliated to this research.

Table 1: Examples of the different types of telephone directories that were made available by BT Archives to this research

BT roll name	Type of content
'bt_900008'	**Typical telephone directory.** Lists all telephone subscribers, organised alphabetically, from one major city or region.
'bt_900011'	**List of Professions & Trades.** Only lists subscribers which are offering their professional services, organised by their trade.
'bt_900591' *(A-H)*	

'bt_900592' *(H-Z)*	**Multi-part telephone directory.** In some cases, a locality has so many subscribers they have to be split across multiple directories, also alphabetically. This mostly occurs for London directories.
'bt_900789'	**Regional directories** – beyond major cities, most other records are combined and listed by regions, which loosely correspond to historic county delineations.

Table 1 provides an illustration of how the data in their raw format were handed over to this research and highlights notable variations in the types of directories that exist, which in turn practically implicate upon the data processing. The scans of individual pages of each individual directory were collated into 'rolls' and rolls that were related to each other (typically by chronology) were then grouped into boxes. The naming of every roll thus importantly includes a unique six-digit identifier that is always preceded by '*bt_*', as well as further information on the time of publication. While the dataset up till 1894 contains only information on subscribers from the Greater London area, national and regional telephone directories emerge thereafter, containing records for subscribers across England, Wales and Scotland and later also Northern Ireland, the last of which this research does not concern itself with. In addition, Table 2 below summarises the eventual results of data capture – a process that will be elaborated on later in this chapter – from the telephone directory scans. The results are also partly indicative of how volumes of telephone subscription in Britain grew in the period spanning 1881 to 1951.

Table 2: Numbers of unique directories and subscriber records that were obtained from the processing pipeline in each decade from 1881 to 1951

Year	Unique Directories	Total Number of Records
1881	7	12,847
1891	3	32,471
1901	5	46,064
1911	11	717,745
1921	7	645,811
1931	11	2,199,345
1941	8	1,719,772
1951	10	2,760,380
Total	62	8,134,435

2.1.2 Data Strategy

In the first phase, this research developed a proof of concept that demonstrates the viability of the routines of data capture and cleaning proposed. This focused on telephone subscriber records from London between the years 1881 and 1911 (inclusive), where the volumes of records yearly are relatively manageable for analysis but also because of the availability of complimentary data in this period. Between 1881 and 1901, digitised individual records from the national censuses of England, Wales and Scotland are available as provided by the UK Data Service through the I-CeM Project (Schurer and Higgs, 2023). By focusing on these years, the accessible Census data as a de facto 'gold standard' against which telephone subscriber records can be compared, in order to ascertain its coverage of the population. In addition, work that has been done into variations in the demographic make-up of different settlements in Britain by Lan and Longley (2021) provides further potential for cross-referencing the data and contextualising the changes in telephone adoption patterns that may be studied through the newly digitised telephone directory data. As an additional year of data to process, 1911 was seen as a sensible cut-off point just before 1912 when automatic telephone exchanges, with capacities to accommodate much larger call volumes than their manual counterparts, were introduced (Post Office Telecommunications, 1970). Logistically, in these early years of telephone operation in the UK, the geographical coverage of the dataset (and thus its volume) was relatively small; the resultant shorter computing times made feasible a trial-and-error approach to developing the pipeline and allowed for the incremental inclusion of improvements, as they were found necessary, without expending excessive computing power and time.

With this focal time period, issues of privacy as defined by the General Data Protection Regulation (GDPR) are also not expected to be a concern as the GDPR protection does not extend to the data of deceased persons: telephone subscribers recorded in these early directories are likely to be deceased as of 2021, given that more than a hundred years have passed since 1911, while mean life expectancy at birth in the UK, which is nearly as high as it ever has been, was 79 and 83 years for males and females, respectively (Office for National Statistics, 2021). Geographically, data capture in this phase focused on London not only because it is the only region for which telephone records are reliably available within this duration, but also that it maintains primacy in the uptake of fixed line telephony through the years of service provision recorded in the archives. Even as telephone adoption rates in major cities other than London began to climb steadily, they remained dwarfed by the numbers for London; it is not until the late 1890s that any other city has its records in a standalone section of the telephone directories. All in all, London proved most suitable for the proof-of-concept work, and given the trickle-down nature of the spread of telecommunications technology (Mahler and Rogers, 1999), it was assumed that patterns of telephone uptake in London would roughly indicate how the same phenomenon would unfold elsewhere in Britain later.

In the second phase, which expanded processing to years of records beyond 1911, data capture proceeded likewise chronologically at decadal intervals, beginning with the most recent London directory of the year and followed eventually by all other available directories. Everchanging directory layouts necessitated continual modification and, sometimes, major changes to the processing pipeline, as fitting to the particularities of the structure of records in directories in the year. To cater to this issue, over time, code written for data capture was developed into a series of generalised functions that are able to adjust themselves to varying applications with but simple changes in their inputs. These functions are detailed in Section 2.2.3 and are also envisioned as a tangible output of this research, with the hope that other research looking to make sense of historical data can similarly employ them.

2.1.3 Organising the Raw Data

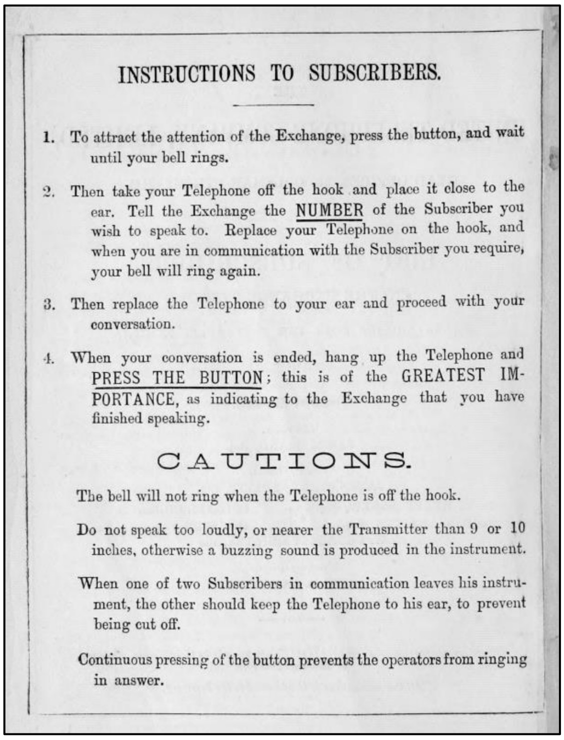
As a final step of preparation before the data were to be processed, the scanned images were sorted based on the type of content they held. Because the primary objective of developing a digitisation pipeline was to enable a largely automated extraction of information of two pieces of information – the registered names of telephone subscribers and their addresses –, some kinds of pages held only information irrelevant to this. This sizeable minority of pages, nevertheless, hold other important information that augment the interpretation of data by providing context. Therefore, prior to its implementation, the pages of directories in the years concerned were manually categorised by their content to facilitate later batch processing. With the enormity of the task at hand – involving the perusal of about 160,000 pages in total – this work was supported by other members of the CDRC research group. Table 3 below details the categories and their respective significance, while Figure 2 gives an example page for each category (excluding 'Blank' and 'Subscriber Records', an example of which has been shown in Figure 1).

Table 3: Categorisation of directory pages prior to pipeline implementation

Page type	Information contained
Advertisements	Full-page advertisements of products or services, containing text that is often accompanied by images
Blank	-

Introductory	Introductory information for telephone directory end-users, for instance "How to use this directory" guides
Maps	Maps (typically hand-drawn) of telephone service provision, such as the location of telephone exchanges and call offices
Subscriber Records	Alphabetically or spatially organised lists of telephone subscriber names, telephone numbers and addresses
Telephone Exchange Information	List of addresses of the main telephone exchanges from which telephony services are provided
Trades Lists	Telephone subscription information organised by profession rather than alphabetically or by subscriber location

Figure 2: Examples of a page from each category of directory content. Top row: Advertisements (left), Introductory (right); middle row: Maps (left), Telephone Exchange Information (right); bottom row: Trades Lists.



INSTRUCTIONS TO SUBSCRIBERS.

1. To attract the attention of the Exchange, press the button, and wait until your bell rings.

2. Then take your Telephone off the hook and place it close to the ear. Tell the Exchange the NUMBER of the Subscriber you wish to speak to. Replace your Telephone on the hook, and when you are in communication with the Subscriber you require, your bell will ring again.

3. Then replace the Telephone to your ear and proceed with your conversation.

4. When your conversation is ended, hang up the Telephone and PRESS THE BUTTON; this is of the GREATEST IMPORTANCE, as indicating to the Exchange that you have finished speaking.

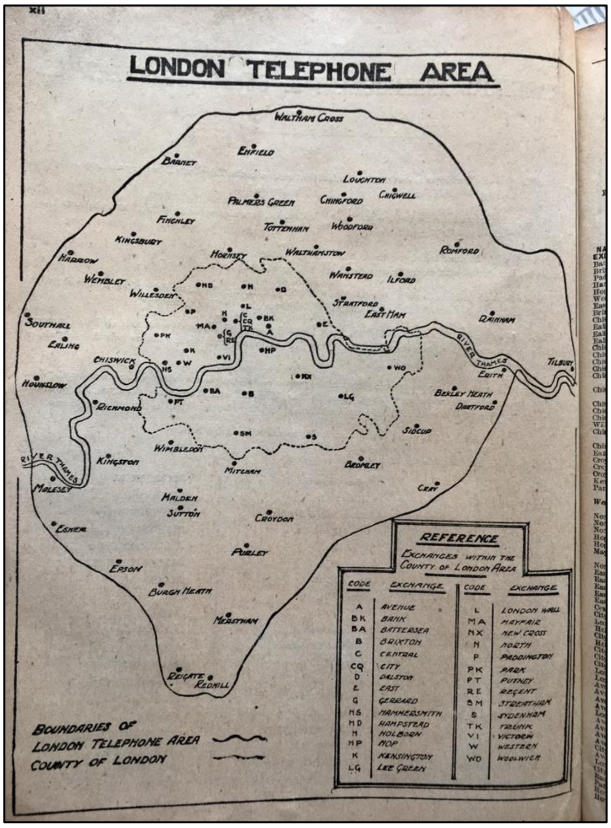
CAUTIONS.

The bell will not ring when the Telephone is off the hook.

Do not speak too loudly, or nearer the Transmitter than 9 or 10 inches, otherwise a buzzing sound is produced in the instrument.

When one of two Subscribers in communication leaves his instrument, the other should keep the Telephone to his ear, to prevent being cut off.

Continuous pressing of the button prevents the operators from ringing in answer.



CLASSIFIED TRADES LIST

For prices and particulars of insertions in this section of the Directory, application should be made to the National Telephone Company's Sole Advertising Contractors, Messrs. Sell's, 168, Fleet Street, London, E.C.

METROPOLITAN PROVINCE.

	TELEPHONE NUMBER.
ADVERTISING CONTRACTORS—	
VICKERS JAMES WM., 5 Nicholas Lane, Lombard Street, E.C, Advertisements received for all Newspapers, etc.	**London— Avenue ... 1303**
ASBESTOS MANUFACTURERS—	
UNITED ASBESTOS CO. LTD., Dock House, Billiter Street, E.C. (*See Advt.*)	London— Avenue **2872**
AUCTIONEERS—	
DOD FRANCIS & CO., Surveyors, Land and Estate Agents, 59 High Street, Stoke Newington, N.	London— Dalston **278**
MOORE C. C. & T., 7 Leadenhall Street, E.C.	London— Avenue......... **335**
BIO-TABLEAUX—	
GIBBONS WALTER, The Bio-Tableaux, 60 Chandos Street, Strand, W.C.	London— Gerrard **3828**
BELTING MANUFACTURERS—	
COBBETT W. WILLSON LTD., 59 Southwark Street, S.E. (*See Advt.*)	London— Hop **25**
BLOTTING PAPER MANUFACTURERS—	
FORD T. B. LTD., Snakeley Paper Mill, Loudwater, High Wycombe. (*See Inset.*)	High Wycombe— Bourne End... **10**
BOLT, NUT AND RIVET MANUFACTURERS—	
PERIAM HERBERT W. LTD., Floodgate Street Works, Birmingham. (*See Advt.*)	Birmingham **373**
BOOT AND SHOE MANUFACTURERS—	
POCOCK BROS., 235 Southwark Bridge Road, S.E.	**London— Hop......... 271**
CURRYER BROS., 69 Oxford Street, W. (*See Advt.*) ...	
BRASS TUBE AND METAL MANUFACTURERS—	
HUDSON, EDMUNDS & CO. LTD., Sheepcote Street, Birmingham	Birmingham **80**

Met.—17

In case of FIRE call for FIRE BRIGADE. No Number required.

CALL ROOMS—*Continued.*

N.W.—*Continued.*

7608	**Kentish Town** . .	H. Conolly, 120, Kentish Town-road.
7403	**Kilburn** . . .	Thos. Bate, Red Tower, High-road.
7448	**Ditto** . . .	Rickett, Smith & Co., West End-lane Station.
7450	**Ditto** . . .	Rickett, Smith & Co., London & North Western Station.

S.E.

9601	**Anerley** . . .	Rickett, Smith & Co., Anerley Station.
4711	**Bermondsey** . .	London Grist Mills, East-lane-wharf.
4857	**Ditto** . . .	Bermondsey Town Hall, Spa-road.
9635	**Crystal Palace** .	Rickett, Smith & Co., Crystal Palace Station.
4664	**Denmark-hill** . .	Ham & Son, 71, Denmark-hill.
8010	**Deptford** . . .	George Collins, 44 & 45, Broadway.
9603	**Forest Hill** . .	The Supply Co., 15 & 17, Dartmouth-road.
9634	**Gipsy Hill** . .	Rickett, Smith & Co., Gipsy Hill Station.
8026	**Greenwich** . .	H. Richardson, 4, Church-street.
8045	**Lewisham** . .	John R. Houlgate, 62, High-street.
4799	**London Bridge** .	Bridge House Hotel.
9368	**Norwood** . . .	Rickett, Smith & Co., Norwood Junction.
9341	**South Norwood** .	Chas. Norwood, "Norwood News," Station-rd.
4823	**Southwark** . .	National Telephone Co., Ld., 3, Southwark-street.
9605	**Sydenham** . .	Sydney Smith, 26, Kirkdale.
9632	**Ditto** . . .	Rickett Smith & Co., Sydenham Station.
9639	**Ditto** . . .	H. Frost, 37, High-street.
9630	**Upper Norwood** .	W. E. Nicholls, 57, Westow-street.
4810	**Walworth** . .	France & Sons (Elephant & Castle Auction Rooms), Draper-street.
8415	**Woolwich** . .	Jayne & Co., 30, Hare-street.

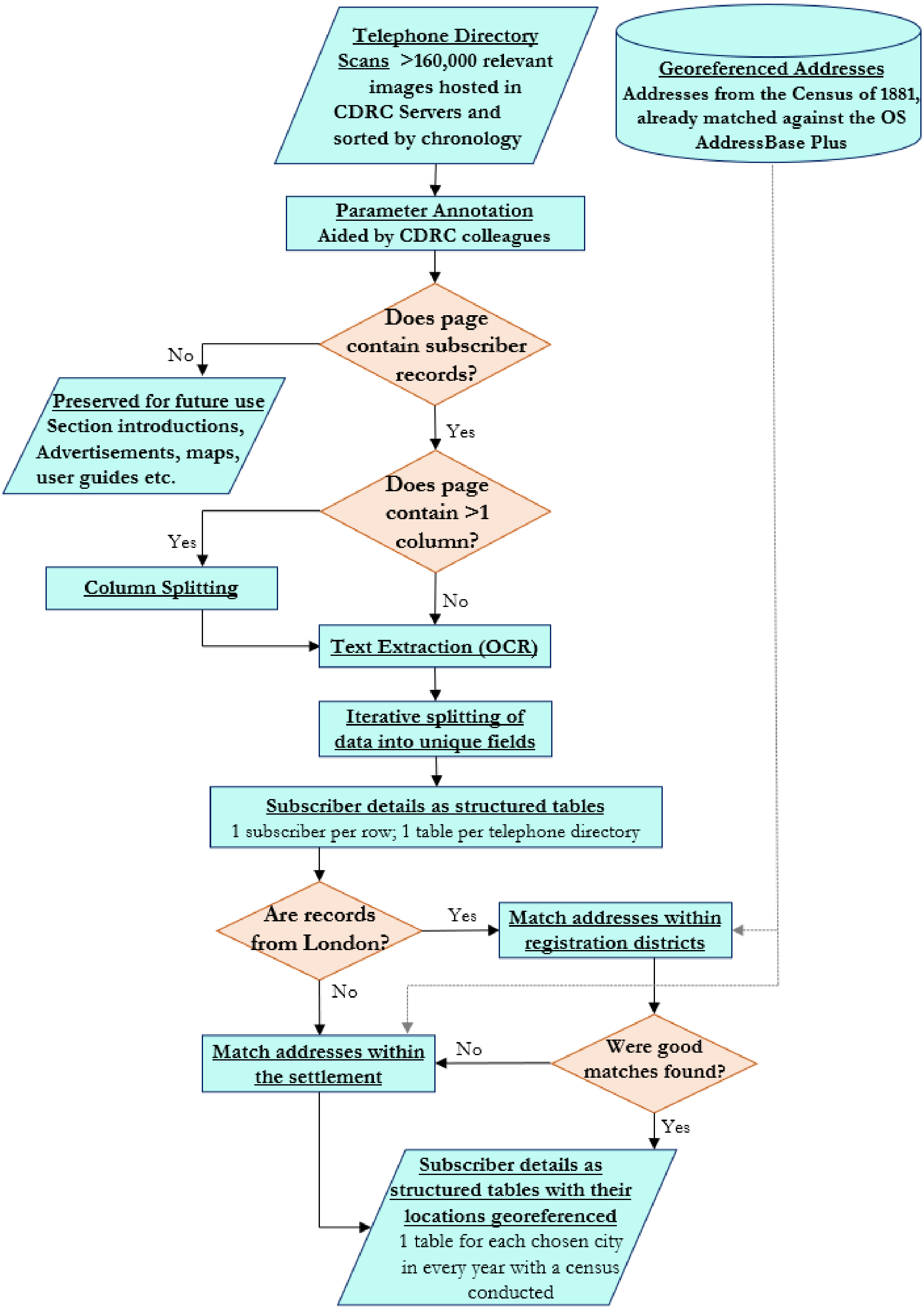
S.W.

9202	**Balham** . . .	James Stevens, 8, Wawman-terrace.
9125	**Brixton** . . .	Clare & Sons, 242, Brixton-hill. [Promenade.
9114	**Ditto** . . .	South Metropolitan Dairy Co., 13, Grand
8562	**Brompton** . .	Moore Bros., 81, Brompton-road.
8714	**Chelsea** . . .	Salmon & Gluckstein, 86, King's-road.
9113	**Clapham** . . .	Dunkley, 101, High-street.
9146	**Ditto** . . .	Rickett, Smith & Co., Clapham-road.
9062	**Clapham Junction** .	Jayne & Co., 71, Falcon-road.
8752	**Earl's Court** . .	German Exhibition.
8617	**Earl's Court-road**	Morley J. W. & Letts, 185, Earl's Court-road.
35052	**Haymarket** . .	National Telephone Co., James-street.
3173	**House of Commons**	**Upper Committee Rooms, Westminster.**
3212	**Pimlico** . . .	Rickett, Smith & Co., Victoria-wharf, Grosvenor-road.
9011	**Putney** . . .	A. W. Taylor & Co., 119, High-street
8905	**Richmond** . .	Ernest Pennington, Grosvenor-buildings.
35006	**Soho** . . .	Breidenbach & Co., 48, Greek-street.

2.2 Data Processing Pipeline

Having described the structure of the data in the previous section, this section proceeds to detail the processing pipeline that has been developed, through which information from the raw dataset is gradually transformed into a tabular format that lends itself to convenience of analysis. This pipeline encompasses a set of procedures common to all editions and its constituent steps are, broadly listed: image conversion, image preparation, Optical Character Recognition (OCR), text cleaning and separation into different fields and, finally, geocoding. While some factors that negatively impact the quality of outputs - such as the formatting of the directories and quality of scans – are non-mitigable, this research has invested a great deal of effort in maximising the robustness of the pipeline in the face of such inconsistencies. Hereafter, each sub-section will outline a single step in the pipeline, the considerations that were had in its development and why the final procedure was ultimately selected over viable alternatives, if any. An overview of the methodology of digitisation can be seen in the flowchart that is Figure 3.

Figure 3: Flowchart of overall methodology in the digitisation of archives



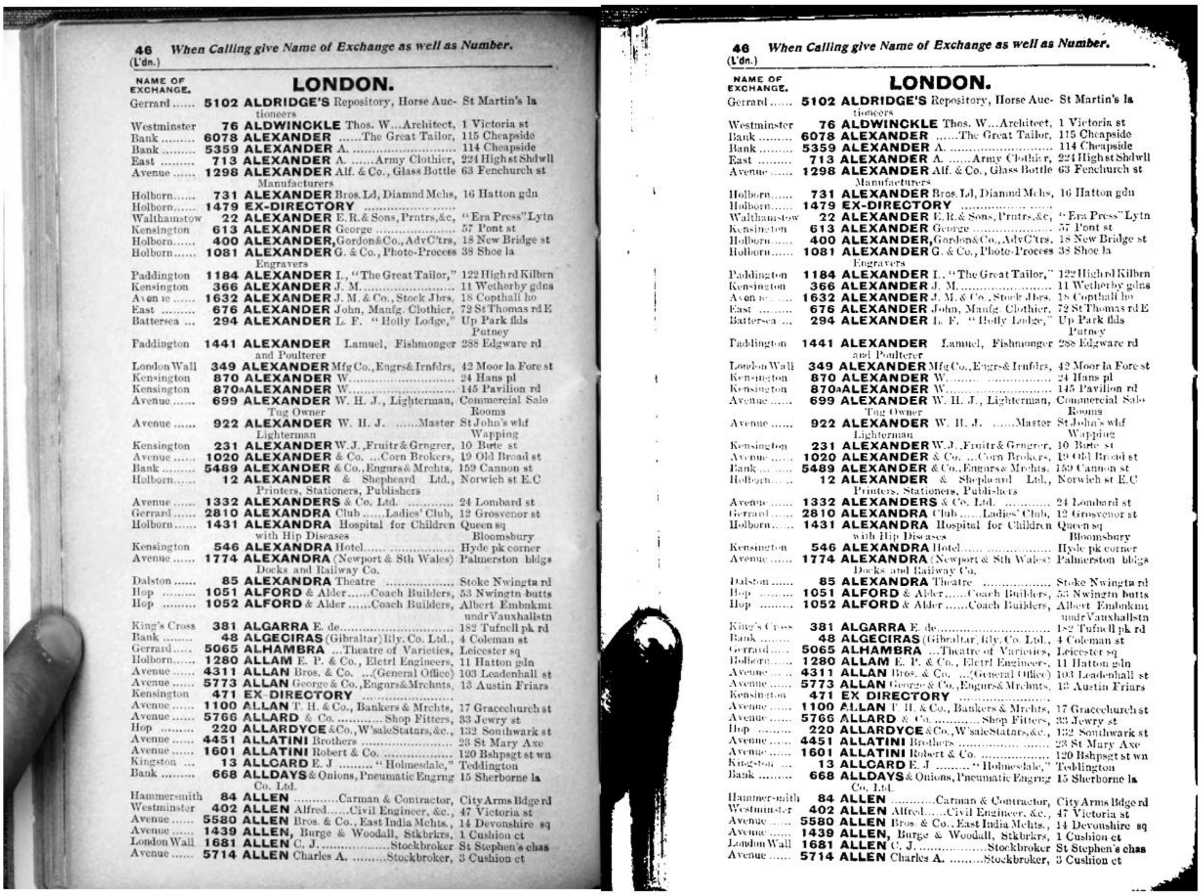
2.2.1 Pre-processing of Images

First and foremost, the pipeline begins with a mass conversion of images belonging to one directory, from scans in the compressed *.j2k* format into *.tif* files, a format that is more amenable for use with common OCR software. Despite this, the scans themselves are kept as *.j2k* files for practicality: this format is an ideal compromise between minimising file sizes and preservation of image quality – both of which crucial for time-efficient and accurate processing of historical documents (Dueire Lins *et al.*, 1994).

Thereafter, every image underwent key operations: cropping of borders and binarization. The former of these aids the OCR engine because it delineates the page extents within which the engine searches for textual information. In doing so, the likelihood of the engine picking up unwanted information, such as page numbers or scribbles on the edge of pages, is minimised. This research employed a trial-and-error approach to achieve this whereby 5 random pages were sampled from each directory and the depth of white space surrounding the main text on four sides of the page was manually noted. Of these pages, the minimum respective values were then used as parameters to crop and export all images from that edition. Random samples of these exports were once again checked to ensure that no essential information was cropped out of these pages.

The other image pre-processing operation attempted was binarization, that is, the conversion of a set of colour pixels into either black or white pixels. In this operation, the algorithm usually determines a threshold grey value, exceeding which individual pixels are changed into white pixels and beneath which they are turned black (Saha *et al.*, 2014). This threshold can either be configured to apply to the entire image (global) or, alternatively, the image can be broken down into smaller areas, for each of which a threshold is set (local) (Ibid.). Because nearly all the scanned pages have uniform colour tone across the image, this research opted for the more straightforward global threshold and Figure 4 below exemplifies how binarisation changes the appearance of an image.

Figure 4: Contrast in the image is increased after binarisation **(right)** ***as opposed to before*** **(left)** ***it. Example of a scanned archive page from 1901.***



Based on prior trials (e.g. Gupta, Jacobson and Garcia, 2007; Reul *et al.*, 2019), it had been anticipated that the heightened contrast in images would better the quality of character recognition. However, during the trial on directories from the Census years between 1881 and 1911, a sample of 100 binarized images that were then passed through OCR software returned poorer performance than their non-binarized counterparts as measured by visual inspection of character misdetections in the outputted text files. A plausible cause may be that in historical documents that have been archived, printed text is often inconsistent shaded, either owing to flaws in the original printing process or due to wear-and-tear that naturally comes with age, thus impeding the effectiveness of running a binarized image through the Tesseract, the engine into which the pre-processed images are later fed into for text capture purposes. The directories also commonly contained pencil strikethroughs and other scribbles in grey. With binarization, the engine tended to mistakenly detect these marks as part of the printed text, thereby distorting the captured text and causing more errors. While what had been attempted was manual pre-processing of images prior to passing them through the software, the Tesseract text capture engine also independently applies some automatic image optimisation that in this case proved more effective than binarizing the images beforehand. For these reasons this research proceeded with the rest of the pipeline excluding binarization.

Finally, some other image pre-processing operations were tested but ultimately did not noticeably influence the results of text recognition and were therefore omitted. These will be briefly covered in this paragraph. Firstly, the scans were de-skewed, or automatically rotated so that the text was vertically upright and thus aiding the OCR process (Tesseract-OCR, 2021). A small but non-negligible proportion of the images from certain directories, especially from 1881, were skewed, possibly because of challenges arising from scanning individual pages bound to many other pages. However, these skews did not, in the vast majority of cases, reduce the quality of character recognition. Secondly, previous research (e.g. Singh and Grewal, 2012) has demonstrated that dilation or erosion of text in images can help in making more recognisable characters that are either too bold or thin, respectively. This particularly relates to historical documents wherein ink bleeding has arisen, the use of now-uncommon fonts or possibly obsolete scanning methods that modern OCR engines are not adept at handling (Tesseract-OCR, 2021). However, this thesis concluded that owing to the relative recency of the source material, both by way of the choice of typography in its production and technology used for scanning, these issues did not pose serious threats to the output. It was following these steps that the images were then inputted into the OCR engine.

2.2.2 Text Capture

Currently, a wide array of OCR software is available on the market and they range from generic and open source to commercial and designed for specific applications, with many proprietary OCR software being variants of more all-purpose OCR engines that were tailor made to suit a particular audience or purpose (Reul *et al.*, 2019). However, while OCR applications on texts using modern fonts and other typographic conventions (for instance, page layout) are at an advanced stage, their application to historical texts still require considerable improvement (Reffle and Ringlstetter, 2013). Following some reviews 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 number of tuning parameters available, and the convenience of licensing arrangements.

With much help of a senior research fellow at the CDRC, this phase of the pipeline was developed following trials of Linux command line interface scripts, the Python and R programming languages and bespoke system libraries that have integrations with one or more of these languages. Technically speaking, the use of all three coding languages would have been able to achieve similar results of mass text capture from the scanned directory images, as wrappers for the OCR engine eventually chosen, Tesseract, existed in both the R and Python languages as 'tesseract' and 'pytesseract', respectively. However, this research found that, in order to benefit from the full range of parameter customisations offered by Tesseract, it was best to run commands directly on the command line of a Linux-operated machine. In particular, when trialling the use of the 'tesseract' package in R, this research encountered difficulties in whitelisting and blacklisting characters for text capture, which were crucial for restricting the range of possible characters in the output of text capture. This problem was likely to be related to the way in which wrappers were written for use of the open-source Tesseract engine in the R language and did not exist when calling on Tesseract from the Linux command line, in addition to how much sparser documentation was for the wrappers in R and Python than in the Linux use-case. Without being able to whitelist or blacklist characters, the OCR engine sometimes captured recurrent and obviously erroneous characters like '€' or '!', which may resemble alphanumeric characters but almost never occurred in the listings of telephone subscribers. In essence, the trade-off to primarily use Linux command line scripts for this phase of data capture was having to learn a new coding language (Bash) in exchange for a smoother and more customisable process of data capture, while leaving the use of the more familiar R language for writing scripts later down in the pipeline.

An alternative that was considered, but ultimately not chosen, was to train an OCR model specifically adapted to the historical telephone directory archives. While this approach would allow for more bespoke tailoring of parameters to best aid character recognition from the scans, it would also take a much longer time since the profiling of historical documents would require considerable manual input from the team of researchers (Reul *et al.*, 2019). The final choice was thus made to use Google's Tesseract OCR engine with finetuning of its parameters appropriate to the particularities of each edition of the directories; this option entailed a process of trial-and-error with almost all parameters that could be tuned in Tesseract OCR. Ultimately, it was found that some parameters had a disproportionate influence on the OCR output when tuned appropriately; these are discussed in further detail hereafter and are namely automatic table detection and page segmentation.

The relevance of automatic table detection pertains to the desired output of this pipeline: a data table in which every column holds a distinct field of information – crucially subscriber name, telephone exchange number and registered address – and every row represents one subscriber. With one such table derived from each page of a directory, these collectively could then be merged to give a digitised version of the directories' records. In theory, this would be a non-issue were the OCR engine able to automatically place information into a tabular format comprising distinct fields and among other research, Gupta, Jacobson and Garcia (2007) have demonstrated the possibility of capturing text as tables from modern or even historical documents, so long as solid lines delineate different columns of information. Unfortunately, similar attempts to replicate this technique on the early directories up till 1911 were unsuccessful as most, if not all, directories lacked lines dividing columns of information and had either texts that ran continuously, or that were inconsistently separated by punctuations such as commas and full stops. Altogether, this lack of consistency made it impossible for Tesseract to detect where one field ended and another started, and this research found automatic table detection unfeasible, rather opting to rely on manual alternatives discussed later in the chapter.

The second important OCR parameter to tune was that for the automatic detection and segregation of within-page columns, or 'page segmentation'. As aforementioned, while the quality of scans generally improves with time, an issue that correspondingly becomes more prominent in later editions is that single pages become more densely packed with information packed into multiple vertical columns. Where before, every horizontal line of text would be reasonably assumed as referring to just one subscriber, the pipeline now must incorporate a way of separating information relating to two or more subscribers within the same row of text. Related to this, Tesseract OCR contains 14 different modes of page segmentation including three modes that detect page segmentation with the help of algorithms and of which two include automatic detection of orientation and written script (Tesseract-OCR, 2021). Results obtained when using these modes varied across pages both within and between directories, with an incomplete detection of column boundaries at best and an output containing a convoluted block of text at worst. This research eventually found that the options for Tesseract to assume either 'a single column of text of variable sizes' or 'a single uniform block of text' worked best because they consistently generated outputs where each record was held on a separate line, albeit without segmentation of the information from different columns. This issue is then dealt with later by querying specific string patterns to separate the information from each record (on a single line) into distinct fields.

Following the completion of trials and of optimisation of the OCR parameters, the majority of the text capture work was handled by the CDRC senior research fellow who had helped me with this process. The system that they put in place was an organisation of scanned directory images on the first level by year of publication, and on the second level by the type of content contained in those pages (as was described in Section 2.1.3). As far as this research was concerned, only the pages containing (partially or fully) information on telephone subscribers were fed into the processing pipeline. They would then run the scanned images through OCR, with one text file being returned for each inputted image. It is this collection of text files, containing the text extracted from the directory scans, that were then cleaned and organised in the following section.

2.2.3 Organisation of Information and the 'teldiR' Package

Following the description of how text extraction from the scans was done, detailed in this section are the next steps in the pipeline pertaining to creating a system of organisation of said information. As they stand, information that had been captured from the directory scans through OCR remain in large blocks of unstructured text in as many text files as there were images originally, and this information would be much more readable if they were to be better integrated with each other. Another manner of organisation that would drastically improve readability would be to separate the different fields of information for each telephone subscriber, for instance to split subscribers' registered names from their telephone numbers. To achieve this on a large scale, the string operations that were to be executed across the numerous directories were typically repetitive, albeit ever requiring some modification, to be used in different orders and for the query of text patterns particular to given editions. For this reason, this research built *'teldi*R*'*, a library of generalised functions in the R programming language, to facilitate the management and processing of text data captured from the historical telephone directory archives. Not only does '*teldi*R*'* facilitate the division of an unstructured block of text into named fields that each contains distinct information, it also provides functionality for ease of execution of operations that are relevant before and after string manipulation. Table 4 below provides an overview of the kinds of operations that the package assists with.

Table 4: Summary of the types of functions in the **'teldiR'** ***package and examples of their use cases***

Category	**Examples of use case**
String operations	▪ `split_strings` and `trim_string` have self-explanatory uses but also allow users to restrict the querying of text patterns to only the first X characters or last X characters of a string *(where X is a number)* ▪ `collapse_RowsUpIf` allows users to conditionally merge two or more rows of data, for when a record in the original directory occupied multiple lines
Process outputs of text capture	▪ `gen_fileLST` lists all text files in a given system folder for easy importation and allows users to easily select subsets of the data, relevant for instance if they are interested in records *(e.g. of a settlement)* corresponding to a defined range of file numbers

Import and export tabular records	▪ `readin_adrDT` imports already processed records (for viewing or further modification), with the option of choosing random samples from the dataset for troubleshooting or trial of new code ▪ `export_fullDT` exports processed records, naming the output file with a consistent format and automatically appends the time of export for purposes of versioning.
General utility	▪ `repl_streetAbbrevs` searches address fields for common abbreviations in throughfare names and replaces them with corresponding long versions ▪ `serialise_recIDs` generates a unique 11-digit identifier for each record which helps when they are to be matched to other telephone records from different years, or to Census records

The primary way in which functions in '*teldi*R' help to organise information on telephone subscribers is first by querying text patterns and then, where there is a match, by truncating, splitting into two or partially duplicating text strings. The queries that are made this way search for user-defined permutations of alphabetical, numeric and special characters present in the results of the text capture. What underlies these queries, in turn, are Regular Expressions (commonly abbreviated to RegEx), which enable, in addition to the above, the use of metacharacters, which are characters able to represent a more generalised pattern or set of other characters, thus drastically enhancing the flexibility of patterns being queried (Campesato, 2019). These functions thus allow users to swiftly and simply manipulate text strings based on a few common arguments, namely the column(s) in the data on which operations should be applied and the RegEx pattern(s) to query, alongside optional customisations which are function-specific.

As shown above, '*teldi*R' is a comprehensive library that facilitates the easy management of the digitised telephone directory records, containing a host of functions to clean, reorganise and add features to the data and is available for public access through this link on GitHub. Using these functions, this research has created datasets which are used in this subsequent analysis, and also comprise an end-product that will benefit other users studying various related aspects of social and historical geographies. End-users may also find the functions in 'teldiR' to be useful, if the directories in which they are interested are either yet to be digitised, or if they would like to improve further on the digitisation of directories from specific periods or locations. In the latter scenario, the processing pipeline prioritises generalisability of operations for application to as many different directories as possible, and it is therefore inevitable that their efficacy will be lower on some directories than others. Altogether, public access to the package means that users will be able to tailor-make improvements to the quality of data according to their needs, independent of this research. More widely, the package may also be useful to the context of other unstructured or loosely structured textual data, which is not uncommon in swathes of recently digitised historical sources. Adding to its general utility is its capability for speedy computation of even large data operations, meaning that it can be applied to datasets larger than the already large dataset of telephone subscriber records: in its development, functions were written on the basis of the 'data.table' package, which has enhanced performance over basic data manipulation functions in the R programming language (Dowle and Srinivasan, 2023).

2.3 Attribution of Geographical Features

One of the foremost possibilities that the dataset of digitised telephone directories affords to research is the ability to locate historical subscribers in time and space; the vast majority of records contain some sort of geographic information, such as the address strings of either their residential or commercial locations and the name of the telephone exchange that serves them. Yet, this capacity alone can rarely be used for spatial analysis and visualisation, owing to its highly imprecise nature and other reasons for its impracticality which are discussed in this section. To become useful, this research argues that these addresses must be attributed with geographical coordinates through approximate text matching with already geolocated databases of addresses, in a process known as geocoding, which is described herein alongside preparation that improves the efficacy of the matching.

2.3.1 Preparation of Address Fields in the Dataset

The initial step of preparation expands on the aforementioned work to separate subscriber information into different fields. Whereas before the pipeline would have left all the geographical identifiers of a record in one field, there is now a need for an even more granular separation of components of the address, such as street name, street number or name of residence, name of area or borough, and regional district/sub-district (where available). Values in the address field of the output data table for each telephone directory entry were further separated into subfields containing these components through Regular Expression (RegEx) pattern querying, although not all of these elements were present in the directories for the different years and separate locations. By splitting up these bits of spatial information, the processing pipeline enjoys efficiency gains through being tasked to conduct the matching of a greater number of shorter strings for each subscriber, rather than of fewer but longer strings (Navarro, 2001). This technique also allows this research better to quantify the geographical precision at which a subscriber address could be matched to the existing database, be it at the scale of an area, a street or the exact street number, something especially crucial given the already considerable uncertainty inherent to working with a dataset as dated as this.

The next preparatory phase entails the standardisation of text in the address subfields mentioned above (that is, excluding street number) with several outcomes in mind and that ultimately harmonises the conventions in the digitised dataset with those used in the database of addresses, against which the former is to be matched. Firstly, many terms common to the naming of throughfares are abbreviated in the telephone directories but must be spelt in their long forms for ease of matching. For example, 'Road' is often abbreviated to 'Rd' and 'Lane' to 'Ln', while some other terms like 'buildings' have had several common abbreviations such as 'bdgs' or 'bldgs', all of which should ideally be considered. In the case of street names, only thereafter is the second standardisation operation applied: converting the text strings into lowercase alphabets for avoidance of mismatches resulting simply from differing letter cases.

A third consideration is that of building in capabilities for the pipeline to deal with sets of alphanumeric characters which the OCR Engine, albeit tuned, is nonetheless inclined to confuse for each other; this strongly implicates the process of replacing throughfare abbreviations with their long forms as just a difference of one character could make an abbreviation go undetected by the algorithm. Furthermore, this relates particularly to the subfield of postal districts, which from the beginning of the availability of data in 1880 have served as an important geographical identifier for London subscribers, especially before the formal introduction and listing of Telephone Exchanges in the directories beginning 1901; the importance of this will be elaborated on later in the chapter. Given that the pool of letters used in London postal districts was finite, it was possible for the pipeline to account for nearly every common misdetection that occurred across the years of information that have been processed. In addition, much effort had to be directed at accounting for common misdetections because of the sheer similarities between some throughfare terms, the prime example being that if 'way' is mistakenly captured as 'wav', then the algorithm detects the 'av' within as short for 'avenue', resulting in an erroneous replacement.

2.3.2 Fuzzy String Matching

Having completed necessary preparations, the data being processed are passed through for geocoding by way of linking subscriber addresses, represented as text strings, to a probable set of geographical coordinates that would subsequently enable spatial analysis and visualisation. As several challenges that stem from the datedness of the addresses captured present themselves, beyond the potential issues with the OCR data capture issues itself, this georeferencing of telephone directory addresses was done through adapting the implementation of the probabilistic method of fuzzy string matching by Lan and Longley (2021): the structure at large behind this matching algorithm was adapted from their work but the code itself was written from scratch by this research in the R language for compatibility with the rest of the pipeline.

To achieve this aim, each entry's street name was matched against a reference database of addresses, that is the subset of addresses in London in the digitised Census records of 1881, which had been perfectly matched to a contemporary address in the Ordnance Survey's (OS) AddressBase. Given that the OS AddressBase is a collection of addresses in present-day Britain, and that road networks throughout the country have undergone massive change since 1881, it is far from guaranteed that the addresses of today would match up with those listed in the historical telephone directories. Thus, this research matched the street addresses of recorded telephone subscribers against the subset of modern addresses in London that had been perfectly matched to an address captured in the digitised Census records of 1881. This is a likelier guarantee that the pipeline would georeference telephone subscribers to addresses that actually existed in their respective eras and avoids a potential linkage to addresses that only became existent decades after the records were made.

Following the string matching described above, the AddressBase address that shared the lowest inter-string distance with each telephony directory address was then selected using an algorithm for georeferencing. When an exact or approximate street name match was found for an entry, a further search for an address with a matching street number would be initiated. In this way, the georeferenced addresses may be accurate either to the street level or the street number level. Moreover, there will exist inherent ambiguity in the geocoding of addresses using historical street addresses because streets may be renumbered, renamed or even demolished with time. Nonetheless, this thesis believes that even addresses accurate to only the street level will still be meaningful to the geospatial analysis they lend themselves to, given that this process already dramatically improves on the much less granular scale of historical geospatial analysis that is hitherto commonplace for the time period.

Fundamental to the georeferencing process is the choice of string-matching algorithm, which helps to create record linkages between the telephone directory addresses and an almost all-encompassing address database. In perhaps one of its earliest incarnations, Wagner and Fischer (1974) detailed three kinds of string operations that they considered most relevant to calculate inter-string distance – that is, a measure of how similar two strings of text are. These operations were namely: removing a character (deletion), inserting a new character (insertion) or replacing one with another (substitution). In some variants, a fourth operation, transposition, is considered, which refers to the movement of a character either forward or backward within the string. Table 5

Table 5 below summarises the differences between some common measures of inter-string distance that have been implemented within the 'stringr' package in R (Wickham, 2023), the coding language in which this analysis was undertaken.

Table 5: Key differences between common measures of inter-string distance implemented in R

String Matching Method	**Allowable String Operations**			
	Deletion	Insertion	Substitution	Transposition
Hamming ** requires strings of equal length*	N/A	N/A	Yes	N/A
Levenshtein†	Yes	Yes	Yes	No
Optimal String Alignment (OSA) †	Yes	Yes	Yes	Yes, once
Full Damerau-Levenshtein†	Yes	Yes	Yes	Yes, multiple
Longest Common Substring (LCS)	** extracts the longest substring common to both strings, keeping the order of characters the same*			

† for these methods, the weight that each kind of string operation carries in influencing inter-string distance can be adjusted manually

Overall, by default, the basic measures of string similarity that utilise this framework calculate the inter-string distance by summing up the total number of such operations that must be applied to transform one string into the other. A distance of zero thus implies a perfect match, while a high number indicates stark dissimilarity. Further, the weights of these operations may be adjusted such that, for instance, an insertion could increase the inter-string distance by twice as much as a deletion would, should this be viewed as optimal.

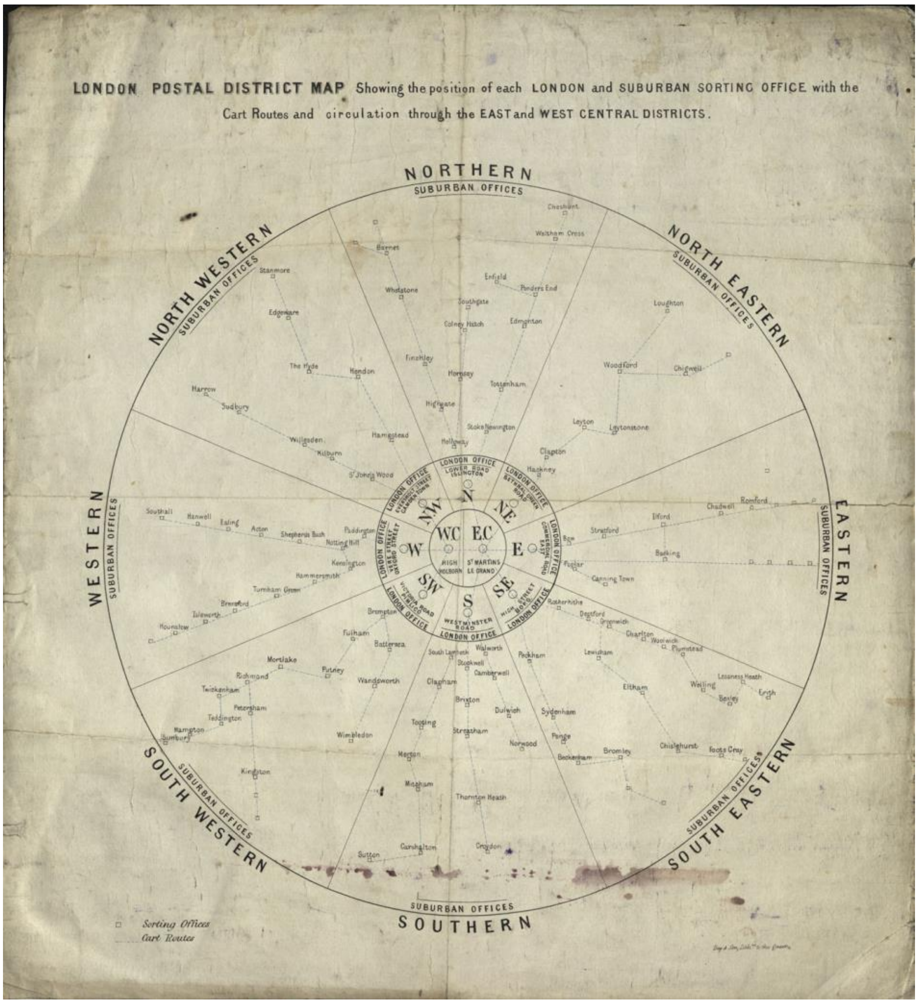
Ultimately, the Hamming and Longest Common Substring methods were deemed unfeasible. The former was deemed so because frequent abbreviations of throughway names and shortening of other street names (such as removing almost all vowels) would create unequal string lengths between the addresses extracted through OCR and those from the OS AddressBase. This reason, alongside the inevitable (albeit often minor) misdetection of characters, would also render it challenging to consistently detect common substrings in addresses from the two sources. Meanwhile, the three remaining string-matching methods detailed in Table 6 were trialled on the London directories from the Census years between 1881 and 1991, with their values for string transposition allowances being the key differentiator. The Levenshtein Distance metric, which does not allow for transpositions, was ultimately chosen; transpositions were done away with because they would be meaningful only if neighbouring characters in text strings were often wrongly ordered, a prime example being human typos arising from the manual entry of a large number of addresses – which did not appear to occur in the original production of the directories. In the context of abbreviated address names and character misdetections owing to the OCR engine, the operations of deletion, insertion and substitution therefore remain more relevant.

2.3.3 Reconciling Different Administrative Geographies

While it is theoretically possible to match each address extracted from the telephone directories to every possible address in the database, the sheer amount of computing power this goal would demand renders it practically unfeasible. To circumvent this, past applications of geocoding through fuzzy string matching have used civil administrative boundaries to conduct a more targeted matching of addresses and also improve the accuracy of matching outcomes, since some popular throughway names like 'High Street' may exist multiply, even in the same city (Lan and Longley, 2019). For this reason, Lan, van Dijk and Longley (2021), when geocoding addresses from historical censuses, used Parish boundaries to delineate the possible matches for each address string in the census. However, while the Census record addresses are tied to historical administrative units, addresses found in the telephone directories are not. Rather, different yearly editions have the street names and numbers of addresses accompanied by different geographical identifiers, most pertinently the postal district markers used in the early directories of London and telephone exchange areas for most directories as the growth of telephone service networks accelerate in the 20th Century. A further necessity for geocoding the telephone subscriber addresses is therefore to devise methods to reconcile the civil administrative boundaries, according to which the AddressBase addresses are categorised, with the other geographies by which telephone directory addresses are listed. Three differing methods are hereafter explained, each used to geocode different sets of directories and to different ends, with a key distinction being that the geocoding process for the three largest settlements in Britain in 1901 (London, Glasgow and Manchester) were intentionally made more comprehensive than for other smaller urban areas.

The first method pertains to the London directories from 1881 (when directories for no other British location existed) and 1891. In these directories, addresses were listed alongside the London postal districts they fell under, which was the original system, implemented beginning 1857, that divided the area within 12 miles of Central London into 10 mutually exclusive sectors, each denotated by a compass direction (The Postal Museum, 2021). The postal districts of each entry's address string would be 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 (indicated by a compass direction) of the extracted address, thereby improving the efficiency of address matching by shrinking the possible pool of address matches from, for example, a whole city to just a number of boroughs or parishes that overlap with the telephone exchange area. This field, however, only exists in later telephone directory editions, beginning with the London directories of 1901.

Figure 5: The original postal district map of London, implemented in 1857-58, that divided the central area of London into 10 zones.



(Source: The Postal Museum, 2021)

Taking cues from the place names present on original postal districts map (see Figure 5) in conjunction with those contained in modern-day maps, this research created a correspondence list between postal districts (of the telephone directories) and the olden-day Registration Districts (of the Census records) in two steps. In the first step, for each Registration District, this research looked up its relative position on a map of London from 1881 to determine an approximate range of compass directions (postal districts) that correspond to the Registration District. Based on this initial indication, the postal districts map was then consulted to see if there was an immediate match for the Registration District in one of the ten sectors, but which admittedly happened only in the minority of cases. For instance, in this way, ‘Poplar’ was assigned to correspond solely with ‘E.’, and ‘Woolwich’ with ‘S.E.’. In most other cases, because there was not a definitive correspondence that could be derived by comparing both maps, approximations were made where each Registration District was matched to multiple possible postal districts, allowing for a greater margin of error, and this applied especially to central and/or larger Registration Districts. For example, ‘Strand’, to the East, borders the historical Registration District of Holborn, which is shown in Figure 5 as likely being subsumed within the ‘E.C.’ zone. However, Strand also stretches further West on the map and thus its potential postal district correspondences were expanded to include ‘N.W.’ and ‘W.’ in addition to ‘E.C.’. Meanwhile, in the Southern half of London, many Registration Districts were listed to correspond with both ‘S.’ in addition to either ‘S.E.’ or ‘S.W.’; not only were many Southern Registration Districts longer and larger than those in the North (thus potentially spanning multiple postal districts), the discontinuation of the use of ‘S.’ in London postcodes today also made it more difficult to ascertain, with the aid of modern maps, how exactly it was bounded, unlike all other historical postal districts. Therefore, assigning the Southern Registration Districts correspondence with typically more than one postal district was seen as a necessary safeguard to implement.

In the second of two steps, the Registration-to-postal district correspondence list was inverted to give a postal-to-Registration District correspondence list, which was ultimately used to narrow down the number of possible address matches for each subscriber’s address. This entire process was inherently probabilistic, given that the postal geographies of London had changed considerably in the decades before 1881 (ibid.) and also that no information or maps of the postal districts that was more specific than had been shown above could be found. As a further measure to maximise the yield of address-linked postal districts for ease of geocoding, many frequent misdetections of the characters occurring in London addresses were accounted for and accordingly substituted (Table 6).

Table 6: Most common misdetections in London postal district markers that were accounted for

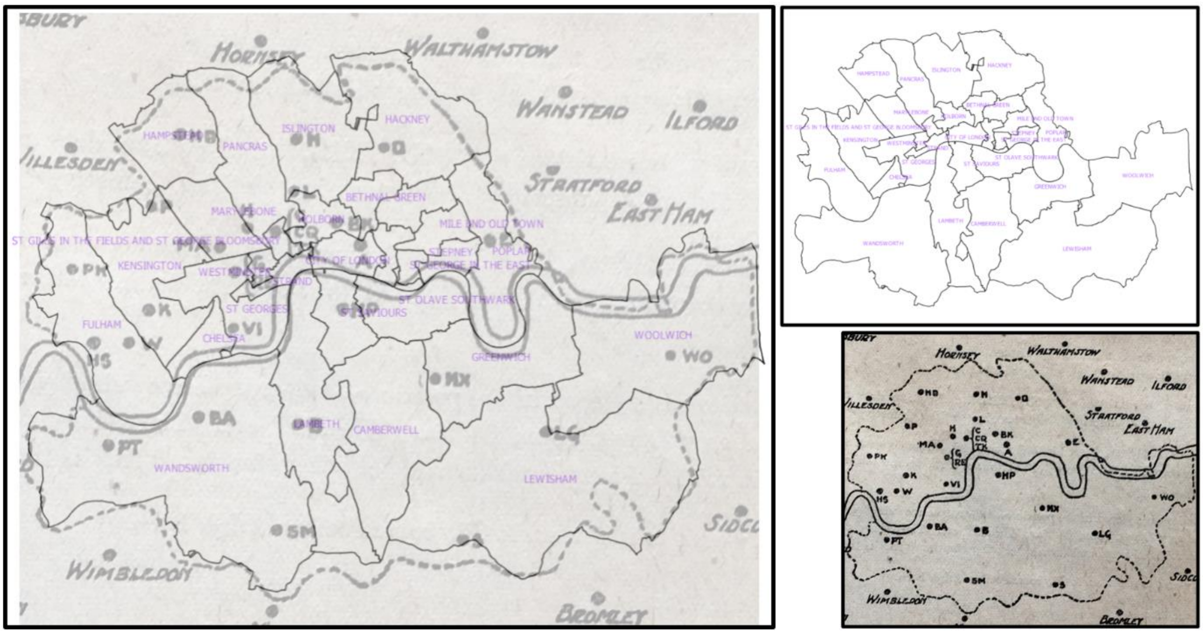
Actual Character	Common Misdetections
S	‘5’, ‘8’
E	‘B, ‘F’, ‘H’, ‘13’, ‘18’
C	‘G’, ‘O’, ‘0’
.	‘,’, ‘ ’, ‘-’

As opposed to address matching which ignored the likely correspondences between postal and Registration District geographies, this approach, crucially, reduced the processing time expended threefold, and was likelier to guarantee that addresses were not allocated rogue matches whose positions in the city was completely different to where they would be, speaking from their postal districts.

The second approach pertains to the London directories in the Census years from 1901 and onward, when telephone exchange areas become the dominant logic of geographic organisation of the records of subscribers across all directories. This approach is largely similar to the first, albeit using telephone exchange areas rather than postal districts. London is chosen as the illustration here because it is by far the most complicated and presenting the highest number of subscriber records to be geocoded each year. From 1901, the notation of subscriber addresses in London also no longer consistently includes their postal districts, which makes the previous method no longer fit for use. Similar to that method, though, is that a list of correspondence between the civil administrative boundaries (Registration Districts) and the telephone exchange geography of London was created. To achieve this, a hand-drawn map of the approximate locations of telephone exchange areas in the 1914 directories (which was the nearest in timeframe to 1901 found) was superimposed upon Registration District boundaries from 1881 (see

Figure 6 below), from which overlaps between the two maps were recorded.

Figure 6: Superimposition of maps depicting London's civil administrative boundaries in 1881 **(right, top)** ***and telephone exchange geographies from 1914*** **(right, bottom)**



Also considered, but ultimately not selected, was the creation of a list of correspondence between telephone exchanges and civil parishes instead of Registration Districts: while parishes contain data at a more granular scale, the number of Registration Districts in London as of 1881 was identical to the number of telephone exchanges in Central London as of 1914 – the two years of data being the closest in time to each other that were found. This made it more likely that the latter two geographical units were similar in the scale of their coverage of areas and populations in comparison with civil parishes. In this way, the establishment of correspondence between the two geographies was limited by the granularity of the telephone exchange geographies. Because this manual effort would link every telephone exchange area to multiple civil geographical units, it also made much more sense for this analysis to be done against 29 Registration Districts, rather than civil parishes which numbered 190 in 1901.

Notably, owing to the density of telephone exchanges in central London, each telephone exchange was given correspondence to a higher-than-usual number of Registration Districts to account for the uncertainty of not knowing where these exchanges' service areas actually end. These would refer to the Registration Districts around, for instance, Holborn, City of London and Marylebone. Further, in comparison to the approach of individual string correction applied onto the postal districts of addresses in the 1881 directory, a second layer of fuzzy string matching was instead chosen to match the telephone exchanges to which each address belonged to with Registration District boundaries. Because the misdetection of characters in the telephone exchange names by the OCR engine were much more varied and irregular than those of postal districts, employing fuzzy string matching proved to be more time-efficient in identifying exchange names which had, for instance, one or two characters misspelt or missing.

The approach used for geocoding records of subscribers from Glasgow and Manchester largely mirrors the string matching described above but excludes the more specific matching of civil and telephone exchange geographies that had been done for London records. Instead, addresses in the telephone directories for each of these settlements were matched against all addresses from the OS AddressBase within the same confines. This simplification was made because of, and permitted by two factors, namely the reduced complexity of these settlements' geographies and their comparatively lower volume of subscribers. Even by 1911, the telephone exchange geographies of Glasgow and Manchester had not been developed enough to warrant the inclusion of maps of their telephone exchanges in their respective telephone directory editions. The replication of the strategy employed on London records was as such made more challenging. In addition, the number of subscribers in London had consistently remained well above that in Glasgow and Manchester through the editions of telephone directories that were digitised, prompting this research to prioritise the finetuning of the algorithm for London over the other two settlements in the limited time that was available.

The third and final approach that was applied to geocode records of subscribers from urban settlements other than London, Glasgow and Manchester was less sophisticated but far more time-consuming. For these localities, the intended outcome of geocoding was simply to be able to enumerate the number of telephone subscribers in each of the settlements outside the top 3, so as to enable macro-level comparisons of growth of telephone subscription between cities. The approach thus entailed manually notating the range of pages in each year's directories that held information about subscribers in each of the top 50 settlements in Britain by population in 1901 (Lan and Longley, 2021) and then attributing to these records an indication of the settlement whose bounds they fell under. Reducing the investment of effort required to finetune an algorithm to geocode subscribers in these settlements also meant that more time could be redirected to other strands of this research.

2.3.4 Geocoding Statistics

The outcomes of geocoding for the records from the three settlements are shown in Table 7 below. The percentage of records in each year and for each settlement which fulfil particular criteria related to fuzzy string matching are listed therein. The leftmost criteria ('matched perfectly to street name and number') is the most stringent, with the matching criteria becoming gradually more relaxed as one moves from left to right within the table. Correspondingly, the proportion of records fulfilling each subsequent criterion is predictably higher than for the criteria preceding it. As an example of interpretation, out of the 12,847 records for London in 1881, 9.02% of records were matched perfectly, by street name and street number, to an address from the reference database of addresses, 39.27% of addresses were matched perfectly by street name (regardless of whether there was a match on street number), and 77.89% of all addresses that year were matched to a reference address with a string distance of no greater than 5.

Table 7: Statistics for geocoding done on the records of each of the three largest British settlements, using fuzzy string matching.

Settlement	Year	Number of Records	Matched perfectly to street name and number	Matched only to street name, with string distance no greater than:				Mean String Distance
				0	3	5	10	
London	1881	12,847	9.02	39.27	65.87	77.89	91.62	2.14
	1891	18,731	9.8	39.68	57.89	71.5	81.32	2.58
	1901	23,820	13.76	46.94	64.14	81.3	81.61	2.91
	1911	311,482	15.93	51.04	71.11	88.21	78.95	2.55
	1921	298,100	14.48	56.77	74.51	86.44	68.32	2.47
	1931	573,440	6.7	25.78	40.93	60.86	64.81	4.38
	1941	436,818	6.65	38.31	49.72	60.82	43.28	3.9
	1951	579,163	7.79	36.98	49.81	61.37	48.82	3.99
Manchester	1901	1,019	5.59	42.2	56.23	77.13	86.85	3.14
	1911	12,128	4.35	50.72	64.57	81.98	71.86	3.28
	1921	21,449	5.1	49.8	62.37	77.27	73.67	3.21
	1931	55,757	5.19	11.7	42.66	58.73	81.66	4.09
	1941	82,380	3.83	8.52	40.36	57.41	79.14	4.19
	1951	139,132	4.79	10.4	45.16	61.68	82.32	4.02
Glasgow	1901	832	9.01	61.54	75.24	88.34	79.45	1.78
	1911	26,921	8.43	59.98	77.46	91.1	76.46	2.09
	1921	22,759	10.27	59.37	75.64	88.13	80.05	2.1
	1931	38,007	5.8	19.83	50.57	67.99	88.6	3.59
	1941	59,664	3.41	11.92	41.76	62.21	83.95	4.02
	1951	91,853	3.66	12.14	47.81	69.27	87.97	3.79

On the whole, across all three settlements, the trend follows that the rate of address matches, no matter which criteria this is evaluated against, falls with the passing of time. For London, the rate of perfect matches on street names (ignoring street number, i.e. in the column 'string distance no greater than 0') actually first increases from 39.27% in 1881 to a peak of 56.77% in 1921, before falling again to 36.98% in 1951). Geocoding for Manchester and Glasgow had only been conducted beginning two decades later than for London, according to when they became listed in the telephone directories; the rate of perfect matches on street names in Manchester starts off comparably to London at 42.2% in 1901, while for Glasgow this figure is markedly higher at 61.54%. However, between the latter three decades from 1931 to 1951, the match rates for both settlements fell considerably more than in London to 10.4% and 12.14% for Manchester and Glasgow, respectively. Yet, when considering the match rates under the criteria of 'street names matched, with a string distance no greater than 3', figures are more comparable across the three settlements in each decade from 1931 to 1951, with figures falling within the band between 40.36% (Manchester in 1941) and 50.57% (Glasgow in 1931).

While these figures are far from perfect, this research argues that they are nonetheless useful in extending the utility of the digitised telephone subscriber records. Particularly for London, on which most effort had been expended, a perfect (based on street names) match rate of 36.98% in 1951 signals that, out of the 579,163 subscriber records in that year, 214,171 records could be placed precisely to their streets of residence. A hypothetical study which made use only of these perfectly matched subscriber records, while undoubtedly having to grapple with issues of representativeness of the data, will already have an immense volume of granular data which it can base its analyses on.

Nevertheless, there is considerable room for improvement of the processes of data digitisation. The better match rates across all settlements in the early decades are likely explained by the fact that the reference database on which address matching was done, derived from the OS AddressBase Plus product, was a subset of modern addresses from the 21st Century that had also been found among records in the national Censuses of 1881. The further away from this year the matching done, the likelier it is that roads would have undergone various forms of renumbering, demolition and rebuilding, or even destruction due to wartime attacks, leading to a reduced match rate across the board. For all settlements, there is also a pronounced jump in the mean string distance of all matched subscriber records between 1921 and 1931. While this research was not able to investigate historical developments in this period which may have resulted in this trend, this may be something that future research seeking to improve on the geocoding of the records in the later years should consider exploring.

Certainly, it cannot be denied that the poor rates of perfect matching (in terms of both street names and numbers) may be partially attributed to flaws in the raw dataset itself, as well as particular mistakes in the text capture process which cannot be easily rectified. As had been discussed earlier in this chapter, this research had endeavoured to continuously and iteratively improve the processing pipeline to help it better deal with problems of common character misrecognition and to improve its accuracy of sorting different information into their respective fields (such as separating street names from street numbers). To further improve the match rates, it may be necessary for future research to modify the pipeline and geocoding procedure to be tailored to the respective areas of their interest. A limitation of the method employed was also that it had relied largely on personal observation to identify needed improvements to the processes of text processing and geocoding, implementing changes most commonly by using Regular Expressions (RegEx) to check for, and correct, common string patterns that were wrongly captured. To this end, while this research had not gone as far, it may be worthwhile for future research to consider the use of machine learning models to help with the detection of errors in text capture. For instance, such a model could be trained to determine if the text capture of a certain record from the scanned archives is accurate, using as training data a subset of the telephone records – both in their raw (scanned) as well as digitised formats – that has been manually annotated for the accuracy of data capture. Were such a model to be successfully developed, recurrent patterns of text misdetection or mis-capture could be much more easily identified, and improvements made to the processing pipeline thus made more targeted, ultimately helping to improve the efficacy of geocoding further down the pipeline.

2.4 Conclusion

In a nutshell, this chapter has outlined the features of the raw dataset of telephone directory page scans as this research received it as well as the numerous steps which had been taken to transform the data into a more accessible tabular format. The choices that had to be made in these processes of software utilised, parameters tuned, and rules adopted and applied over their alternatives are justified; the challenges faced in the digitisation are mitigated as far as possible or, where immitigable due to the inherent properties of the dataset itself, the reasons therefor are elaborated; the constraints of time available in this research or computing power are also discussed as they relate to the compromises that had to be made in the production of the dataset. While still possibly riddled with shortcomings, this research argues that the outcome of the digitisation pipeline is still sufficiently robust for the purposes of analysis of, and subsequently making conclusions about, the spread of telephone uptake in Britain as well as the adjacent themes of demography and urban change in historical Britain. Furthermore, the dataset itself will be published and made accessible to the public by request, meaning that it can afford its utility to a variety of research interests, whether academic or otherwise, that may focus on much more specific subsets of the data.

3 Telephone Uptake Across Urban Britain

Within geographic literature and that of related fields, investigation into the emergence and subsequent spread of the uptake of new technologies has been commonplace. Among the earliest and most well-known examples of this process are the conceptualisation of the idea of Diffusion and Innovation (Rogers, 1962) and, later, Hägerstrand's (1967) further development by arguing for the inherence of a strong spatial element to this process. In his definition of diffusion, Rogers (1962) stressed the importance of social networks as a medium through which an innovation is made known to and thereafter becomes adopted by more members of the population. One simple application of this in reality, Hägerstrand surmised, is that new adopters tend to be those already living near the early adopters and thus receive face-to-face exposure of its potential (Hägerstrand, 1965). In this way, geography is already important to the diffusion of a new technology at the local scale and at the wider national scale, it would be all the more critical in determining how the innovation spreads through, for instance, the strategies of expanding their service provision from region to region that the service providers implement.

An investigation into the spread of telephones would not serve solely to trace the rise of what would turn into an everyday household object by the end of the 20th century, but also provide some context to urban growth in Britain. More specifically, telephones have been considered important markers of the development of urban settlements in more ways than one, an example of which originates from another prominent theory in locational geography. In introducing the Central Place Theory as a means to conceptualise how settlements relate to one another in space, function and relative size, Christaller (1966) incorporated in his operational definition of settlement centrality the number of telephones contained in a given settlement. In similar vein, Hägerstrand (1965) used the volume of telephone traffic as a proxy variable for the extent of intra-urban interaction between individuals, as measured by records of telephone calls made between telephone exchanges in select Swedish settlements. The introduction of telephones can be said to have revolutionised the ways and brought down the barriers with which people communicate and exchange their ideas such that in the present, well past its heyday, its long-lasting impact of fixed-line telephony on telecommunications is still felt through the development of mobile telephony that it arguably prompted.

The previous examples of studies that inquired about the spread of an innovation had usually been conducted relatively contemporaneously to the advent of the innovation which they attempted to study. In this way, they helped to describe patterns and monitor changes in the development of said innovation in the present while also potentially proving useful in influencing future planning decisions to be made by the provider of technology or the authorities that are trying to monitor and regulate its uptake – as was the case for British telephony, where the telephone in its early phase of operation was actually seen as a competitor and threat to the revenues of the state-owned telegraph services (Holcombe, 1906). What is unique, then, about what this chapter seeks to achieve is that its time perspective on the study of innovation diffusion is retrospective: as of the time of writing, fixed-line telephones can be said to have passed the different stages of initial uptake, peak adoption and finally a decline into obsolescence. While being far from current, this chapter benefits from access to a virtually complete database of how telephone adoption transformed in one geographical entity from start to finish. Nevertheless, this chapter also endeavours to build upon an already well-established tradition, albeit with the aid of novel data on arguably one of the most influential advances in telecommunication history – the fixed-line telephone.

The newly digitised telephone directories represent a novel opportunity to study the specific pattern of innovation diffusion pertaining to telephone adoption in Britain as its network coverage increases over space and time. Conventionally, one would expect that technologies would first reach the largest urban centres, where there are concentrations of both people and the desire for, and ability to afford, these initially luxury goods. With the coverage of the digitised telephone directories spanning an overwhelming majority of directories published in Britain from 1880 through most of the 20th Century, it is also therefore technically possible to interpolate an understanding of the nature and speed of changes in the volumes of telephone subscribers across the country. Where historical geographical research before the turn of the century tended towards "locality studies" that emphasised the importance of local circumstances in determining how processes unfold (Dennis, 1991), the new dataset this research has produced can allow for a more big-picture overview of the spread of the telephone. This chapter thus provides a starting point in the thesis through macro-view description and analysis that compares the variations in uptake of fixed-line telephony across the urban settlements and regions of Britain.

3.1 Methodology

3.1.1 Calculating Number of Subscribers

The process of calculating the telephone subscription figures in different parts of the country is described hereafter. As a precursor, in the relevant years (that is, those contemporaneous with national censuses), all available telephone directory scans were digitised in line with the processing pipeline described in Section 2.2. As this section's analysis was designed to prioritise breadth of insight over depth, some steps in the pipeline were skipped, notably geocoding. The attribution of more specific locational details through the geocoding of subscriber records (as described in Section 2.3) enables analysis in other parts of this thesis to delve deeper into the spatial analysis of select larger settlements. However, for this analysis, it was more important simply to know how many records were successfully digitised from each directory, within a defined page range. In the following steps, I perused the directory scans page by page and notated important information.

Table 8: Example of the manual notations made to help enumerate the telephone subscribers in each settlement.

Section Heading	Perfect Match?	Matched Settlement	Start Page	End Page	Directory	Section Index	Shared Area(s)
Edinburgh and District	Partial	Edinburgh	267	368	bt_900701…	-	-
Glasgow & District	Partial	Glasgow	87	255	bt_900700…	28-	Paisley
Glasgow and District	Partial	Glasgow	89	261	bt_900701…	-	Paisley
Hull District	Partial	Hull	894	939	bt_900698…	26A	-
Leicester	Yes	Leicester	577	634	bt_900696…	17B	-
Leicester & Derby	Partial	Leicester	551	644	bt_900697…	-	-

As shown in Table 8, the first step to enumerating telephone subscribers in each settlement involved a manual listing of the section heading names, and notating the page ranges which they occupied in each directory. This would be the primary information used to subset records belonging to a settlement, which are then tabulated to give an initial count of subscribers residing there. Additionally, the 'Section Index' corresponding to each relevant settlement was noted down to streamline perusal of the directory (given that some directory sections had geographical coverage which was irrelevant to this research). Then, I ran a code script which searched, in the directory section headings, for occurrences of the names of the 50 most populous settlements. While some of these were perfectly matched (as is the case for the section 'Leicester' in Table 8), many others were partial matches, as is for how Edinburgh the settlement was matched to a directory section entitled 'Glasgow and District'. In cases similar to the latter, mostly where section headings included the words 'and District', I looked into the relevant directories for indication of the geographical extent of its coverage. Where it was found that a given section additionally covered another settlement listed in the 50 most populous, I noted their names down in the 'Shared Areas' fields of the table. This would facilitate attribution of telephone subscribers to individual constituent settlements, a process which is described later in this chapter. The process described above was repeated for every telephone directory in the years pertinent to this research to generate subscriber counts for each settlement of interest, as far as information was available.

That being said, the process of enumerating the subscribers in each of the selected settlements is certainly not without problems. Broadly, a key challenge is that the geography of telephone exchanges, according to which the records in telephone directories are organised, do not always neatly align with civil geographical boundaries used in British urban centres. To illustrate this problem, we consider the following example: up till 1921, due to smaller overall subscriber volumes, the subscribers in smaller settlements adjacent to Manchester such as Oldham, Rochdale and Stockport were listed in standalone sections of the telephone books. However, beginning 1931, these settlements became subsumed under the heading of 'Manchester' for purposes of telephone book organisation, although they remained distinct urban centres throughout this period. Were the analysis to take the telephone exchange area of 'Manchester' as holding records only of subscribers in the civil geographical definition of Manchester, this figure would likely grossly overestimate the actual number of telephone subscribers there.

Changes were also effected on both the civil urban geographies and the telephone exchange geographies every so often, more often than not out of sync with each other, resulting in further obstacles for the analysis to harmonise the delineation of areas being analysed. At other times – becoming increasingly frequent with the passing of time, as the size of individual directories grow – the list of subscribers in a smaller settlement may even be altogether absorbed into that of a larger, nearby settlement, which in turn poses difficulty in differentiating how many subscribers actually belong to each of the constituent areas. Some settlements – most notably in the case of Liverpool and Birkenhead – even had their subscribers listed together in every year of data relevant to the analysis. In order to account for these stark variations in geographical definition, the raw subscriber counts that the initial analyses yielded had to be adjusted; even if there was not a highly precise way to do this, this research made attempts to bring the results closer to what actual figures might have been. These attempts, in the form of applying multipliers to the subscriber counts, are described hereafter.

3.1.2 Moderating the Settlement Counts by Applying Multipliers

As described above, this research had to take additional steps in order to adjust the telephone subscriber counts attributed to each settlement, with the bigger obstacle being the inconsistencies between the urban geographies used by civil authorities and those in the telephone books. Two options were strongly considered to overcome this, the first of which was to individual sort subscriber records in sections of the directories that were 'composite', that is, comprising subscribers from multiple urban settlements, into their constituent settlements. This method would have used the data captured in the 'Telephone Exchange' field to determine which urban settlement (by civil geographical definitions) a particular telephone subscriber lived in. In some cases, this process would be simple, such as how subscribers in 'Paisley' were almost always registered under the 'Paisley' telephone exchange area, even in the directories in which they had been subsumed under the banner of Glasgow. However, this process would be much less straightforward in many other instances since each settlement could contain more than one telephone exchange, with the number of exchanges in service only growing with the expansion of the national telephone network. Moreover, this method would rely on a field in the processed data that, as described in Section 2.3.3, is inevitably prone to immitigable errors in data capture. Opting for this method to granularly attribute telephone subscribers to their actual settlements of residence – without a reliable and efficient method of checking for the accuracy of the 'Telephone Exchange' information in the first place – was deemed as far from ideal.

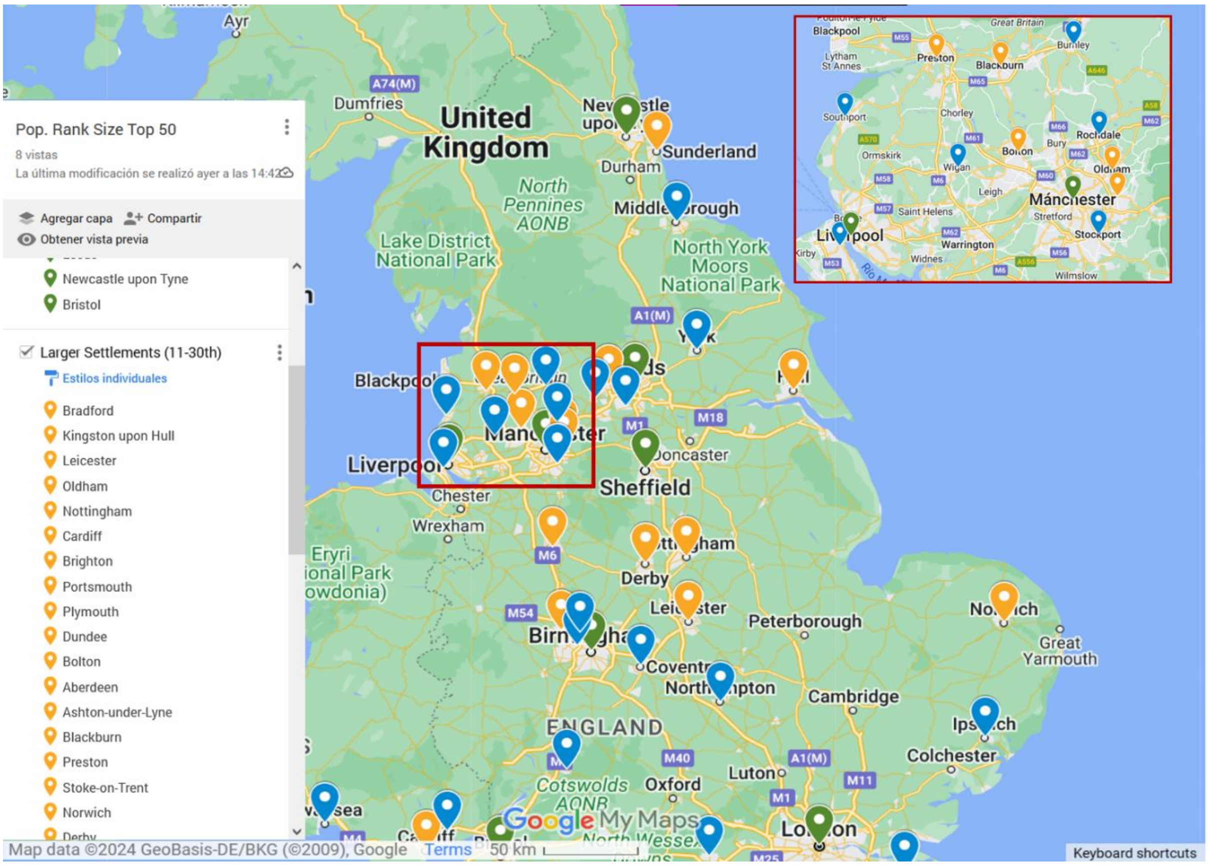
The second option, that was ultimately chosen, was to break down such composite telephone directory sections into their constituent settlements based on a mixture of algorithmic methods and heuristics. The broad principle here was to allocate the subscribers in each such section of the directories to their constituent settlements, proportionally to the relative size of their urban areas. As a pre-requisite to applying these calculations, this research spent considerable effort in consolidating relevant information from within the directories and other sources, namely the groupings of urban settlements (ranking among the Top 50 in terms of settlement size) that were adjacent and thus likely to share sections in the telephone directories. This was done by manually perusing the scanned versions of their pages and annotating the co-occurrence of settlements in an Excel spreadsheet to be used later in the analysis. I had to adapt the method of this derivation as the information presented in the directories changed with time, and these alterations are summarised in Table 9 below.

Table 9: Description of various ways in which the composition of settlements for each telephone directory section was derived, by year of directory

Year	Method of Derivation
1881	N/A *(data available only for London)*
1891	**Section headings:** A small number of section headings contain the names of more than one settlement, most notably the section "Liverpool & Birkenhead".
1901	**Lists of settlement names:** Available at the beginning of each directory section is a list of all areas therein being served, ordered alphabetically. Smaller settlements, more than usual than in other years, tend to have their own sub-sections, while others are subsumed under the heading of a larger settlement or a wider region (e.g. "South-West Lancashire"). Just 2 of 53 settlements are subsumed under sections with the name of other settlements.
1911	**Lists of settlement names:** *as in 1901*. The lists of settlement names tend to be, however, much more comprehensive than before.
1921	**Section headings and Lists of settlement names:** *as with previous years*. However, there were very few sections shared between major settlements in these editions.
1931	**Index of towns:** A comprehensive index of towns is present at the beginning of each telephone directory from this year. Looking up the names of smaller settlements, readers are led often not to a standalone section belonging specifically to that settlement but rather to one with the name of a neighbouring and larger settlement.
1941	**Network coverage maps:** Present before the list of subscribers in each section are hand-drawn maps comprising the names of telephone exchanges encompassed within this service area. These names are used to infer if telephone subscription to smaller settlements have been subsumed under the name of a neighbouring and larger settlement.
1951	**Network coverage maps:** *as in 1941*

As an important aid to this process, this research also mapped the locations of the 53 unique settlements on Google Maps to help myself identify the settlements likely to have been subsumed under the heading of a neighbouring settlement (see Figure 7, top). On some directories, the map could even be used in conjunction with hand-drawn maps in the directory that indicated more precisely the geographical extent of each telephone exchange area (e.g. see Figure 7, bottom). The settlement markers and ellipses drawn in the following two figures were colour coded by population rank size, with the ten largest settlements being coloured green, the 11th to 30th largest settlements yellow and finally any settlement ranking 31st or lower being coloured blue.

Figure 7: Map depicting the 53 settlements relevant to this analysis on Google My Maps **(top)** ***and example of how the mapFigure 7 is used to help identify settlements which are subsumed under the heading of larger, adjacent settlements*** **(bottom)*****.***



** Location pins are colour coded by rank size in terms of resident population as of 1901 as follows:* ***green*** *(1st – 10th),* ***yellow*** *(11th – 30th) and* ***blue*** *(31st – 50th and including unranked settlements). Inset map zooms into the area around Liverpool and Manchester. An interactive version of the map can be accessed via [this link].*

**Base image taken from page 8 of the scans of the telephone directory 'bt_900859_box115' from year 1951. (Source: BT Archives, 2021)*

Having identified the telephone directory sections which needed splitting up, this research chose to use relative population size as the means to allocate telephone subscriber numbers proportionally to individual settlements, as the availability of data to measure historical urban size online rendered this the most credible option. Information on the reported populations of settlements with more than 50,000 residents was obtained from the Great Britain Historical GIS Project (2017). Because this information was only available for 1901, 1911 and 1921, the population sizes used for calculating multipliers was that of the year closest to the year of subscription data concerned - meaning that, crucially, the combined sections from the 1921 directories and onward all used the 1921 settlement populations for calculations.

As an illustration, Table 10 depicts how multipliers were calculated to separate subscribers in the 'Birmingham and West Bromwich' section of the 1921 directories. Because 1921 was one of three years for which population size data for the 50 largest British settlements was available, the calculations utilised the 1921 population sizes of the aforementioned settlements. Taking each settlement's population, divided by the combined population of the two settlements, gave the multipliers of 0.926x and 0.074x for Birmingham and West Bromwich, respectively. These ratios were finally used to split the 6,946 subscribers in the concerned directory section, to give the final figures of 6,432 subscribers in Birmingham and 514 subscribers in West Bromwich.

Table 10: Worked example of how multipliers were calculated to allocate the subscribers in a section proportionally to its constituent settlements

Settlement	**Population** *(1921)*	**Multiplier Derivation**	
Birmingham	919,444	919444 / (919444+73647) =	**0.926x**
West Bromwich	73,647	73647 / (919444+73647) =	**0.074x**

With the passing of time, telephone provision expanded from the city centre of Birmingham to encompass nearby, smaller settlements. Beginning in 1931, the directory listings for *"Birmingham and District"* often included the subscribers of Walsall and Wolverhampton, in addition to West Bromwich, which in this analysis were treated as entities standalone from Birmingham. Subscriber numbers from listings of *"Birmingham and District"* were thus divided in a similar manner to that in Table 10, proportionally to their populations, except across the four settlements instead of two.

Admittedly, an underlying assumption of the method of multipliers proposed is that, per unit area, the rates of adoption of telephones among the residents of the adjacent settlements were equal. This is not likely to be the case since there are often stark socioeconomic divides in neighbouring settlements, especially where one dwarfs the other(s) in population. For example, there are both more likely to be more residents, and a greater ability to afford telephones, in Glasgow City than Paisley, or in Birmingham than West Bromwich. The outcome of proportionally dividing the subscribers of these combined areas is an underestimation of the subscription rates in the largest settlement, and an overestimation in the smaller settlement(s). However, this research argues that a misestimation of individual settlements' subscriber counts, as in this case, is still superior to the alternative of being unable to separate subscribers in the "conjoined" settlements – in which case there would be no data points for the individual settlements in the years that are affected. The other method considered, of attributing individual records in affected directory sections to a settlement based on the captured 'Telephone Exchange' information, also appeared to introduce even more unquantifiable uncertainty owing to the data capture process.

Having addressed the larger issue of inconsistent geographical boundaries, multipliers were also used to remedy a less frequently occurring issue in deriving subscriber counts. Where the first kind of multipliers was used to split up the subscribers in one directory section and allocate them to two or more settlements, the multipliers referred to below were used to consolidate the subscriber counts for a single settlement, if it had featured in more than one directory sections in the same year. Ordinarily, for each such settlement, I calculated the subscriber numbers in every relevant section and took the arithmetic average of these figures to derive the final tally. The assumption underlying this calculation, however – that each section pertaining to said settlement was a full listing of all its subscribers –, was not always applicable, thereby necessitating the need for multipliers.

In any one year for a given settlement, multiple directory sections may hold information about the same settlement, most commonly because the records are updated commonly at intervals of between six months and two years. For instance, it was sometimes the case that directory sections for larger settlements listed changes and updates in subscribers since the last publication. In this way, subscriber counts from such directories would comprise at most 20% of the total subscription in those settlements, as exemplified by Liverpool in 1941, where one directory section, 240 pages long, was supplemented by another directory section that was only 20 pages long (see Table 11). The default operation – of averaging the number of subscribers contained in these two directories – would thus produce far from a reasonable estimate of the total subscription volume in Liverpool that year, since it assumes both directory sections were complete listings in themselves. The multiplier that was applied in this case (see Table 11) served to undo this erroneous averaging operation by, in effect, assigning weights to each contributing directory section based on their relative lengths (1 and $^{20}/_{240}$), in place of giving them equal weights, thus giving a multiplier of 1.85x to be applied to Liverpool in 1941.

Table 11: Worked examples of how multipliers were applied to settlements which occurred in more than one directory section, in the same year

Settlement	Irregularity	Multiplier Derivation	
Liverpool (1941)	One directory section, labelled "Liverpool Area Supplement", is an update to the other listing of the Liverpool area. It spanned just *20* pages, compared to the other's *240* pages.	$(\frac{(1+^{20}/_{240})}{2})^{-1}$ =	**1.85x**
London (1931)	The 2 directories were split by starting alphabet, and each only contained about half *(0.5)* of all subscribers.	$(\frac{(0.5+0.5)}{2})^{-1}$ =	**2x**

**Note: These multipliers work on the assumption that settlement counts from the different directories concerning one settlement have already been mathematically averaged, and thus that they each hold a roughly equal number of that settlement's subscribers. These multipliers therefore work by undoing that default calculation.*

Along similar lines, a peculiarity for London data, beginning 1911, was that the settlement's subscribers numbered so many that they had to be consistently split across multiple directories, based on the starting alphabet of their registered names. In 1931, for example, the London telephone subscribers were listed in two separate volumes, with those with names beginning A-K in the first, and names beginning L-Z in the second. The default averaging operation would assume that each contributing directory was a full listing of all of London's subscribers that year, when they in actuality only held about half of all subscribers each. The multiplier that was eventually applied was 2x (Table 11), and in effect allows for the summation of the counts from each directory that year, rather than conducting an averaging operation.

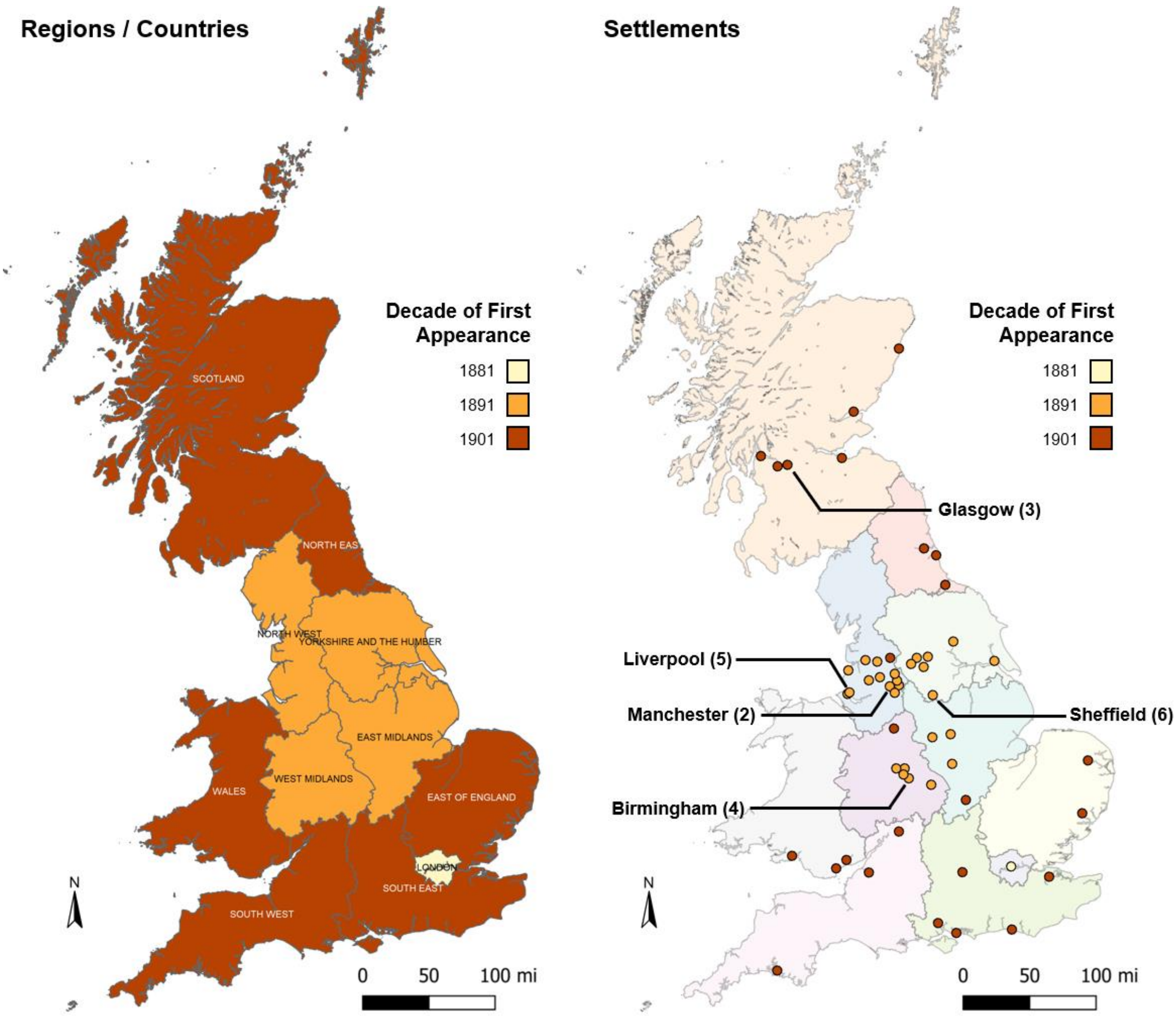
In summary, the calculation of telephone subscriber counts in each settlement required modification in several instances due to reasons of misalignment between civil and telephone exchange geographies, and because some directory sections contained only partial listings of the telephone subscribers of a given settlement. By default, when multiple directories contained subscriber information for one settlement in a given year, this research took the average count of records in these directories, assuming that the listings in each directory was complete. However, corresponding to the two kinds of exceptions aforementioned, this research found it sensible to apply two kinds of multipliers to moderate the final estimates produced. The specifics of the settlements and the years to which multipliers were applied can be found in Appendix Table 1.

3.2 Regional Variations

Having described how this researched obtained the counts of telephone subscriptions and moderated them based on aforementioned concerns, this following section outlines regional differences in telephone adoption. The analysis relates to 53 urban settlements that were chosen based on their position within the Top 50 in Britain in 1901 in terms of one of these criteria: population size, number of households, diversity of surnames represented and number of street segments (Lan *et al.*, 2021). While the overwhelming majority of settlements are top ranked by all measures, few smaller settlements make the list just one or two times, thus giving rise to the appearance of 53 settlements across the four Top 50 lists. Because of their primacy in the urban network of Britain, often in terms of multiple of these criteria, it is thereby sensible to assume that they would be at the frontiers of telephone adoption in the country before it trickles down to other more peripheral parts of the country, even if this sample would not be representative of the country, especially of rural Britain.

Figure 8 below illustrates the initial expansion of the service areas of the telephone service. Beginning with only service provision in London in 1881, by the next decade, the service coverage expands to four English regions from the Midlands and upward, with the exception of the North East. This pattern of expansion in the second decade of widespread operations seems sensible for telephone providers in retrospect: together, the West Midlands, East Midlands, North West and Yorkshire and the Humber account for 30 of the chosen 53 British urban settlements and of which but three settlements were recorded to not have had received telephone services by 1891. By 1889, three British telephone companies had amalgamated to form the National Telephone Company, which then went on to acquire other smaller telephone companies with regional service niches (British Telecommunications, 2021). This strategy therefore probably enabled them to consolidate their market dominance in a market comprising a very considerable number of the most densely populated areas in England, with another development supporting this theory being that the consolidated company had in 1890 unveiled a trunk circuit connecting London and Birmingham, thereby opening the routes of telecommunication between the capital city and the Midland and Northern Counties (ibid.). Finally, the four said English regions that were first recorded in the directories in 1891 also encompass four of the five most populous British settlements in 1901 (excluding London), further reaffirming the likely attractiveness of the combined area for the then-rapidly expanding companies.

Figure 8: Maps showing when different regions, countries, and settlements in Britain were first served by the telephone network

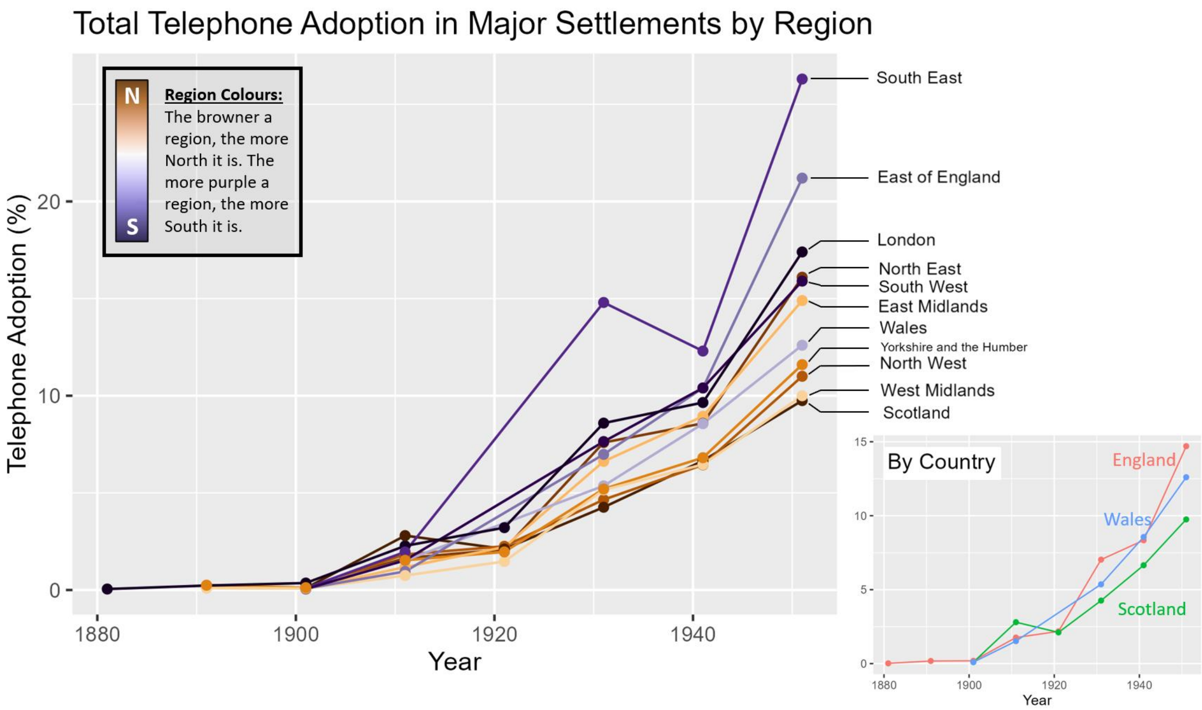


Compared to figures published by Calvo (2006), which estimated that over 95% of the 60 largest British had had access to the telephone by 1892, the findings from the digitised telephone directory archives present a more delayed picture of the diffusion of the telephone in Britain. These differences are likely attributable to the fact that the coverage of the telephone directory archives is not complete in the earliest two decades of service, when the telephone industry in Britain was still highly fractured and many telephone companies held provincial licenses to operate in different parts of the country (The Economist, 1904). For this reason, British Telecommunications, as provider of the raw data used in this research and the successor of the Post Office which had come to manage all telephone services following nationalisation in 1912, may not have had access to early records of telephone subscribers in regions served initially by other licenced companies, for instance Scotland (by the National Telephone Company) and the South East and South West of England (by the South of England Telephone Co.) (Kingsbury, 1915).

By 1901, telephone service provision had extended to cover all remaining, populous settlements which had not already been served a decade prior, notably including coastal settlements around England, the South of Wales and Scotland. What also stands out from Figure 8 is that, with the exception of Glasgow, which neighbours Paisley, other settlements that had just appeared in the directories in 1901 seem visibly more isolated from other populated settlements. In contrast, those that had already been recorded in the decade prior tended to be proximal to other settlements among the 53 chosen, to name some examples: Liverpool and Birkenhead, Birmingham and West Bromwich and also Manchester with Oldham, Rochdale, Ashton-under-Lyne and Stockport. This virtual twinning of settlements has practical implications on the analysis this chapter reports, especially since the telephone directories sometimes make no overt distinction between subscribers of these actually separate settlements.

Beyond the start of service provision, the growth in telephone subscribers in major settlements over the time period 1881 to 1951 are shown in Figure 9 below, as grouped by regions (only for English settlements) and subsequently countries. For purposes of this analysis Welsh and Scottish settlements were not sorted into smaller geographical units as they numbered but 6 and 3, respectively – figures near the average representation of individual English regions – and sorting may leave some units only containing 1 or 2 settlements, which this analysis did not deem meaningful. The larger chart in Figure 9 employs a diverging colour scheme to highlight the ordering of the regions of England and two other countries geographically from North to South.

Figure 9: Line chart of overall telephone adoption rates across time in regions (left) and constituent countries (lower-right) of Britain, 1881-1951



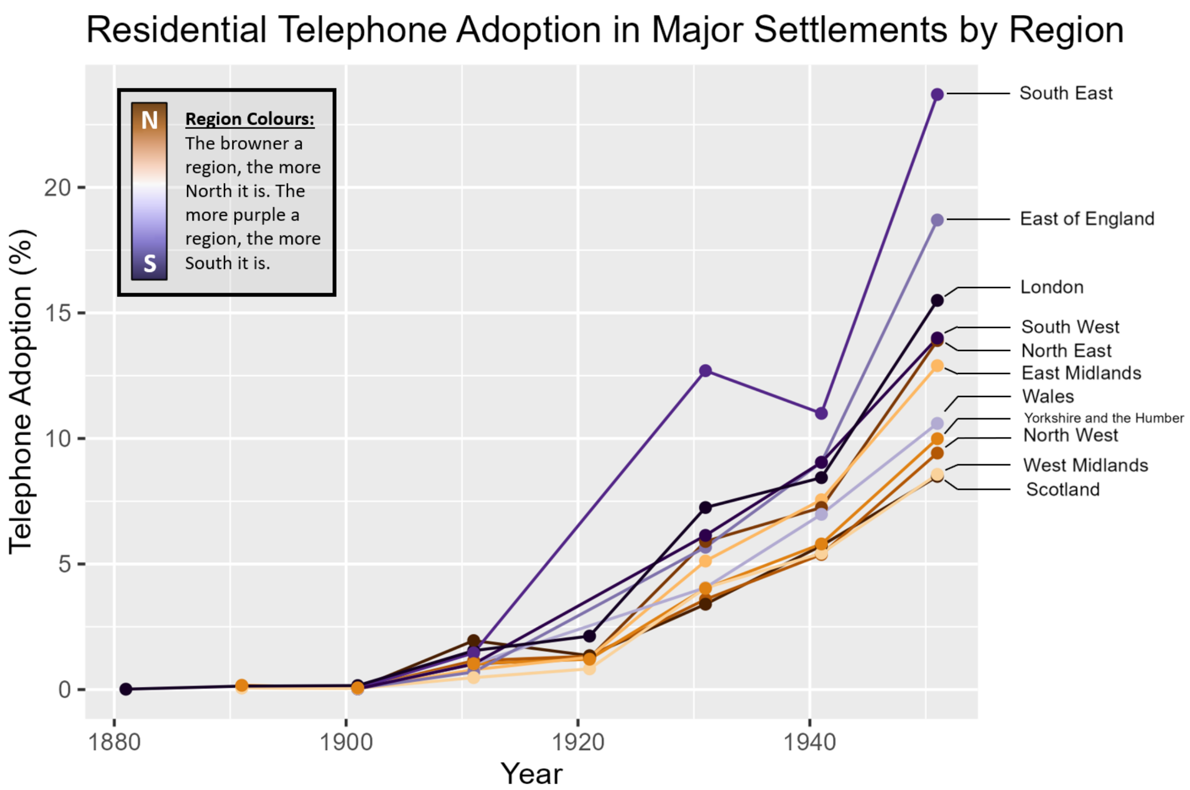
Calculations of adoption rates were made by dividing the number of processed subscriber records per settlements by the actual (1901-1921) or extrapolated (all other decades) settlement population sizes, where extrapolations were done by assuming that the decadal change in population of settlements equalled that of the UK country they were subsumed under. Using the baseline figures from 1901 to 1921 was deemed as most feasible because only in the 1921 Census Report did there exist a collated list of population sizes of the most populous settlements in England and Wales, covering 88.7% of the settlements of interest; the alternative method of copying and pasting the figures for each settlement for each decade manually appeared much more prone to human error and therefore less credible. As the rates of population change in the most populous settlements in each country would weigh disproportionately on the overall rates of change, this research believes that the use of the latter as approximators of the former is a reasonable trade-off between time-efficiency and accuracy of analysis. It is likewise that this research chose the population size rather than number of households as the denominator for the analysis, despite residential telephone subscriptions typically being made at the household level.

On the whole, total telephone adoption rates in the major settlements of Britain followed a steady upward trajectory with the South East of England being consistently ahead from 1931 and which in 1951 had a number of registered telephones equal to 26.3% of its resident population.

It is followed then by East of England and London with telephone subscription rates of 21.2% and 17.4% of its population, respectively. Interestingly, London ranks just third in subscription rates in 1951 even though it leads every other settlement by far in terms of absolute telephone subscription numbers, as this chapter later shows. Rather apparent from Figure 9 is that while the South East leads all other regions in adoption rates beginning 1931, this figure dips in 1941 before rising again. One possible explanation is that it is a region which consisted of exclusively mid-size to less populous settlements, with Brighton ranking most highly at 17th and Reading the lowest at 39th. The cumulatively smaller total population of settlements in this region could have resulted in greater instability when used as a denominator in the calculations of adoption rates, with a further exacerbating factor that the grouping of Chatham with other settlements in sections of the directory changed almost every ten years, making it more difficult to keep calculations of its subscription rates consistent. Within England, a division also emerges between the Southern, more avidly telephone-adopting regions and those less so in the North. Demarcating the South as regions below the Midlands, the Southern English region with the lowest telephone uptake of 15.9% is the South West, a figure just 0.2% shy of the North East, the region with the highest uptake in the North. This is notable as Southern regions (excluding London), despite receiving telephony services later than most of the Northern regions, over half a century becomes by far the keener adopter of telephones.

At the country level, telephone adoption rates in the largest settlements of Scotland consistently trail the average of counterparts in England and Wales, with Scotland averaging an uptake of 9.75% in 1951, just lower than the slowest adopter among the English regions of West Midlands at 10%. Wales, meanwhile, had an adoption rate of 12.6%, making it below average when compared to English regions and the figure for England as a whole was 14.7%. The patterns of overall telephone uptake are largely echoed in the uptake of residential telephones (see Figure 10) and this is expected since in 1951, residential telephone subscriptions accounted for 87.5% of all subscriptions made in the 53 chosen settlements.

Figure 10: Line chart of residential telephone adoption rates across time in regions and constituent countries of Britain, 1881-1951

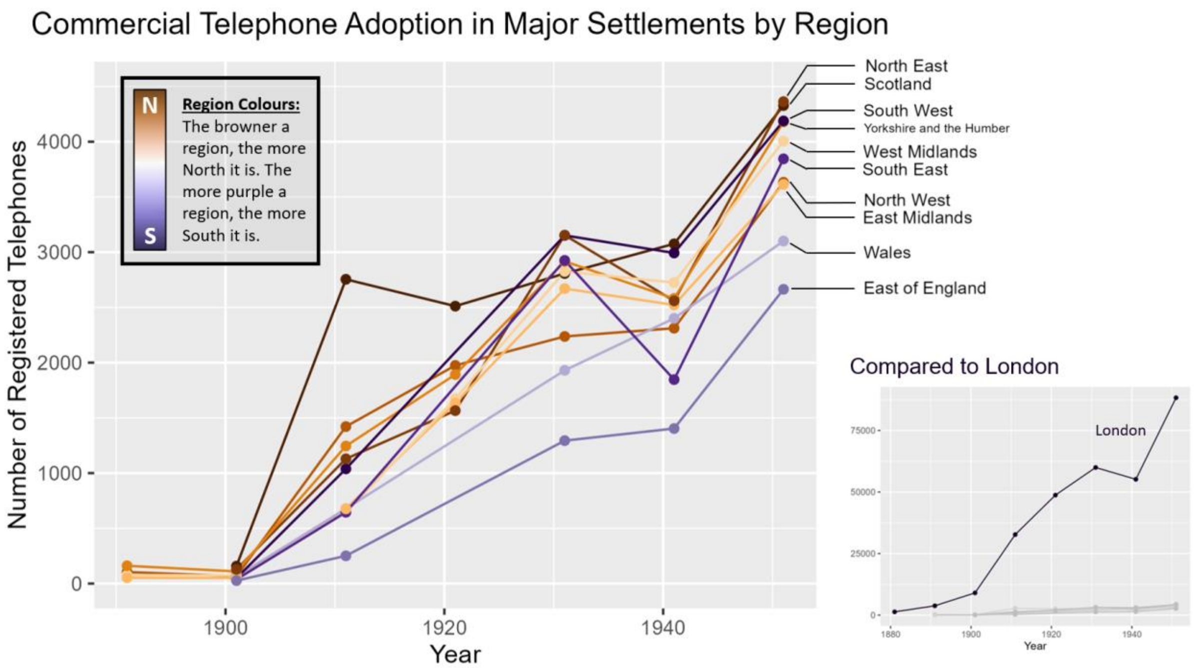


While the focus of this research has been telephone usage as they relate to residential settings, and while the majority of telephone subscription records was residential in nature, this research nonetheless found a brief exploration of patterns of commercial telephone subscription worthwhile. Especially for interactive technologies, including for the telephone, their visibility to the population segments that had yet to become adopters was one important factor in influencing their potential and eventual decision to use the technology (Jahangir and Zia-ul-Haq, 2023). In this way, commercial telephone adoption could be argued to exert some influence on residential telephone uptake, as a person would have more awareness and potential interaction with the telephone if the businesses they patronised already had one. Conversely, a higher rate of residential telephone adoption would also likely incentivise commercial entities to adopt the telephone in order to better reach their customer base.

To pre-empt this exploratory analysis, commercial telephone subscribers were first separated from residential subscribers through conducting keyword searches in the 'name' field of the digitised directories dataset. As the processing pipeline was modified for robustness to the directories of more and more decades, common terms that very probably denoted a subscription record was commercial in nature were pooled together into a list that was then iteratively used to reclassify the records at different stages of this analysis. Some examples of these terms included

‘& Sons’, ‘Limited’ alongside the names of various professions and industries like ‘Attorney’ or ‘Bank’. Although this way of classification was far from deterministic, there was no other feasible way to undertake this task without involving the laborious hand-coding of millions of records. The results of this analysis are shown in Figure 11, which charts the total number of commercial telephone subscribers per settlement included in the analysis per region or country. In contrast to the previous charts, the numbers shown therein are absolute figures, rather than ratios, because this research saw the commercial use of telephones as much less directly tied to the population or household size of a settlement than that of residential subscriptions. Moreover, the absolute number of commercial subscribers in a settlement could also indicate the occurrence of commercial activity with links beyond the settlement in which it was based hence also reducing how meaningful such a relative measure as before would be.

Figure 11: Line chart of number of commercial telephone subscribers across time in regions and constituent countries of Britain, 1881-1951

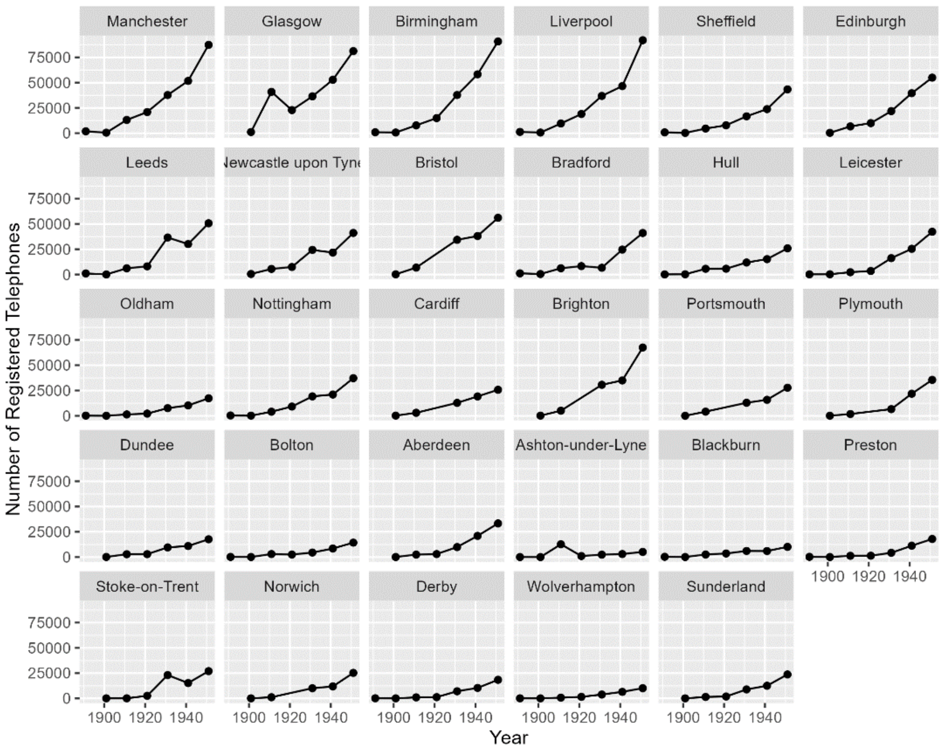


It is clear from Figure 11 that the disparity of commercial telephone subscription volumes between the extremes of regions/countries is much less pronounced than the equivalent disparity in residential adoption rates. Where for the latter case, the top region (South East England) had nearly three times the adoption rate of the bottommost (Scotland), the region with the most commercial subscribers, North East England, has just approximately 1.4 times the number of subscribers in the region with the fewest, Wales. A reversal of an earlier trend also takes place, with mostly Northern English regions and Scotland ranking in the upper half when excluding London: all Northern regions but the East Midlands surpassed an average of 4,000 commercial subscribers per settlement in the tally. The results thus hint at a distinction of the nature of telephone subscription in the South being more personal or residential, and of that in the North being more commercial, notwithstanding the fact that merely 1 in 8 of all records included in this analysis are of commercial subscribers. Finally, the relegation of London to the smaller subplot was a decision made due to its markedly different scale of commercial subscription volumes, with its figure of 88,290 commercial subscribers in 1951 dwarfing even the total of all the other countries and regions combined.

3.3 Variations Across Settlements

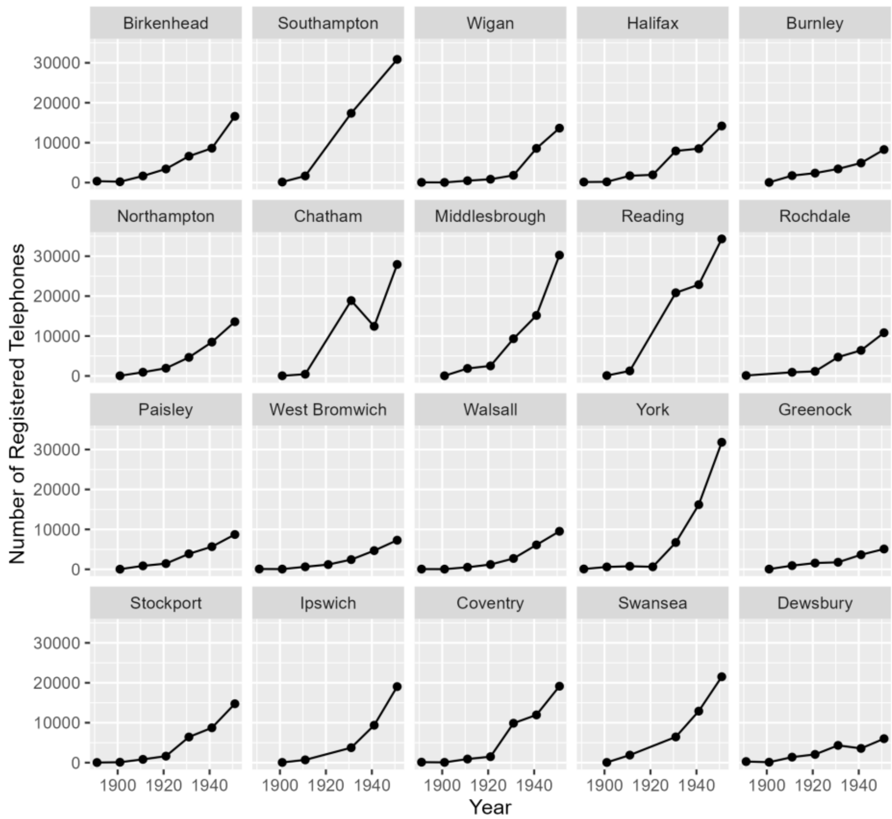
This section proceeds to build on the regional descriptive analysis presented in the previous section at the more granular level of individual settlements, with Figure 12 showing patterns of growth of telephone subscription in the 30 most populous settlements in Britain as of 1901 in descending order of population. London once again is excluded from this plot as the volume of its telephone counts as of 1910 far surpasses the number of subscribers in each of the next four largest settlements by at least tenfold. Almost without exception, telephone uptakes increase continuously across the 29 settlements over the early 20th century and the volume of subscribers appears roughly proportional to the population of the settlement. There exist kinks in the outputs where subscriber counts increase one decade and then fall the following decade, such as in Glasgow and Ashton-under-Lyne in 1911, Leeds and also Stoke-on-Trent in 1931. A possible explanation is that these named settlements have in common a tendency to be combined in sections of telephone directories with the listing of subscribers of other settlements and the method of resolving this complication described in Section 3.1.2 could have proven unable to do so perfectly.

Figure 12: Faceted plot of total number of telephones per settlement for the 30 most populous settlements in Britain as of 1901 in descending order (excluding London)



Similar to the variations in telephone adoption volumes are the variations in rates of growth according to expectation: the gradients for the graphs of the largest settlements such as ranked between 2nd (Manchester) and 19th (Plymouth), show marked increases in the latter decades despite having initially begun with a much gentler slope. This trend alludes to the S-shaped curve of innovation diffusion that Ryan and Gross (1943) was among the first to postulate, in doing so asserting that the diffusion of an innovation takes place in five stages; in its second stage, innovation diffusion accelerates as a new group of early adopters begins to buy into the invention, while ultimately still remaining in the minority of society. Among the aforementioned settlements, a distinction can also be seen in the timing at which this acceleration of uptake began: in the graphs for settlements that are more highly ranked and for which data was already available in 1891 as a point of comparison, such as Manchester, Birmingham, Liverpool and Leeds, the kinks in the graph reflect an accelerated uptake that can be said to have begun since 1901. What was a flat line segment depicting relatively small increases in subscription volumes from 1891 to 1901 in the graphs representing these settlements, give way to a more pronounced positive gradient from 1901 to 1911, signalling more rapid increases in subscription in each decadal period. This corresponds to the introduction of the charging-per-call service model in 1900, replacing in many parts of the country the previous mode of flat-rate subscriptions, which likely pushed many potential subscribers who had been on the fence about the telephone to finally adopt one; this hike in the demand for telephones as well as in the number of calls made in Britain mirrors what had happened after similar changes were made in the US in the mid-1890s (Connected Earth, 2006). These observations, again, allude to the process of innovation diffusion being hierarchical in nature, as it cascades from the urban settlements where more of the population and resources concentrate to settlements lower down in the urban hierarchy. For example, it is shown that this pattern occurs in Leeds and Leicester only in 1921 and Bradford and Plymouth in 1931.

Figure 13: Faceted plot of total number of telephones per settlement for the 31st to 50th populous settlements in Britain as of 1901 in descending order



Patterns for the uptake of telephones in the settlements ranked 31st to 50th in terms of population in 1951 are shown in Figure 13. Total subscriptions in some of these less populous settlements actually equal and in some cases even exceed those of some settlements ranked in the first 30. Especially outstanding are settlements which by 1951 have registered more than 30,000 telephones, namely Southampton (32nd), Middlesborough (38th), Reading (39th) and York (44th). Accompanied by markedly steeper gradients of telephone growth, their adoption rates are all the more notable considering their smaller sizes. It was, in general, difficult to find historical sources that provide detailed accounts of how the telephone was introduced to and received by the society and economy in these markedly smaller British settlements. However, in the case of York, the lowest-ranked settlement to exceed 30,000, a possible explanation in the disproportionate uptick in telephone may be its route centrality on the railway network. Both physical communication networks (as is that of the telephone) and transport networks are products of the same social networks and processes that have existed historically (Sawyer, 2005). When railway traffic passing through York, as well as their speeds of carriage, dramatically increased in the 1850s due to both Southbound (to London) and Northbound (to Scottish cities) routes that were introduced, York's popularity as a stopover destination saw a corresponding hike, even if the resident population in the settlement remained modest (Victoria County History, 1961). The increased importance of York as a transport hub in the following decades may therefore explain why the settlement punches above its weight in terms of subscription rates to the telephone.

Finally, the counts of telephone subscribers in different settlements will hereafter be validated against accounts that this research has found in various literatures around the use of the telephone. Mentions of telephone subscription volumes in British settlements, especially other than London, were hard to come by. Sources which report some indication of telephone usage volumes in London are depicted in Table 12, alongside a comparison of how these figures match up against a tally from the information digitised by this research, either from the same year or in the year nearest to the data point presented by the source which is also available in the processed dataset created. Where there were such mentions, they usually related only to the same few settlements such as Glasgow, Portsmouth and Guernsey, the services of which had remained in the charge of local/provincial telephone companies until much later than 1912, when control of most other telephone services had become centralised under the wing of the Post Office. These are presented in Table 13 and in a similar format to the table used for validating subscriber counts in London.

Table 12: Collation of historical sources containing figures relating to subscription volumes in London between 1879 and 1913

Source and Further Details	Reported Figure in Source	Information from digitised dataset		Margin of Error
		Year	Figure	
London had 1.6 Subscribers per 1,000 population in 1879. (Table 1 in Calvo, 2006).	1,797,486* / 1,000 * 1.6 = **2,876**	1881	602	-79.1%
"By 1882, there was one phone for every 3,000 people in London; by 1890, the ratio was up to one in about 800…" (Hamill, 2010: 277).	1,797,486* / 3000 = **599**	1881	602	0.01%
	4,211,743* / 800 = **5,265**	1891	5,596	6.29%
"For instance, in the metropolitan area of London [there were] 70,000 subscribers… at the beginning of the twentieth century." (Calvo, 2006: 421)	**70,000**	1901	16,253	-76.8%

"By 1910-12, there were some 600 thousand phones in Britain and a quarter to a third of these were in London." (Hamill, 2010: 277)	**150,000 ~ 200,000**	1911	102,619	-48.7% ~ -31.6%
In 1910, London accounted for 46.4% of all subscribers in the UK. (Table 2 in Calvo, 2006).	**46.4%**	1911	35.27%	-24.0%
"As late as 1913, London accounted for over one third of the telephones in the entire country." (Perry, 1977: 76)	**33.3%**	1911	35.27%	5.92%

** population figures for London in these years were taken from the Great Britain Historical GIS Project (1917), as their source numbers were not stated in the respective sources.*

Data points for London were found in every decade from 1881 to 1901. With the exception of some statistics cited – that is, of figures for 1882 and 1913, as derived from Hamill (2010) and Perry (1977), respectively – the margins of error of the estimates produced by the dataset digitised in this research are negative, meaning that statistics generated from the dataset this research created have a recurrent pattern of underestimating the actual number of telephone subscribers in London at a given point in time, assuming the figures in the references are accurate. With the largest margin of error, the number of subscribers in London based on our data in 1881 (as the closest point in time) underestimates by 79.1% the number of subscribers that were in London in 1879, according to Calvo (2006). This could be explained, firstly, by the fact that two telephone companies – the National Telephone Company and the Post Office – were in fierce competition in the London telephone market then (Perry, 1977) and that the archives upon which this research is based only contains subscriber data for one of these companies, thereby underestimating the overall subscriber numbers. However, by virtue of subscription volumes being lowest in the early years of the telephone, estimates for 1881 may be susceptible to a greater degree of error than that for a later decade. Estimates produced by this research for the subscription volumes in subsequent years had margins of error of between -76.8% and +6.29%, and this means that the coverage of the directories made available to this research is more than likely to not include the full range of telephone subscribers from London in the years up till 1911. However, with how nicely rounded some of these figures were, with wordings that sometimes made them appear as though they were estimates rather than exact data (Table 12), it is also likely that they should not be treated as precise validation data, but rather as ballpark estimates of the actual numbers.

Furthermore, the issue of determining an accurate cross-reference of the number of telephone subscribers in a given year is illustrated by the references in Table 12 to data for 1881 and 1911. In trying to validate the subscriber counts produced by this research for 1881, the statistics given by Calvo (2006) and Hamill (2010) already give estimates that differ by a magnitude of slightly greater than 4, of 2,806 and 599, respectively. In estimating the proportion of all national telephone subscribers that London subscribers accounted for, the estimates given by Calvo (2006) and Perry (1977) were also quite starkly different, at 46.4% and 33.3%, or one-third, respectively. Although both contrasting figures in these two cases must have some basis, this research could not definitively make a conclusion on which was more accurate, either because some of the sources they cited were not publicly available sources (and also usually very dated material), and/or because the figures used to calculate some of these summary statistics was not made available. Most notably, it was not clear which population figures of London, or where they were derived from, was used to calculate the rates of subscription. The population figures that this research used in its calculation to estimate a numerical figure based on statements they made are those from the digitised Census Reports from the Great Britain Historical GIS Project (2017).

Table 13: Collation of historical sources containing figures relating to subscription volumes in settlements other than London between 1905 and 1907

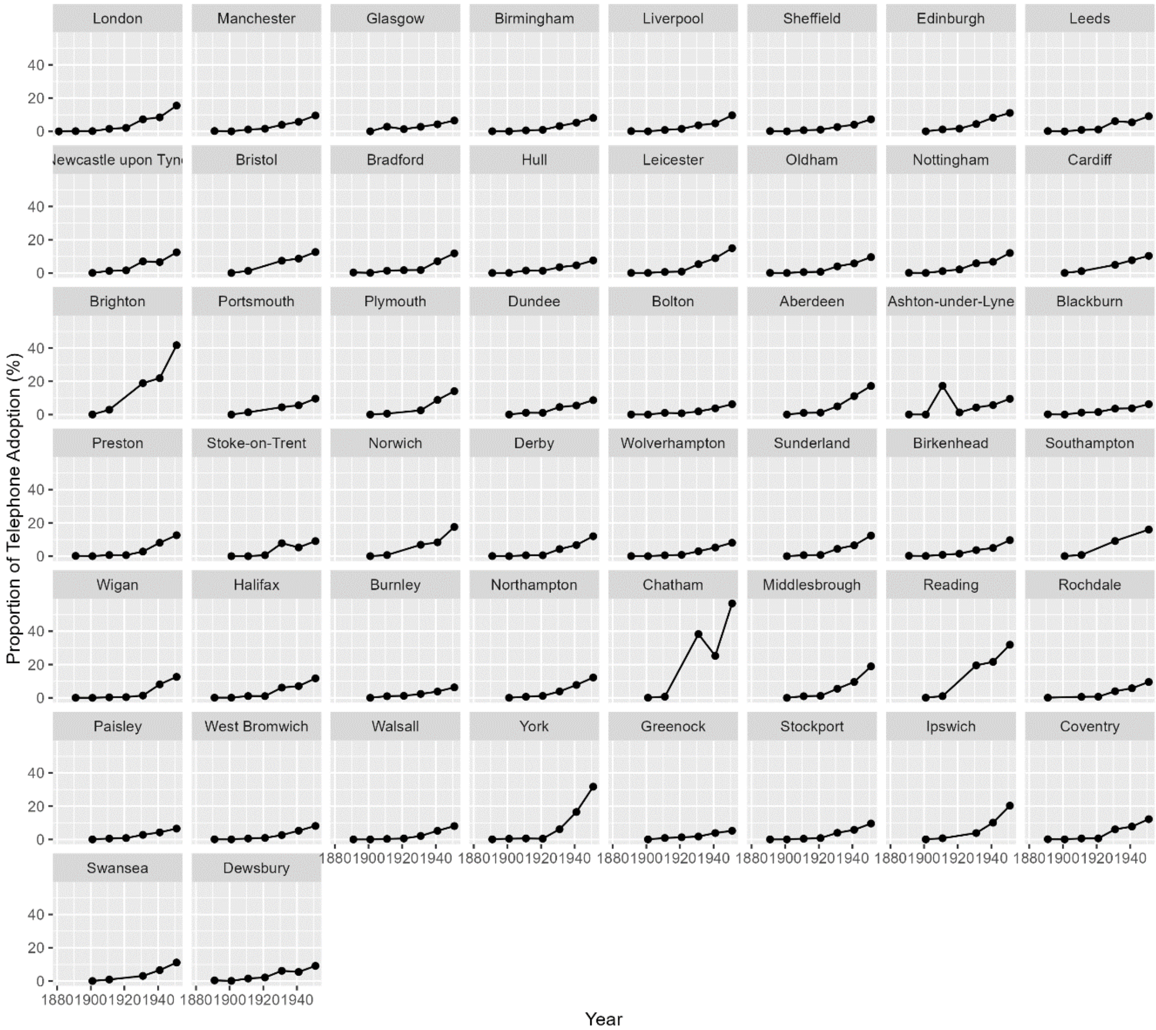
Settlement and Year	**Source and Further Details**	**Reported Count in Source**	**Count from digitised dataset**		**Margin of Error (%)**
			Year	**Count**	
Glasgow, 1905	*In 1905, there were 12,300 telephones.* (Appendix 1 in Perry, 1977)	12,300	1901, 1911*	20,998	70.7%
Glasgow, 1907	*"1907: The number of subscribers to the Post Office provincial exchanges… [in] Glasgow [was] 11,103."* (Hemmeon, 1912: 236)	11,103	1901, 1911*	20,998	89.1%
Swansea, 1905	*In 1905, there were 1,400 telephones.* (Appendix 1 in Perry, 1977)	1,400	1901, 1911*	2,606	86.1%
Brighton, 1905	*In 1905, there were 1,900 telephones.* (Appendix 1 in Perry, 1977)	1,900	1901, 1911*	2,078	9.4%
Hull, 1907	*"Hull and Portsmouth were the only towns maintaining municipal telephonic systems in 1907, Hull having 2128 telephones in use and Portsmouth 2553."* (Hemmeon, 1912: 236)	2,128	1901, 1911*	2,912	36.8%
Portsmouth, 1907		2,553	1901, 1911*	2,078	-18.6%

* *average taken of 1901 and 1911 counts to give an estimate of the 1905/1907 figures*

Table 13 shows validation for a select few settlements other than London. As mentioned before, these accounts were in general even rarer than those for London and were usually available for settlements whose telephone service provision had not been absorbed into the operations of the largest telephone companies, which refer to the United Telephone Company, the National Telephone Company and the Post Office. Because these settlements had, at the point of enumeration, retained their independent local telephone systems, their subscription information would be listed separately from, rather than included as part of much more spatially aggregated totals. The trend for all settlements listed above but Portsmouth is the reverse of that for London, whereby the estimates based on the dataset created in this research tended to overestimate the number of telephone subscribers, by as little as 9.4% for Brighton and by as much as 89.4% for Glasgow in 1907. Keeping in mind that the relevant citations found and listed in Table 13 cover a very limited time period, it was difficult to pass judgment on why the discrepancies might have come about. At the same time, some of these references seem to contradict each other: while, as seen from charts presented earlier in the chapter, subscription volumes in all settlements rose with time, almost without exception, the estimates given by (Perry, 1977) and (Hemmeon, 1912) yet suggest that the subscription volumes of Glasgow fell 9.7% between 1905 and 1907, if the figures are taken to be comparable. All in all, while contradictions exist, the validation data that had been collated for London and the other settlements alike act as a sanity check to the data generated by this research and confirm that, at the least, they have been in roughly the same order of magnitude as the data referenced across the historical literature.

Hereafter, the discussion of findings shifts in focus to the rate of telephone adoption for residential use in all selected settlements. Figure 14 summarises the findings and displays information for all top 50 most populous settlements while also including Coventry, Swansea and Dewsbury which while were unranked in terms of population size, had made the top 50 settlements in one of the three other criteria set out by Lan *et al.* (2021).

Figure 14: Faceted plot of residential telephone adoption rates for the 53 most populous settlements in Britain as of 1901 in descending order

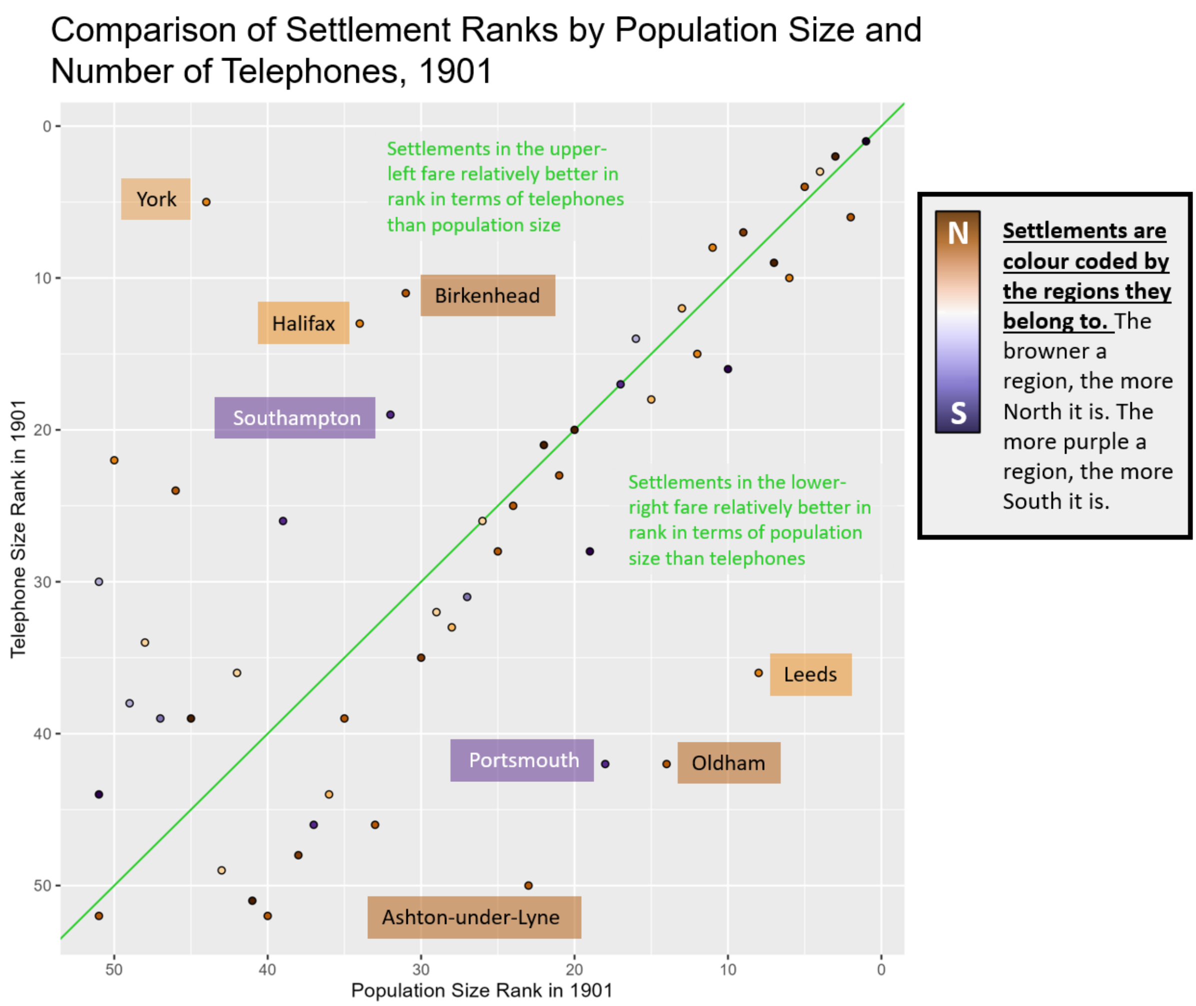


The vast majority of settlements have an adoption rate of between 10% and 20%, with most of the settlements ranking between 2nd (Manchester) and 11th (Bradford) having a residential adoption rate of close to 10%. This overall figure would also feasibly tally with the official statistic that by 1970, or nearly twenty years after the last data point, just 35% of British households had access to a household telephone (Office for National Statistics, 2019).

Some irregularities in the results proved challenging to correct for. For one, sudden spikes or dips in the graphs of Ashton-under-Lyne (1911) and Chatham (1931-41) reflect likely misestimations in the proportion of subscribers in composite directories that belong to these respective settlements. The graphs for some settlements also omit missing data points, notably Southampton in 1921 and 1941 alongside many other settlements in Southern England and Wales in 1921. This is attributable to the fact that, as the number of telephone directories published annually grew, the schedules at which new directories for different regions were published became staggered and irregular. It was thus the case that the couple of missing data points in 1921 belonged to regions which did not receive updated directories in that year. The strategy used to attempt to remedy this shortfall was to extend the data capture as far as possible temporally, so that general trends could be observed in spite of missingness.

Finally, Figure 15 shows a comparison of the 53 settlements as ranked by population size and by number of telephones in 1901, where settlement points lying on or close to the green diagonal line are proportionally ranked on both measures. Conversely, those lying further away from the diagonal are ranked disproportionately higher on one metric than the other.

Figure 15: Comparison of settlement ranks in terms by different measures of size, 1901



On the whole, there does not appear to be apparent partitioning of the settlements in the chart, based on their geographical locations. Of some of the most major deviations in Figure 15 that are highlighted, two settlements are from the South of England, Southampton and Portsmouth. Like most other highlighted settlements, they are mid-sized settlements ranking near the middle of the Top 50 list in population rank. However, unlike the other Northern settlements, Southampton and Portsmouth belonged to a region – South East of England – which only in 1901 was listed for the first time in the telephone directories and thus were later to receive telephone service coverage than their Northern counterparts. This relatively late introduction of telephone services to Southampton and Portsmouth might explain why their population rank in comparison to their telephone size rank in 1901 was rather imbalanced.

Turning to the Northern settlements, the anomalies could be divided into two groups, the first of which were generally more populous settlements that had underperformed on the telephone subscription metric, namely Leeds, Oldham and Ashton-under-Lyne. An explanation for this undersubscription may be derived from their historic functions as industrial urban centres. Oldham and Ashton-under-Lyne, in particular, were centres for the cotton and textile industries, with the former even having once been coined the most productive cotton-spinning town in the world (Gurr and Hunt, 1998). While this manufacturing role was likely to draw large working-class crowds from neighbouring villages to Oldham and Ashton-under-Lyne, this was not necessarily the segment of society that needed, or could pay for, a telephone in its early years. This predominance of manufacturing employees may explain why the three aforementioned settlements ranked more highly in terms of population than in terms of telephone subscription rates.

The second group of anomalies in the North were mid-sized to smaller settlements that had proportionally higher telephone size ranks, namely Birkenhead, Halifax and York. Birkenhead, which for the entirety of the period of interest in this research, had always been listed in the same directory sections as the adjacent, larger settlement of Liverpool. As had been discussed in Section 3.1.2, the telephone subscription rates in Birkenhead, as the smaller of the twinned settlements, would have been overestimated as a result of the way multipliers were applied. This positive bias for Birkenhead might have been exacerbated by the fact that Liverpool, as a major international port, would also have more links than usual to, and thus more need to communicate with, other parts of the country (and the world). This would probably have translated into an increased propensity for telephone subscription among the residents of both Liverpool and Birkenhead. While not consistent through the years, Halifax was sometimes also listed together in the directories with the adjacent Bradford, which had nearly triple its population in 1921, and this might have led to a similar positive bias in the subscription numbers for Halifax.

Like Liverpool (and thus Birkenhead), the disproportionate rates of telephone subscription in York, relative to its population size, could also be explained in part by its high level of transport connectivity. In the mid-nineteenth century York benefitted considerably from the expansion of the rail network in England, together with its advantageous position which put it on many national railway routes (Victoria County History, 1961). However, the massive industrial expansion that occurred in other settlements did not occur and instead, York's economy remained primarily anchored by smaller enterprises (ibid.). Rather than attract swathes of people that were unlikely consumers of the telephone – as in the case of Oldham and Ashton-under-Lyne –, York's development kept its population growth relatively stable, while its incumbent population, many of whom were small business owners, more resembled the profile of an early telephone adopter. On the whole, historical urban functions may provide clues as to why some settlements defy expectations in the comparison of their population and telephone rank sizes.

3.4 Limitations

The analysis described above remains admittedly prone to inaccuracies due to immitigable problems, the largest issue of which is the uncertainty in the definition of settlement boundaries. Nonetheless, this analysis has, through extensive consideration, endeavoured to implement numerous safeguards which would align its results and conclusions as closely as possible to ones that would be arrived at were there perfect information available. The main approach adopted to partially remedy unavoidable sources of uncertainty was to consider in each situation multiple plausible scenarios to simulate how results would be altered with a change in a given variable. Where an input to the calculations do not drastically affect the results, this research sought to underline the assumptions that were made in carrying out those calculations and show that they were reasonable. Where the scope of possibilities would produce results that diverge from each other too much, a conservative approach was favoured, notably in that the analysis would rather underestimate, rather overestimate the growth in subscription numbers and proportions. To begin, the first matter below pertains to the figures for settlement population used to calculate telephone adoption rates.

Figure 16: Comparison of the calculated telephone adoption rates in select settlements, according to different methods of extrapolating settlement population

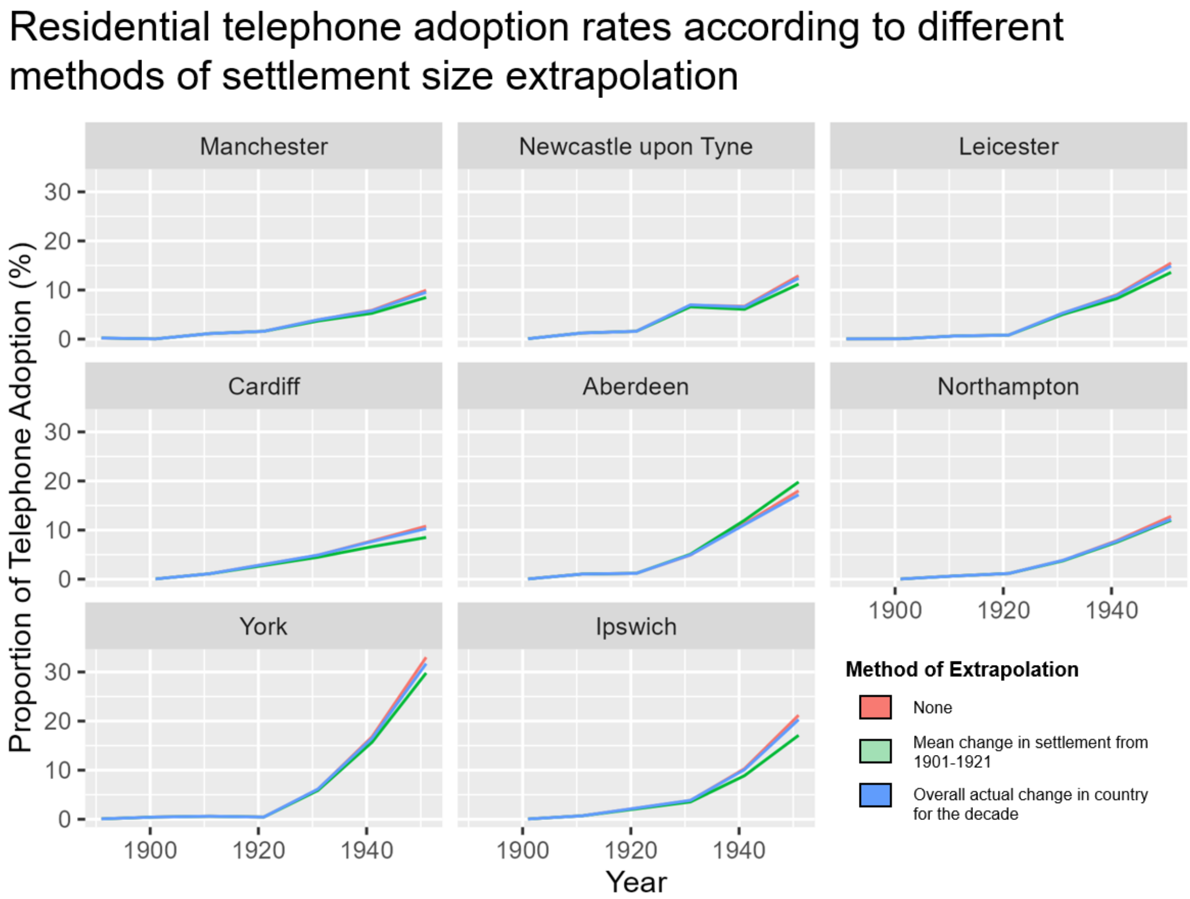


Figure 16 depicts comparisons between the residential telephone adoption rates in 8 chosen settlements as calculated using three different baseline figures for settlement population. The settlements were chosen to represent not only settlements across the spectrum of population sizes, from higher to lower, but also the different British countries. Because population sizes for the largest British settlements was only readily accessible for every decade between 1901 and 1921, this analysis opted to mathematically infer the settlement populations in all other relevant decades. As seen, in the majority of instances, the method of no extrapolation yields the highest estimates for adoption rates, albeit being minimally different from the other estimates. It also seems that estimates of telephone adoption proportions are more variable across the three scenarios in smaller settlements like Ipswich and York as compared to larger settlements like Manchester or Newcastle upon Tyne. This first method takes as the denominator for calculations of years 1881-1891 the settlement population as of 1901, and for all years following 1911 the population as of 1921. Given that this assumption of no population growth beyond 1911 is highly unlikely, it is safe to say that the results overestimate adoption. The second method assumes that population change in every settlement in every decade is equal to the mean decadal change for that settlement in 1901-1921, while the third projects onto individual settlements the overall population change of the UK country they belong to. The former most often yields an estimate of telephone uptake lower than that derived from the latter, indicating that decadal population growth in most settlements between 1901 and 1921 outstripped the overall growth of their respective countries in the period after 1921. The exception to this is Aberdeen in Scotland, which reflects an extrapolation of the population loss, rather than growth, that faced most major Scottish settlements between 1901 and 1921. Overall, the method of population size extrapolation does cause the numbers calculated to vary slightly but did not have an impact so large as to alter the trajectory of the results altogether, allaying fears that the uncertainty in settlement population figures would seriously impact the accuracy of this analysis.

The second major set of limitations relates to attributing geography to subscriber records in settlements that had constantly changing telephone exchange geographies. Prime examples of these are Stoke-on-Trent, which until 1931 was subsumed under the directory heading for *'Potteries Area'* or similar, and Greenock, which would sometimes be subsumed under Glasgow and at other times under a collection of a much wider area named *'Scotland West'*. In such cases where there was no affirmative way to deduce the total population size of the areas being served by the settlements involved – and thus no way to calculate a multiplier with which to moderate the subscriber counts for that settlement – a multiplier of 0.1x was applied to the number of subscribers ascribed to the aforementioned settlements. This was seen as a safer alternative to the possible gross overestimation the figures by including the tallies of subscribers in large tracts of neighbouring areas, rural or urban, to just one settlement. In fact, when this scenario was trialled, the total number of telephone subscriptions at times exceeded the entire populations of the settlements, which therefore invalidated this approach completely.

3.5 Conclusion

In sum, this chapter has illustrated one use case of the newly digitized dataset of telephone directory records. Through a process of attribution of high-level urban geography to the records that is in theory straightforward, yet in application was riddled with many necessary considerations to make, this chapter demonstrates how the dataset enables a spatiotemporal overview of how fixed-line telephony uptake grew across Britain in the 19th and 20th centuries. The analysis focused on adoption of telephones within the 53 top-ranked urban settlements as of 1901 and presented a supplementary conception of urban rank order based on the number of telephones registered in each town or city. When this analysis was repeated at the spatially more aggregated level of modern administrative regions, the results affirmed that strong regional differences emerge in the patterns of telephone adoption, notably the North-South divide among English regions with the North seeming to be more conducive to commercial adopters of the telephone, and the South to residential subscribers. With painstaking effort invested into anticipating and accounting for the limitations of the methods and data used, this analysis furthers understandings of the innovation diffusion of telephones in Britain through quantification. Additionally, this chapter also illustrates how rates of telephone adoption varied across the first 70 years of its service in Britain and were affected by the demographic and urban landscapes of the places in which they proliferated.

4 Further Exploratory Analysis

This fourth chapter presents the results of exploratory analyses that have been undertaken; it represents an initial foray into the possibilities that the newly digitised dataset brings for a retrospective analysis of several topics related to telephone ownership. Extended in different directions, this chapter explored an array of topics: in parts, it tries to draw on and build upon research that has already been done, on population data that are similarly structured to the telephone directory records and from a similarly historical context. Notably, this chapter looks to the historical Census data that have been digitised in the recent decade (Schurer and Higgs, 2023), as well as research that has made use of these data (e.g. Lan and Longley, 2021), as ways to enrich interpretations of the telephone subscriber data. Meanwhile, in other parts, this chapter considers how the dataset created could be analysed in its own right, for instance examining the feasibility of linking telephone subscriber records through time, in an attempt to emulate what Van Dijk *et al.* (2021) had done in a more modern context.

What ties together these seemingly disparate analyses is the common goal of evaluating the potential usefulness of the telephone subscriber data for future research. Alongside the discussion of outcomes of exploratory analyses that were done, this chapter also details the methodological considerations that had to be made and, as before, reflections on the assumptions that were made and how they implicate on the conclusions that could be made from the data. Where the preceding chapter aimed to give a bird's eye view perspective of the broad differences in telephone adoption across regions and settlements in historical Britain, this chapter is more focused on making inferences at the smaller scale of settlements and about the variations that exist therein. Broadly speaking, this chapter is divided into two parts, the first of which looks at telephone subscribers at the level of individual records, before the scope of analysis widens in the second part to an exploration of subscription patterns at an intra-urban scale. Ultimately, this chapter is intended to further demonstrate potential applications of the digitised telephone directories dataset that may serve as inspiration for future research in related disciplines.

4.1 Individual-scale Analysis

Having previously in this thesis noted granularity as a key advantage of the telephone directories dataset, this section first explores how its records can be analysed at an individual scale, drawing on research that has been done on datasets of a similar format, in particular the historical

Censuses. It begins with a comparison of the telephone subscriber and Census datasets, noting how differences in their original production, as well as later digitisation, ascribe them with different limitations. It then proceeds to test the feasibility of carrying out one-to-one matching on telephone subscriber records, firstly with contemporaneous census records to establish demographic provenance, and secondly with telephone subscriber records in the following decade to examine the extent of temporal continuity in the dataset. The section finally concludes with a reflection on the results, as well as what they possibly signal for future research.

4.1.1 Census Records as a Point of Reference

The database of telephone subscriber records, without any further modifications, already significantly widens the scope of opportunities for quantitative social science analyses in historical perspective. This is because little information is at all available at the scale of individuals and at the national scale for the larger part of the 20th Century: perhaps the sole other source of such information, the national Census of Britain, is bound by confidentiality guidelines which restricts disclosure of granular Census data until 100 years after it has taken place (Office for National Statistics, 2011). Consequently, this implies that, while these records are largely accessible for the 19th Century – during most of which the telephone had not been commercially released –, at the time of writing, the latest granular Census records that could theoretically have been released during the undertaking of this research would be those from the Census of 1921. As it then stood, the relevant Census records that were available through the Integrated Census Microdata (I-CeM) project were those for England and Wales from 1881 to 1911 and for Scotland from 1881 to 1901 (Schurer and Higgs, 2023).

Precisely due to this scarcity of available historical population data for Britain, this research argues that the digitised census records serve as a fundamental point of reference for how the telephone subscriber records could be used, helped only by the similarities that exist in the formats in which these datasets were originally captured. For one, they are recorded at the scale of individual households, which grants the user of these data the ability to conduct analyses at a more disaggregated level than the spatial units commonly used in quantitative historical analyses, such as countries, regions or cities. The telephone directory data are likewise recorded at the individual subscriber level, thereby also allowing for more disaggregated analysis, albeit of a smaller sample of historical populations, that is, the consumer bases of the telephone across the country. Secondly, like the census records, the historical telephone directories were also published at regular temporal intervals, with each edition meant to be an update of the records contained in the preceding edition.

In this way, both the census and telephone subscriber records offer users the potential to investigate change and continuity of enumerated persons within the same geographical boundaries, again at the individual, rather than spatially aggregated, scale. However, key differences that emerge between the two datasets bear practical implications on how analysis can be done, and conclusions reached, using the telephone subscriber data. These differences – stemming mainly from the processes of original data capture, and of how subsequent digitisation of the hardcopy data has been attempted on each dataset – will be discussed through a series of examples of research that has been conducted using data from the Censuses of England, Wales and Scotland.

Firstly, subjectivity in the original process of data capture is inherent to the Census data but which is not so much a problem for the telephone directories dataset. Studies which utilised data from national Censuses in Britain have sought to understand inconsistencies in the data capture process owing to various social factors. For example, McGeevor (2014) discussed the intentional exclusion of women's work in the Censuses by male enumerators and male heads-of-household, which might result in systematic geographical differences in how women's occupations are listed, depending on the sex of the responsible enumerator and/or head-of-household. This concern seems to be echoed by You (2020), who highlights the understatement of a woman's economic contribution when she is listed merely as 'butcher's wife', despite being equally as involved in commercial activity. Logistically, whenever historical Censuses were conducted, thousands of enumerators were simultaneously tasked with collecting data for different subsets of the British population and, while questions they asked the respondents were formulaic, the responses they received and ways of recording this information were not standardised; these inconsistencies were so pervasive that, as an example, there were more than 300 ways of expressing the occupation 'Blacksmith' (Schurer and Higgs, 2023). Altogether, the issues of bias and lack of standardisation that existed in data capture for the Census do not constitute worrisome problems for the telephone directory dataset, as the latter was both much more coordinated and smaller in scale in its data capture and dissemination (through the few telephone companies that operated), while also seeking to capture far less information that lent itself to subjectivity in description.

Secondly, the contrasting methods of digitisation of source data between the Census and the telephone directory archives have given rise to different limitations in the final data products created. While the use of Census data, owing to its reliance on manual transcription, has been prone more to human errors, the telephone directory data, having undergone a mostly computerised process of data capture, are bound more by limitations of the technology employed. Using Census data, researchers, in order to enable efficient analysis, had to manually digitise the parts that were relevant to their interests, whether academic, commercial or personal. The painstaking nature of this preparatory work meant that, typically, the range of records to be digitised had to be restricted, either/both by taking a representative sample from the population and/or confining the geographical extent and period of interest. For instance, McGeevor (2014), in her inquiry into how women's occupations were recorded in the Census, limited the area of interest to one English county, Hertfordshire, while also taking a 19.3% sample of records from a related county trade directory to digitise as cross-reference material. The completion of the I-CeM project in 2017 has tremendously reduced the amount of manual work required prior to using the Census data, but researchers' particular interests may still demand considerable manual transcription. For example, Williams *et al.* (2020), in their study of institutionalised poverty in historical Britain, also created a dataset from 10% subsamples of workhouse populations from every Census between 1851 and 1911, so as to enable an analysis of temporal trends, while keeping the workload of notating and categorising the data manageable.

Altogether, the Census data, as illustrated by the examples above, has often required users to put in numerous hours to transform the source data into something that fits their intended use-case. The potential for human errors in this transcription abounds, but nevertheless, quality checks between the digitised and source material can be feasibly conducted, either with sufficient resources designated in the project (that is, researchers' time) or if the scope of digitisation is limited. In contrast, the approach that this research adopts for the telephone directories dataset, in relying on the OCR technology as the first step of text data capture, is admittedly more prone to computer errors that are harder to both detect and resolve. While, as with the Census, source material is readily available for quality checks to be done on the outputs, such an endeavour was deemed as too time-consuming for one researcher to undertake, given how the volume of telephone directories grew exponentially with time, and especially given how this thesis sought to maximise the temporal coverage of the output dataset within a limited time span. Without a feasible and time-efficient way to conduct manual checks on the digitised records, this research acknowledges that there is considerable uncertainty in how accurate information contained in the resultant dataset is. This is a significant limitation of the telephone directory data as they stand, in comparison to Census data, and will be factored into the conclusions that are reached in the exploratory analyses in this chapter.

4.1.2 Record Linkage to Census Records

As a first attempt to extend the functionality of the digitised subscriber records, this research conducted trials to contemporaneous and similarly digitised Census records, with the trial analysis undertaken focusing on London, Manchester and Glasgow – the three largest settlements by population in 1901 in Britain – through the context of which initial findings will be demonstrated. Telephone subscriber records from these settlements were matched against records from the national Censuses of England, Wales and Scotland, which provided various aspects of demographic and socioeconomic information of the British population such as measures of family size, living conditions and occupations they held. Taken together with information on telephone subscribers, these data enable a more in-depth understanding of the characteristics of people who were more likely to adopt the telephone in its first decades, more than plainly sheer subscription numbers, and also provides some measure of telephone adoption rates at geographic scales smaller than entire settlements. Such linkage at the level of individual subscribers or households (as telephones tended to be rented) can help mitigate against the problem of ecological fallacy and thus further our understanding of the heterogeneity of neighbourhood structures in Britain's past.

Sets of records were linked using two pieces of identifying information: their last name and their street addresses. In the case of matching records of London, a third field of data was used to narrow the pool of potential matches – the Registration District of the geocoded addresses of subscribers – and this was deemed necessary not just to increase the likely accuracy of matching but also to optimise use of computing power. As with the process of geocoding records earlier described, the basis of record matching was fuzzy string matching. However, this record matching importantly differs from the fuzzy string matching used in the earlier geocoding process in that it matches on two pieces of information – subscriber surname and address – rather than solely the latter. In the earlier process of geocoding, the chosen match is the address from the OS AddressBase reference data that is the lowest string distance away from the telephone directory address. In contrast, for the record matching in this section, the chosen match is the Census record is which has which is the smallest string distance away from the telephone subscriber record, when the inter-string distances for both the surname and address fields are totalled up. This is expanded on below.

In the case of surname matching, the 'subscriber name' field of each record was in some cases further split into two values if there was a comma in that field. The most common scenario in which this happened was when a subscriber's occupation was listed alongside (usually following) their initials and surname, for instance 'E. CLARKSON, ELECTRICIAN', which would for this step be split into the distinct values 'E. CLARKSON' and 'ELECTRICIAN'. Both values would be standardised into the lower case, then each is matched to all surnames in the pool of Census records that serve as potential matches. String distances would then be calculated for pairings between these two strings and each distinct surname present in the Census records, with it being more likely in this example that the string 'E. CLARKSON' returns generally lower string distances than 'ELECTRICIAN', since the former contains a person's surname while the latter an occupation instead. Another, albeit rarer, scenario in which this further splitting of the 'subscriber name' field proved useful was when unwanted commas had been captured by the OCR software, either correctly (as commas were commonly used separators of information) or mistakenly, and had not been removed in earlier steps as the 'name' field was not yet relevant to any analyses.

In the case of address matching, two information fields from each telephone subscriber record are also used as potential candidates with which to match onto Census records, namely the original address string as captured by the OCR software, and the address string from the OS AddressBase that had been appended to the subscriber record as the best match during the process of geocoding. For this reason, the margin of error is increased because the algorithm now has to consider the inter-string distances between the values for two fields for every pair of possibly matching records and a perfect match becomes all the less likely. This sub-section will thus explain how the procedure was developed and how uncertainty in its results were quantified.

The solution that this research eventually arrived at was a general-purpose code script for record linkage. Affording flexibility in its use by way of input data and parameters, most crucially leniency on the threshold for string matching, this script was found to be necessary for two reasons. Firstly, the digitised telephone directories data often comprised more than one version of the same piece of information, and an easily customisable algorithm would facilitate quick comparisons of the efficacy of using these different versions for matching. Secondly, this feature also efficiently enables the continuous trial-and-error testing of different string distance thresholds, so as to understand their impacts on matching rates. Since the entire processing pipeline of the telephone directory scans entail a non-negligible amount of uncertainty in its accuracy, being able to make tweaks to the process with the simple change in one line of code is paramount in allowing this project to evaluate the results of linkage under different scenarios and thereby better gauge the uncertainty involved.

The generalised function also allows for the record linking process to be broken up into stages, with each stage using surname and/or address variables for linkage in varying combinations. Specifically, the street names of subscriber addresses exist in both their raw, unmodified forms as the outputs of text capture from the scanned images and in their 'perfect' matches derived from the OS AddressBase+ that were joined onto the subscriber information during geocoding. Similarly, the last names of subscribers also exist as both the direct outputs of the processing pipeline and another in which they have been matched with a comprehensive list of all names existing in at least one of the national Censuses between 1881 and 1901 as was compiled by Lan *et al.* (2021). Table 14 below describes the parameters that the function takes and how each pertains to the outcome of record linkage.

Table 14: Description of parameters of the generalised function for record linkage

Function Argument	Description
`tld_A;` `tld_Z`	Input data frames containing telephone subscriber information for the earlier decade (`A`) and later decade (`Z`)
`snmCol_A;` `snmCol_Z`	Variable names within the respective data frames that contain residential subscribers' surnames.
`adrCol_A;` `adrCol_Z`	Variable names within the respective data frames that contain the street names of subscribers' addresses.
`snm_Thres;` `adr_Thres`	Numeric value representing the maximum acceptable string distance to apply in the matching of surnames and street names of addresses between telephone directories. The default value is 3.
`lnk_Stage`	Numeric value that indicates the stage of linking records, i.e. combination of fields of surnames and street names used. Also used to create matching flags for the matching process to be retraceable.

Of particular note are two variables which correspond to thresholds for the fuzzy string matching of names and addresses. There is a fine balance to be struck in determining these thresholds: while a near-zero threshold would maximise confidence that the results are making only non-erroneous matches, it must include built-in tolerance for common inter-string differences that arise during data capture. It was found necessary to allow for separate thresholds for the matching of surnames and addresses because the problem of abbreviations pertains only to the latter and thereby demands the possibility of choosing a more lenient threshold. Moreover, captured last names tended to comprise fewer characters than street names (as did subscriber names than full addresses), meaning that a one-unit increase in the inter-string distance threshold would affect matching of last names more and more likely to transform a word into something completely different. The thresholds ultimately chosen were a string distance of 2 and 3 for surnames and street names, respectively. The precision of record linkage, admittedly, would also rest in part upon the accuracy of the prior geocoding of subscriber information.

4.1.3 Record Linkage Outcomes

Table 15 shows the results of record linkage between the digitised telephone records and those of the contemporaneous Census. For each year and for each settlement, the total number of records eligible for matching refers to records that both belonged to the chronologically latest directory of the year for the settlement, where multiple directories existed, and that were deemed to be residential in nature by the processing pipeline. Because this research had only obtained access to the granular Census records of England and Wales from 1881 through to 1911, and of Scotland from 1881 through to 1901, matching of records was done for the three largest cities for the periods for which a substantial number of telephone records existed for each of these settlements. The records for Manchester from 1891 were not included, as they numbered fewer than 100. Additionally, the telephone directory records of Glasgow in 1911 were matched to the Census records of 1901 as the 1911 Census records for Scotland were not accessible.

Table 15: Proportion of records matched at different accuracies between telephone directories and contemporaneous Census records

Accuracy of Linkage	London				Manchester		Glasgow	
	1881	1891	1901	1911	1901	1911	1901	1911
Number of eligible records	547	4,610	7,577	55,552	547	4,965	447	13,209
1: Perfect matches for both surname and street name of address	0.01	<0.01	0.17	0.13	0.05	<0.01	0.04	0.02
2: Perfect surname match; match of street name within string distance of 3	0.08	0.15	0.14	0.14	0.14	0.11	0.21	0.13
3: Perfect match of street name; surname match within string distance of 2	0.19	0.03	0.24	0.30	0.45	0.22	0.11	0.10
Fallback: Imperfect matches for both surname and street name, within string distance of 5	0.31	0.37	0.28	0.24	0.26	0.45	0.43	0.43
Overall proportion of records matched *(excluding 'Fallback')*	**0.30**	**0.19**	**0.55**	**0.58**	**0.65**	**0.35**	**0.37**	**0.24**

As seen from Table 15, the overall match rates between telephone subscriber and Census records were not ideal. Excluding the 'Fallback' category which had a much more lenient matching threshold of string distance 5, the match rates for most years and in most of the settlements were under 50%. There is a marked difference in temporal trends in Glasgow and Manchester as opposed to London, despite the fact that, in their respective first years of appearance in the directories, the number of records eligible for matching in all three settlements are similar. For London, match rates on the whole increasing from 30% in 1881 to 58% in 1911 (dipping to 19% in 1891), while for Manchester and Glasgow they fall by 30% and 13%, respectively, between 1901 and 1911. By 1911, the last year for which Census records are available, the rates of perfect matches between telephone directory and contemporaneous Census records in terms of subscriber surnames and street names of their registered addresses are also markedly higher in London at 13% as compared to less than 1% and 2% for Manchester and Glasgow, respectively.

With data across four decades available London, as opposed to two for the other settlements, it becomes easier to spot that 1891 appears as a clear outlier. When the general trend for match rates shows an otherwise upward trend with time between 1881 and 1911, it instead falls 11% between 1881 and 1891 and, upon closer inspection, this dip corresponds to the change in telephone subscriber records that were matched perfectly on street names, but imperfectly – within a string distance of 2 – on surnames (category '3' in Table 15). Figures on the latter measure change similarly, dropping from 19% in 1881 to just 3% in 1891, before again increasing to 24% and 30% in 1901 and 1911, respectively. A likely reason for this decreased match rate is that addresses in the London directory were formatted particularly differently from those in the Censuses and in the London directories of other decades. For one, it seems that the OCR algorithm struggled to accurately read most postal district markers in this edition – which had been placed at the end of addresses, but in a different font and smaller font size. Secondly, addresses from 1891 were also formatted very differently from those from other directory editions, notably with the inclusion of non-standard information, namely of the wider district or area in which a subscriber lived, in addition to the standard components of an address that are street name, street number and postal district. Upon examining the digitised outputs, this research even noticed that a non-negligible proportion of addresses contained two or more street names, for instance the address entry '10, London-street, Fenchurch-street', which indubitably added to the challenge of accurate address matching. A combination of these factors was thus likely to have contributed to the drastically lowered match rate in London for 1891.

Comparing across the three settlements, something that the diverging temporal trends likely point to is the difference in how well harmonised their telephone exchange geographies are with their respective civil administrative geographies. In London, as shown earlier in Figure 6, a map of the collective service area as defined by the 1914 edition of the directories lined up very neatly when overlaid on a map of Registration Districts – an official type of civil boundaries used for enumeration through the later part of the 19th century – in London from 1881, despite a significant time difference in production of these materials of over 30 years. In turn, these Registration Districts corresponded spatially to the Registration County of London that had been digitised by the Great Britain Historical GIS Project (2017).

Meanwhile, the same certainty in the equivalence of telephone network and civil geographies cannot be had for Manchester and Glasgow, which in comparison were much smaller spatial extents and thereby were given less detailed documentation of their service areas in the directories. For Manchester, the pool of possible Census records to match on was initially taken from the twin Registration Districts of Manchester and Salford, which corresponded to sections of the telephone directories which, until 1911, were entitled 'Manchester and Salford'. However, when this returned an abysmal rate of matches of 36% in 1911, the possible pool of matches was expanded to include all records in the Registration County of Lancashire; in addition to the two aforementioned, Lancashire also contained Registration Districts including Ashton Under Lyne, Oldham and Rochdale, many of which would be featured in the list of names of regions covered by the 'Manchester' section in directory editions beginning 1921. A similar problem existed for choosing the pool of Census records to match on for Glasgow, where a telephone service area that was not initially clearly defined beginning 1921 began to incorporate telephone subscribers in neighbouring areas, notably the settlement of Paisley. However, unlike Manchester, Scottish Parishes were used as the main civil geographical units used for analysis of Glasgow as the historic Registration Districts of Scotland had not been available in a digitised format as they are for England and Wales. The range of Census records that were then deemed as possible matches were those in historic Parishes that had some overlap with the boundaries of modern-day Glasgow City. Ultimately, this inability to precisely pinpoint the extent of Census records that should be availed for matching for Manchester and Glasgow, unlike for London, would mean that the match rates for these settlements are likely overestimations, as some subscriber records might be matched to Census records outside areas then-serviced by the telephone.

One final difference to note from Table 15 is that, between 1901 and 1911, the decrease in overall matching rates for both settlements is attributable to different factors. In Manchester it is mainly due to a reduction in records matched perfectly on their street names but not surnames (45% to 22%), while in Glasgow it is due to fewer records being matched perfectly on their surnames but not street names (21% to 13%). It is less clear why this decrease was so pronounced in Manchester but a possible cause, based on skimming the data, is that 1911 Census records from Lancashire (which contains Manchester) tended to also include additional information in their address fields, such as city or region name, more so than the records of London and Glasgow. This different convention might have thus affected string matching accuracy by introducing information that was present only in the Census dataset but not the other. Conversely, because the matching of 1911 telephone subscriber records of Glasgow was done against the Census records of 1901, the reduced match rates are likely a result of both an influx of new immigrants and outflow of migrants. This appears to be in line with the net population decrease across nearly every major Scottish settlements between 1901 and 1921, as indicated in the tables of the Census Reports of Scotland of 1901, 1911 and 1921 (Great Britain Historical GIS Project, 2017). With these observations, something that warrants further investigation is whether the omission of the initials of subscribers' first names and street number, both of which are not always reliably recorded or captured in the directories, impacts the accuracy of matching significantly.

4.1.4 Record Linkage across Time

Having trialled the establishment of linkage between telephone directory and Census records, this research sought, using an adaptation of the method outlined in Section 4.1.2, to also link up telephone directory records referring to the same subscriber in different time periods. Linkage of subscriber records with those from the Censuses were conducted for the period between 1880 and 1911, that is, between when the first telephone directory in the research archives was published and the last decade in which granular Census records had been released. As a contemporaneous source of population data that was, similarly to the telephone directory dataset, recorded at the level of individual households and had national coverage, the Census held potential to shed light on the geodemographic profiles of urban telephone subscribers. However, beginning 1921 (keeping in line with the previous focus on Census-contemporaneous years), such granular Census data were not yet accessible, at the time of undertaking this research, to serve as a 'golden standard' that indicated the full extent of the population at any given time in history. Other kinds of data that could help to contextualise telephone subscription growth patterns still existed – such as aggregated data on facets of urbanity, population and the economy from the decadal national Census reports –, but of the disaggregated data, all that was available was data in the newly digitised telephone directory dataset.

The rationale behind this attempt to establish temporal links amongst telephone subscriber records was that doing so will facilitate an examination of the continuity, or lack thereof, of the populations represented in the dataset. This, in turn, draws upon recent research that deal with more modern sources of population information not originating from official government sources, that become increasingly pervasive near the end of the 20th Century. A prominent example of this – the Linked Consumer Register conceptualised and executed by Lansley *et al.* (2019) – shows that it is possible to give authoritative estimates on population statistics and measure change thereof in contemporary Britain by bringing together annual, individual-level electoral roll datasets from 1997 to 2017 with other sources of non-officially collected data, such as those derived from land registries and consumer files from the private sector. Van Dijk *et al.* (2021) then leveraged on this dataset to better understand the and extent nature of residential moves in Britain, by matching records of individuals at different times, using forename-surname pairings in the data; an individual is considered to have moved if different residential addresses are appended to their records in the two time periods, and considered to have stayed put if their address remains constant over time.

The example cited above served as inspiration for this research to imitate the method of record linkage used, so as to understand the residential stability of telephone subscribers. This was envisioned to add to the pool of available information about each subscriber, that is, whether they were existing subscribers (having already been present in an earlier edition) or new subscribers to the current edition. At a more aggregate scale, this was further envisioned to enable the residential telephone subscription patterns in different geographical areas to be quantified through comparisons of the proportions of old and new telephone subscribers, although the exploratory analysis in this chapter did not end up going as far.

As a proof of concept, the scope of this linkage was limited to records within London between telephone subscriber records ten years apart and occurred in the census years between 1901 and 1941. As with before, London was chosen because of its dominant share of total subscription volumes within Britain: even as this dominance waned over time, as late as 1926, London alone still accounted for 28.9% of all telephone ownership in Britain (Kingsbury, 1915). In terms of years of records selected, this analysis remained aligned with the previous analyses in focusing on telephone directory records from years that were contemporaneous to the historical Censuses, with the view that this would maximise temporal coverage of the data used. This was also done to optimise time-efficiency, since the records of London, which are considerable in number in each year, had already been digitised and geocoded for the Census-contemporaneous years, and thus no records had to be additionally digitised solely for the purpose of this analysis. In hindsight, however, it might have been more sensible to conduct analysis on a smaller scale, with records from directly adjacent years rather than those 10 years apart from each other; after all, the samples of telephone-subscribing populations in a locality were likely to be more similar, with fewer removals and additions, if compared across a 1-year period than across a 10-year period.

Nonetheless, the results of record matching, as done across 10-year periods, is presented hereafter alongside a reflection on the outcome attained.

Table 16: Proportion of records matched at different accuracies between telephone directories of successive decades.

Accuracy of Linkage	**London**			
	1901	1911	1921	1931
Number of eligible records	7,577	55,552	64,891	381,916*
1: Perfect matches for both surname and street name of address	0.09	0.07	0.02	<0.01
2: Perfect surname match; match of street name within string distance of 3	0.05	0.05	0.02	<0.01
3: Perfect match of street name; surname match within string distance of 3	0.15	0.16	0.14	0.15
4: Imperfect matches for both surname and street name, within string distance of 3	0.21	0.24	0.30	0.26
Proportion of all yearly records matched	**0.50**	**0.52**	**0.48**	**0.41**

**for the year 1931, a random sample the size of the number of eligible records in 1921 (64,891) was chosen to test this analysis.*

Table 16 displays the proportion of subscriber records in the latest London directory of every decade which had found a plausible record match in the directory published ten years after. The accuracy of record linkage is then described as falling within one of four categories, ranging from a perfect match of both surname and street name ('1') to a non-perfect match of both variables ('4'), but that still fall within the pre-defined matching threshold of string distance 3. This more lenient threshold was chosen, as compared to the threshold of 2 earlier used for matching telephone subscriber records to Census records, to accommodate for the fact that the telephone directory records are considerably more error-prone than those in the Census: digitisation of the latter involved less automation and much more manual input and monitoring in the process and so matching between two sets of telephone subscriber records would involve a greater margin of error. As anticipated, the match rates here are on the whole lower than those in Table 15 despite the leniency built in. From 1901 through to 1921, roughly 50% of all records in the latest directory of London subscribers are matched to a record in a directory edition ten years later, before falling to 41% in 1931. The number of perfect matches shows an even more marked decline over time, from 9% in 1901 to 2% in 1921 and to under 1% in 1931. In addition, as a proportion of all records matched, the proportion of records that are matched under category '4', that is, imperfectly (with a string distance between 1 and 3, inclusive) both by their surnames and their address street names, constantly trends upwards; over half of the records from 1921 and 1931 that were matched fall within this category.

All this evidence above seems to hint at the fact that the largely automated process of data capture of the telephone directory information, while sufficing in its accuracy for earlier applications, may not yet be suitable for use in record matching across time. Previous analyses in the third chapter hinged upon two factors: firstly that the number of subscribers in each telephone directory could be tabulated, and secondly that these records could be attributed to the settlement within which their corresponding telephone subscribers resided. Putting in place manual checks to safeguard the accuracy of attribution of the latter, while tedious, nonetheless remained feasible timewise, as it primarily involved ensuring that the range of pages corresponding to each urban settlement, in each relevant directory, had been recorded accurately. The comparatively more gargantuan task, to create the optimal conditions for accurate record matching, was in this case to check that each individual subscriber record, especially its name and address fields, had its text accurately captured and successfully passed through the processing pipeline described in Chapter 2. Given the sheer number of records for London alone, this time commitment was not one that this research could afford to make in order to further but one aspect of the exploratory analyses undertaken in this chapter.

This research had initially thought that the comparatively higher volumes of subscriptions for London in comparison to other settlements would have made potential record matching easier by way of widening the pool of potential matches for each subscriber record to be matched. Instead, the trials showed that record matching became less precise with the passing of time, which also resulted in the pool of potential matches being expanded. A closer examination of Table 17 also reveals that the decline in precision of record matching was more likely attributable to a reduced accuracy of surname matching, rather than of geocoding (or street name matching): the proportion of records in category '3' (matched perfectly on street name, but not on surnames) remains relatively constant around 15% in each year, whereas that of records in category '2' (matched perfectly on surnames, but not on street name) starts off already lower, at 5% in 1901, and falls to below 1% in 1931. This lends credibility to the proposition made earlier in this chapter, that the accuracy of capture and digitisation of subscriber surnames is more difficult to ensure: not only are these strings generally shorter (and thus impacted more greatly by every additional character that is erroneously captured than longer strings), subscriber surnames are also harder to identify from within the wider 'subscriber name' fields than street names are to identify from 'subscriber address' fields, not least because people's names are much more loosely bound by convention than street names. A central implication of this exploratory analysis on the temporal linkage of telephone subscriber records is that, if this idea is to be taken further by future research, it will probably have to either be reduced in its scope of analysis, and/or as a pre-requisite to the matching process involve greater time dedication to checking for the accuracy of data capture of the subscriber records, particularly in the field of subscriber names.

4.2 Intra-urban analysis

The previous section of this chapter has explored the feasibility of analyses that could be conducted on the scale of individual records of telephone subscribers, looking to extend their utility through attempts to link them one-to-one to records from contemporaneous national Census of England, Wales and Scotland, as well as to telephone records in a different time period. While the results showed some promise of these approaches, they simultaneously indicated that, in order for the analyses to attain a more ideal level of credibility, considerable refinement would be needed in the capture and digitisation of said subscriber data for the subsets of records relevant to the particular research interests that guide the analysis. Having reached these conclusions on the feasibility of working with the data at an individual scale, this section will proceed to explore how the telephone subscriber data can be analysed at the more aggregated scales of streets and urban districts to reveal intra-urban patterns of telephone subscription. As before, the analyses in this section will also draw upon the examples of the three largest British settlements as of 1901, that is, London, Glasgow and Manchester. To supplement an understanding of historical telephone uptake, this section will attempt to uncover not only how overall figures of telephone subscription trend, but also which kinds of people were more likely to adopt the telephone, and how these patterns differed within and between the three aforementioned settlements.

This section begins with a visualisation of the spatial disparities in telephone adoption through mapping location quotients of telephone subscription. This is done by aggregating the telephone subscriber records to the level of districts or parishes, and depicts the parts of the different settlements that are laggards or leaders in telephone adoption rates, relative to the average of the settlement. Another measure of inequality of access to the telephone is then presented, examining how evenly distributed subscribers on the telephone network of each settlement are across the streets of that settlement, and how these patterns change temporally. This chapter finally attempts to enrich understandings of telephone subscription profiles in each named settlement through the incorporation of a geodemographic element to analyses. By drawing on existing historical research on Census data from the same period, this analysis attributed each telephone subscriber records with the geodemographic profile of the street that they were registered to and subsequently examines how these compositions varied within and across the named settlements. The focus here, at large, is to chart as far as possible the changes over time or the persistence of trends of telephone usage that are reflected in the digitised telephone directories and other data which help to contextualise it.

4.2.1 Location Quotients of Telephone Adoption

This first analysis uses the concept of Location Quotients, borrowing a concept popular in adjacent disciplines such as Regional Studies, to quantify variations in a quantity between the smaller parts of, as well as in relation to the overall figures of, a larger spatial unit (Flegg *et al.*, 1995). Although the digitised dataset is indeed granular and allows for potentially similarly granular insights to be derived therefrom, the previous chapter exploring the feasibility of one-to-one linkage showed that, without significantly more manual work put into overcoming the shortfalls of the imperfect data capture process, the efficacy and accuracy of these efforts would be rather restricted. Producing location quotients allows for an alternative, admittedly less precise, way to quantify the differences in telephone adoption rates across both time and space. Put into context, location quotients here serve to compare the telephone adoption rates between urban geographical units for the settlements of London and Glasgow. Manchester was excluded from this analysis as, with the information available and at a scale comparable to the ones used for London and Glasgow, there were only two distinct spatial units (Registration Districts), which made a calculation of location quotients not meaningful in comparing variations therein.

First and foremost, residential telephone adoption rates for each geographical unit in each aforementioned settlement were calculated by dividing the total number of residential telephone subscribers by the estimated capacity for telephone adoption in each Registration District (London) or Scottish Parish (Glasgow), as measured by the number of households. The derivation of the first quantity is described hereafter. Following the digitisation of telephone directory records, an algorithm probabilistically assigns a label indicating whether a subscriber is residential commercial based on querying for string patterns that reveal this. This was an iterative process wherein more keywords were added to the list of queries as more kinds of commercial telephone subscribers were discovered in the dataset, to improve the accuracy of labelling. Commercial subscribers are disregarded and then, the remaining records are aggregated according to the civil geographical units they belong to, given that the earlier process of geocoding would have attributed each subscriber record to an address that, in turn, belonged to a Registration District or Parish. Another consideration in calculating the number of residential subscribers was the chronology of telephone directories published: in cases where multiple volumes of telephone directories were published in a single year about the same settlement, the one published last was taken to guard against double-counting of records.

In selecting the denominator for this calculation – the estimated capacity for telephone adoption –, this research sought to approximate as accurately as possible the maximum potential for residential telephone adoption in small areas and several options were considered. Of these, the most easily accessible data that were also available throughout the period of concern to this analysis was the residential population of each geographic area. This information had been used to calculate, in the previous chapter, the adoption rates of the 53 British settlements that had ranked within the Top 50 on at least one of four criteria in 1901. In that earlier case, raw population figures had been used as the basis of calculations because neither of the latter two options presented hereafter were available for all 53 settlements. Population counts by settlement for the Census years between 1901 and 1921, inclusive, had been collated for each of these 53 settlements within the Census Reports of 1921, making this information not easily accessible, but also the only option for that scenario with complete coverage of the pertinent settlements, which was necessary to enable a consistent comparison of adoption rates across the board.

With this analysis concentrated on a much smaller geographical area – London and Glasgow – than prior analyses, other options availed themselves and using the population figures no longer remained the sensible option: not only did telephones for the most part of this dataset's timespan remain financially out of the reach for the majority of the British population, it was also even less likely that individual households had use for multiple telephone lines. Telephones were usually adopted at the level of the household and so the second option considered was to use the number of people marked as 'head of household' in the digitised Census records. This option would better account for the fact that some members of the household would not feasibly become telephone adopters during its nascence such as children, live-in servants or housekeepers. However, it was also common well into late nineteenth-century England for more than one household to occupy the same apartment, flat or even room. Given that it was more deprived segments of the population that had to live under such arrangements, it seemed unwise to assume that each of these households living under one roof had the potential to adopt one telephone line each.

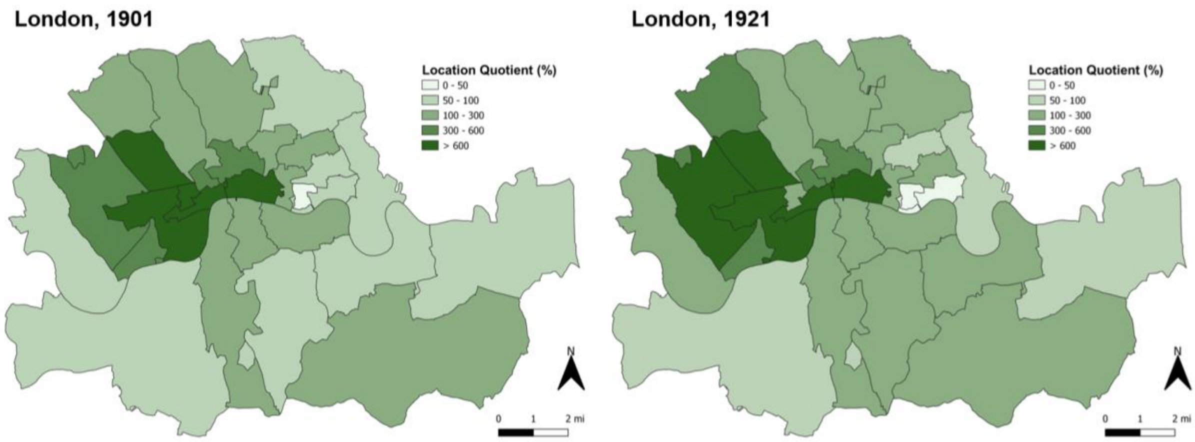
Furthermore, the availability of this 'head of household' data would be limited by the release schedule of the granular Census records, making comparisons with years beyond 1911 more difficult using this calculation method.

Ultimately, the data chosen was as the denominator was the number of inhabited buildings in each area, taken from population tables published in the decennial Census reports (Great Britain Historical GIS Project, 2017), seeing that the household was the most common unit for fixed-line telephone adoption. These figures were collated by this research for each Census year from 1881 to 1921, inclusive, by manually copying-and-pasting figures for each settlement and each year separately, thereafter, combining them. While this was more straightforward for London, whose constituent Registration Districts were usually conveniently listed in the same page for each year, piecing together the information for Glasgow was more tedious. The Scottish Parishes that formed the City of Glasgow were, for most years, listed under separate Burghs that were in turn subsumed under four separate Registration Counties: Dunbartonshire, Lanarkshire, Renfrewshire and Stirlingshire (ibid.). Going through this trouble to collate figures for two settlements over the 5-decade period was manageable, but doing so for all 53 Settlements would have been unfeasible timewise, and therefore why this analysis differentiates itself from the analysis in Chapter 3 in its choice of what constitutes the maximum potential for adoption of the telephone in their respective calculations.

Admittedly, this may overlook the fact that some inhabited buildings did not have real potential for telephone adoption, buildings that served as servants' quarters in particular. Nevertheless, it was determined that the occurrence of such buildings would be much lower than instances of multiple households living in the same building but not individually subscribing to telephone services, a strong indicator of this being the advent of payphones shared by entire apartment complexes at the beginning of the 20th century ('Dial B for Britain: The Story of the Landline', 2017). Consequently, this meant that the loose rule of one telephone per household applied even less uniformly and using the number of inhabited buildings as an approximant of telephone adoption potential was seen as a better alternative to using the number of heads of households, especially in a period where available information was scarce even at the spatially aggregated level.

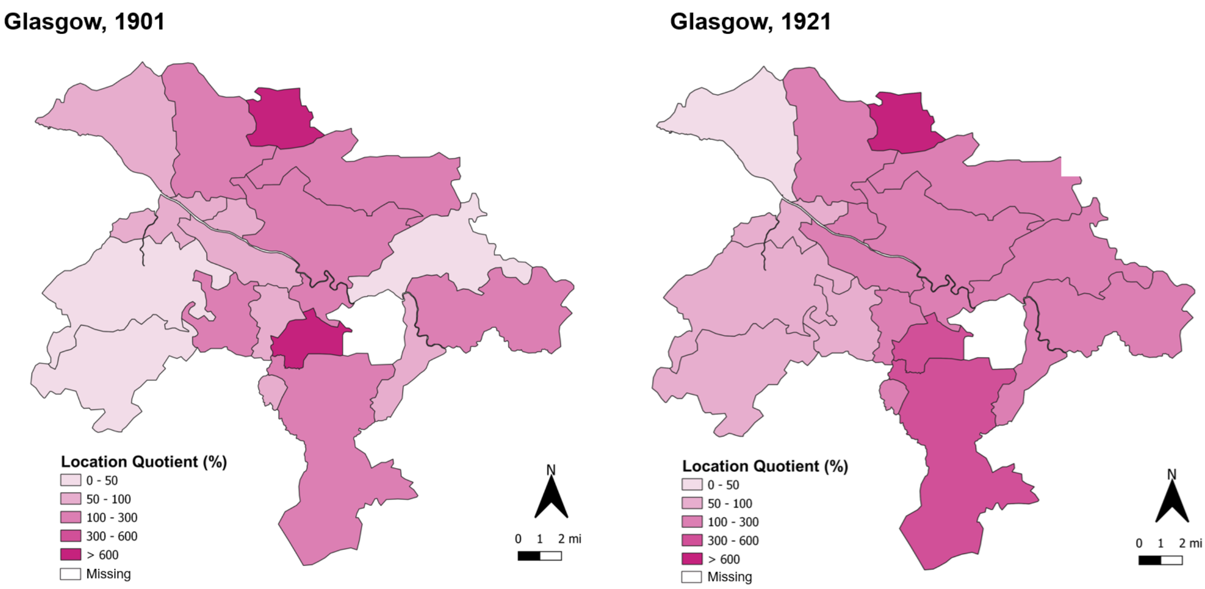
As a first indicator of the differentiation of telephone uptake in intra-urban contexts, Figure 17 and Figure 18 depict maps of London and Glasgow which compare the proportion of residential telephone adoption per constituent Registration District and Scottish Parish within each settlement, respectively, against one another. A location quotient value of 100% would mean that telephone adoption within a geographical unit in the city is on par with that of the city as a whole, while values exceeding 100% would mean that telephone uptake per unit population is higher than the average of the city. In this instance a map of Manchester was not included because it would have contained just two Registration Districts, Manchester and Salford, which would be too few to make for meaningful comparison using a settlement-wide location quotient. The two time periods with data depicted are 1901, chosen to represent the first year of inclusion in the directories for Manchester, Glasgow and many other settlements outside London and 1921, the first year of telephone subscription data that was captured after the introduction of automatic telephone exchanges in 1912 (Post Office Telecommunications, 1970) that prompted accelerated growth in subscription numbers across the country.

Figure 17: Maps of the location quotients of residential telephone adoption, comparing the Registration Districts of London between 1901 and 1921



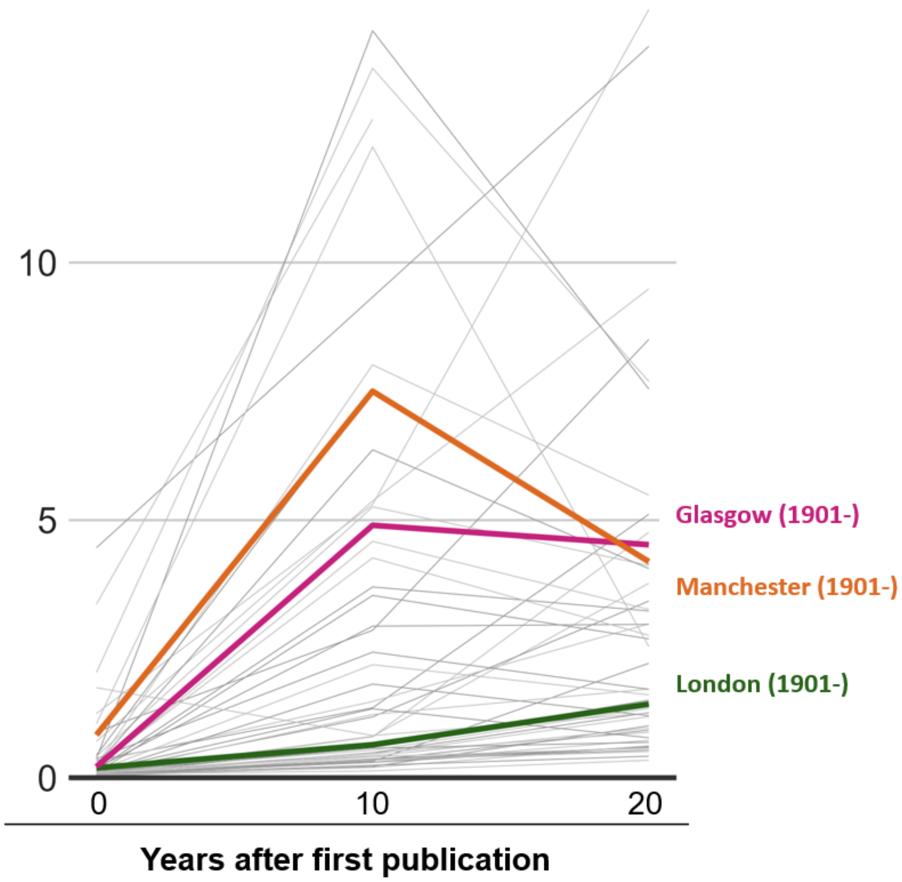
In London, trends in the variation of telephone subscription uptake across the Registration District seemed to hold roughly constant, even with a 20-year gap and the introduction of automatic telephone exchanges (Figure 17). The core of Registration Districts with the highest rates of telephone adoption relative to the others remains in the just North of the settlement centre, with districts such as City of London, Strand and Westminster remaining in the highest bracket of location quotients in excess of 600% across both time periods. Districts surrounding this core to the North, like Hampstead, and West, like Kensington and Chelsea, also mostly become among the most avid adopters with location quotients of over 300% by 1921. Conversely, the districts with the lowest relative telephone adoption rates are located further away from this core, with Fulham and Wandsworth to the West and Poplar and Woolwich to the extreme East. In Glasgow in 1901, the Parishes with the highest location quotients of residential telephone uptake are mostly positioned in a vertical line through the centre of the settlement (Figure 18), notably including the Parish of Glasgow, which consistently accounted for the largest share of subscribers in the settlement. By 1921, this pattern extends into a distinct East-West divide whereby all the Parishes with location quotients above 100% are in the settlement's centre or East while those with location quotients under 100% lie exclusively in the West, including the Parishes of Abbey, Renfrew and Neilston and excluding Cambuslang, to which no telephone subscribers were geocoded in either year. Together, a common theme that the maps collectively echo is that intra-urban patterns of inequality of residential telephone access persist considerably over time for both the largest settlements of England and Scotland.

Figure 18: Maps of the location quotients of residential telephone adoption, comparing the Scottish Parishes of Glasgow between 1901 and 1921



In place of a map, a comparison between Manchester and the former two settlements can be seen in Figure 19, where telephone adoption rates for the constituent geographical units (Registration Districts or Parishes) of the three settlements, and for them as a whole, have been adjusted for the year of their first inclusion in the directories. Having received telephone service coverage much later than London, Manchester and Glasgow nonetheless had faster increases in their settlement-wide adoption rates in their first 20 years of recorded telephone service, with both achieving an adoption rate of just under 5% by 1921 as opposed to London's equivalent figure 2.1% in 1901. The spike in the graph of Manchester after 10 years of recorded service was most probably caused by a redefinition of the Manchester Registration District in 1911, whereby its registered population in the Census tables jumped approximately sevenfold to just over 700,000. Together with the fact that Manchester in the telephone directories of 1901 serviced just the area of the two districts of Manchester and Salford, this redefinition is what likely caused the irregularity.

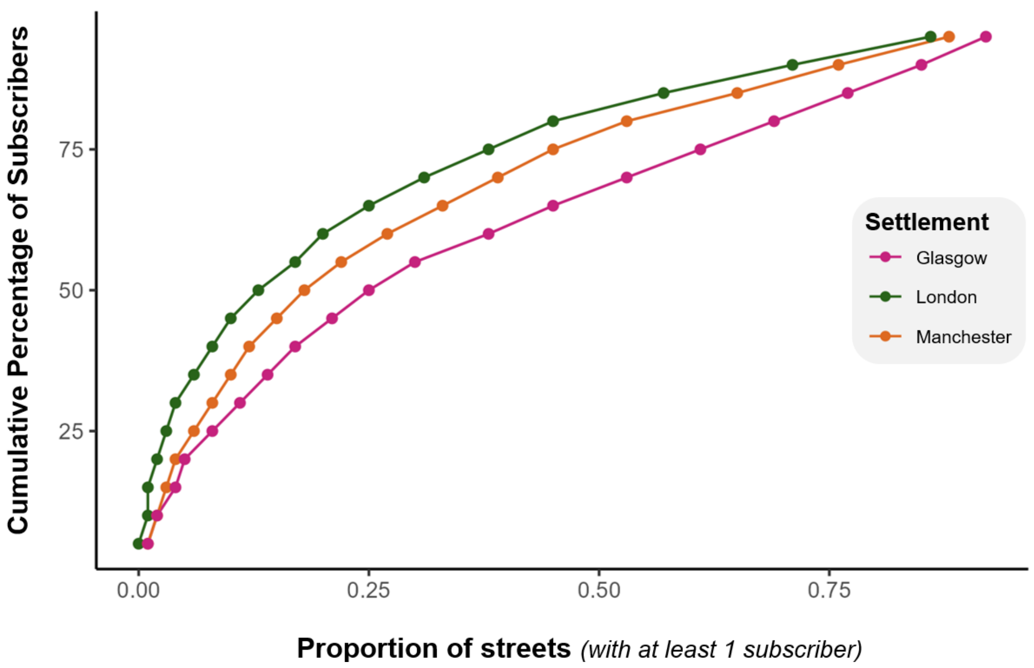
Figure 19: Comparison of telephone adoption rates in the three settlements, adjusted for when they were first included in the telephone directories



4.2.2 Concentration of Telephone Access within the Street Network

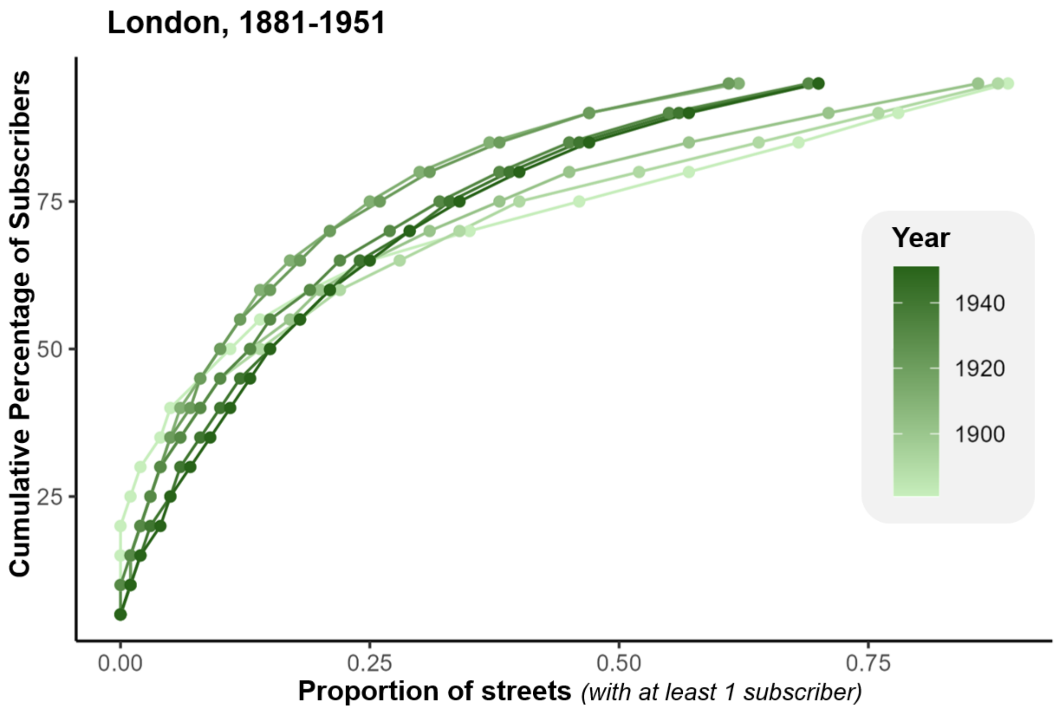
Another gauge of the differing patterns of the innovation diffusion of telephones within settlements is how geographically concentrated or dispersed telephone subscription patterns are. While the previous charts could show rough distributions in the geodemographic types of subscribers and the subscription volumes on different street, they could not more precisely summarise this information using a quantitative measure. Having attributed to every telephone subscriber in the three largest cities residency in the most proximal street segment, this analysis summarises the distribution of residential subscribers across the street network with a series of cumulative frequency percentage charts. Ordering the street segments that contain at least one telephone subscriber in descending order of their subscriber counts, Figure 20 shows that the graph for each of the three settlements begins with steeper gradients, indicating the presence of the few streets that are top-ranked in terms of number of resident telephone subscribers, before this tapers off with more numerous streets but which are less extensively serviced by telephone networks. In 1901, the first year of telephone service in Manchester and Glasgow, telephone subscriptions were more dispersed across each settlement with the top 18% and 25% of streets accounting for 50% of all residential subscriptions, respectively, whereas this figure is 13% for London. This difference becomes less pronounced when looking at the percentage of streets which account for 95% of all subscribers. Nonetheless, the relative order of dispersion among the settlements remaining the same, this figure being 86% for London, 88% for Manchester and 92% for Glasgow.

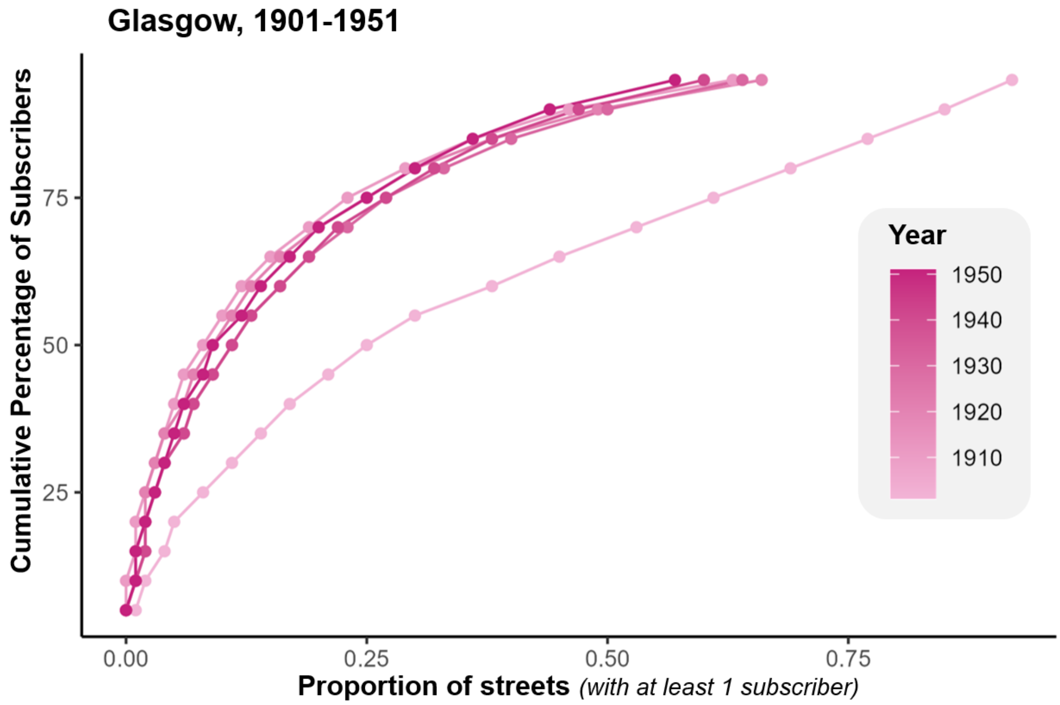
Figure 20: Cumulative frequency (percentage) chart of the distribution of telephone subscription across street segments of each settlement in 1901

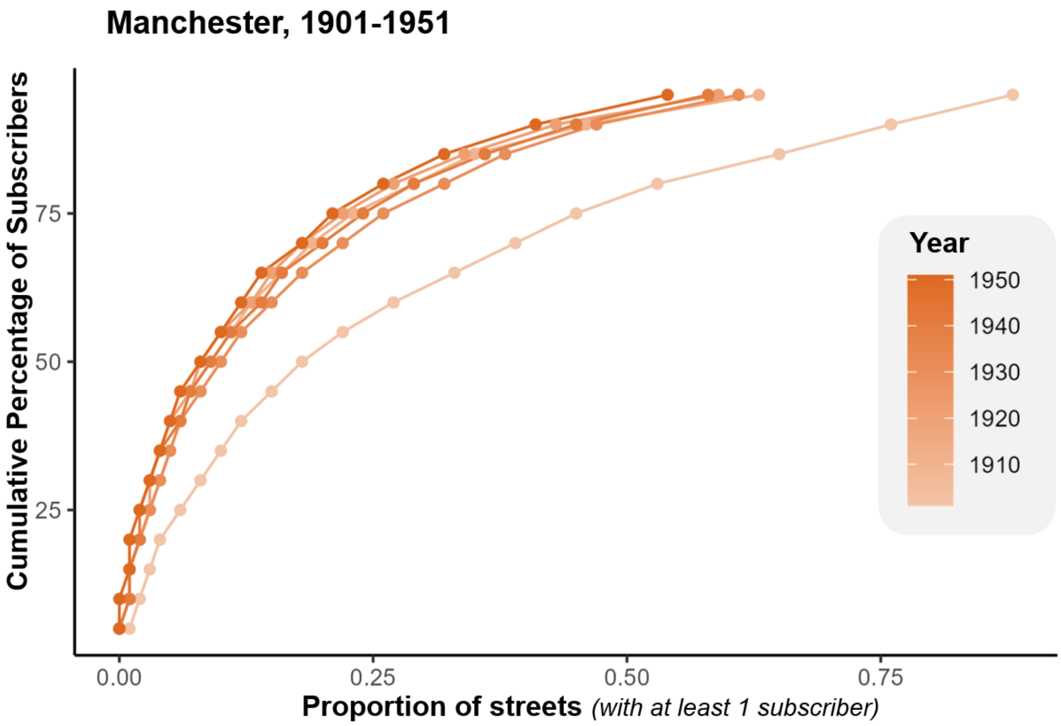


Extending the analysis above to the time periods beyond 1901, the sub-figures of Figure 21 depict how the patterns of concentration of telephone subscribers across the street segments encapsulated within the different settlements vary across time, with data beginning in 1881 for London and in 1901 for Glasgow and Manchester. At a glance, notable is that the patterns for all three settlements post-1901 are distinctly different to those in 1901 or before. In Manchester and Glasgow, the graphs for telephone subscription patterns in 1901 take a form that maintains a constant gradient for much longer than expected, in comparison to the graphs for all in the following years. This is the probable result of there being very few residential telephone subscribers in these settlements and the cumulative frequency graph takes a more expected form in the latter years. For London, however, it is less clear why the patterns of concentration and dispersion of subscribers are so markedly different prior to and during 1901 as opposed to after. With every passing decade up till 1901, telephone subscriptions within London become more concentrated in fewer streets: 89% of the streets most populated with subscribers accounted for 95% of all subscribers in 1881, with this figure dropping to 60% in 1911. Thereafter, from 1911 to 1951, the trend reverses and telephone subscriptions become gradually more dispersed again, the same figure from before rising from 60% to 70% in 1951.

Figure 21: Cumulative frequency (percentage) chart of distribution of telephone subscription across street segments of each settlement through 1951







The patterns in Manchester and Glasgow are clearly the reverse of those in London, with telephone subscription only becoming more and more concentrated in fewer streets with the passing of years. In 1911 63% of Manchester's streets hold 95% of all its residential subscribers and this falls to 54% by 1951; in Glasgow in the same years, the figures similarly fall from 63% to 57%. As compared to London, this may reflect how the telephone subscription remained concentrated in parts of these settlements that had already been dominant in 1901 – the South of both Manchester and Salford Registration Districts and the Parish of Glasgow, respectively. All in all, these make residential telephony more spread out across the urban street network in London than Glasgow or Manchester which might point to a mature telephone service network that has also undergone more expansion over the longer period of time over which it has been served by telephones. This may also suggest that although overall rates of telephone adoption were rapidly increasing across the country and first and foremost in the three largest settlements, the growth in adoption in places where telephones were already frequently adopted maintained a steady pace and, in some cases, grew at a faster rate than in the more newly serviced areas.

4.2.3 Attribution of Geodemography to Subscriber Records

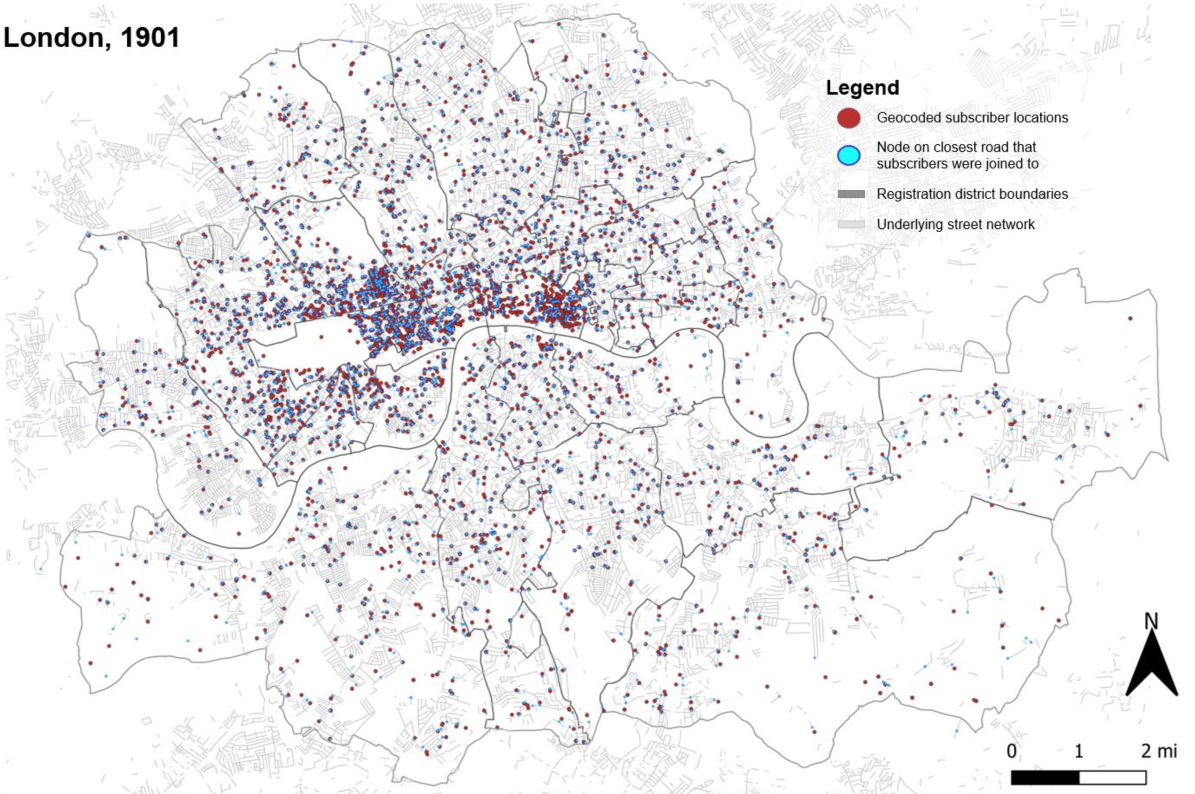
Through the quantitative lenses adopted in the preceding two sub-sections, this thesis had explored variations in the dataset of telephone subscribers: by mapping location quotients, one could glean how spatially unequal the adoption of telephones was within each settlement; by charting the spread or concentration of telephone subscribers across the streets of each settlement, one could visualise the nature of the change of their telephone networks and whether they democratised the access to telephone locally, or did otherwise. Yet, something that has not been explored is the diversity of reasons for why a residential subscriber might adopt the fixed-line telephone, as before every residential telephone subscriber which treated as functionally no different from one another. The following analysis therefore seeks to explore how the telephone subscriber data could be combined with population data from existing research, so as to enable a deeper understanding of what kinds of people tended to subscribe to telephones as residential subscribers, and how they may be distributed in the intra-urban context. More than mere volumes of telephone subscriptions, this analysis sought as well to examine the traits of telephone adopters and how adoption may, for instance, be geographically concentrated for different types of residential telephone subscribers in different areas and this in turn draws upon the idea of geodemographic analysis, which, simply put, is ‘an analysis of people by where they live’ (Sleight, 1997: 16). Although geodemographic analysis is more commonly used in the context of data emergent from the most recent two or three decades, in this analysis, this was operationalised by capitalising on the work that Lan and Longley (2021) have undertaken to characterise small areas in Britain between 1881 and 1901, with data derived from the digitised historical Censuses.

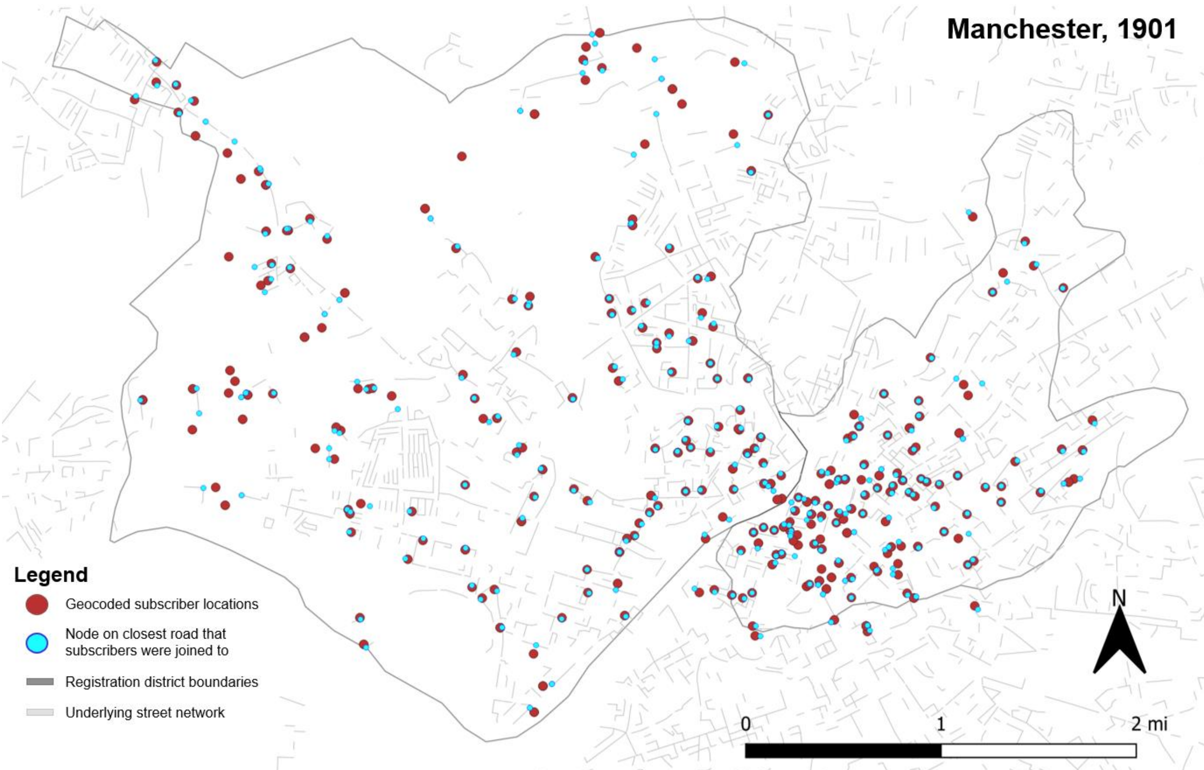
Having collated information relating to aspects of employment and occupation, residential mobility, as well as demography, Lan and Longley (2021) used the k-Means clustering algorithm to summarise major differences across a large number of variables in numerous spatial units (in this case street segments). Their analysis established a six-category typology to represent distinct and then-contemporary social structures, with only five being important to the urban context – since the sixth category was labelled 'rural residents'. The result was that each urban street segment in Britain had been allocated to one of these categories as a way of indicating what the profile of their residents appeared to be and these categories are, namely: 'high social status households, 'sales and service families', 'artisanal communities', 'hard-pressed production families' and 'poverty and casual employment'. These category labels are by no means definitive, nor are they all-encompassing, of the characteristics of the historical populations of Britain. However, such a classification can serve as a starting point for this research, and potentially other interested research in the future, to add a layer of differentiation to the residential telephone subscribers, which until this point had been characterised almost only by their geographical location. This also enables an exploration of patterns of the spread of telephones through a dimension other than mere subscription volumes.

The method of doing so, as well as the input datasets required, will be hereafter described. The three datasets used in this analysis were as follows: shapefiles of the modern street network of Britain, the allocations of these street segments to one of the aforementioned historical geodemographic categories (in a tabular format) and the digitised information of residential telephone subscribers that had been produced by this research. The shapefiles of British street segments were obtained by this research from the Ordnance Survey Open Roads product, which covers most of the modern street network of Britain and were the files with which Lan and Longley (2021) had created the aforementioned geodemographic classification. This research first subset the shapefiles to give only the street segments within London and Glasgow, the settlements of interest. Then, the geodemographic classifications for each street segment, which were separately obtained from the authors, were joined to the street segments through a unique identifier field as defined by the Ordnance Survey, to give street segments that each corresponded to one of six geodemographic categories. Thereafter, the records of residential telephone subscribers, whose geographic attributes exist in tables in the form of coordinate pairs, were converted into shapefiles in which every point corresponded to one residential telephone subscriber, and one shapefile existed for each settlement and each decade. The final step was to match, through a spatial join operation, the point representing each subscriber to the node on the nearest street segment as measured by perpendicular distance. The subscriber is then linked in the dataset to the said street segment and, for purposes of this analysis, is taken to be residing on that street; the geodemographic classification attached to the street segment is then transferred and attributed to the point representing the subscriber.

The sub-figures of Figure 22 show the distribution of telephone subscribers vis-à-vis points on the nearest street segment to which they were joined across the three chosen settlements in 1901. The data used to plot the figures were chosen to be from 1901 because it is the year in the first decade in which telephone subscriber information is available for all three settlements. Moreover, as the density of subscription in London increases substantially in 1911, the differences in the density of subscribers in London's centre and outer districts would become too large for the viewer to be able to easily perceive at a glance, even as the intra-urban patterns in London in 1911 would likely continue the trends observed in 1901, as shown earlier in this chapter. Given that the geodemographic classification was originally created using Census data from the years 1881, 1891 and 1901, an analysis using telephone directory data from 1901 would be contemporaneous and thus more comparable with the original Census data.

Figure 22: Distribution of telephone subscribers (in red) and nodes on nearest street segment onto which they were linked (in blue) by Registration District or Parish





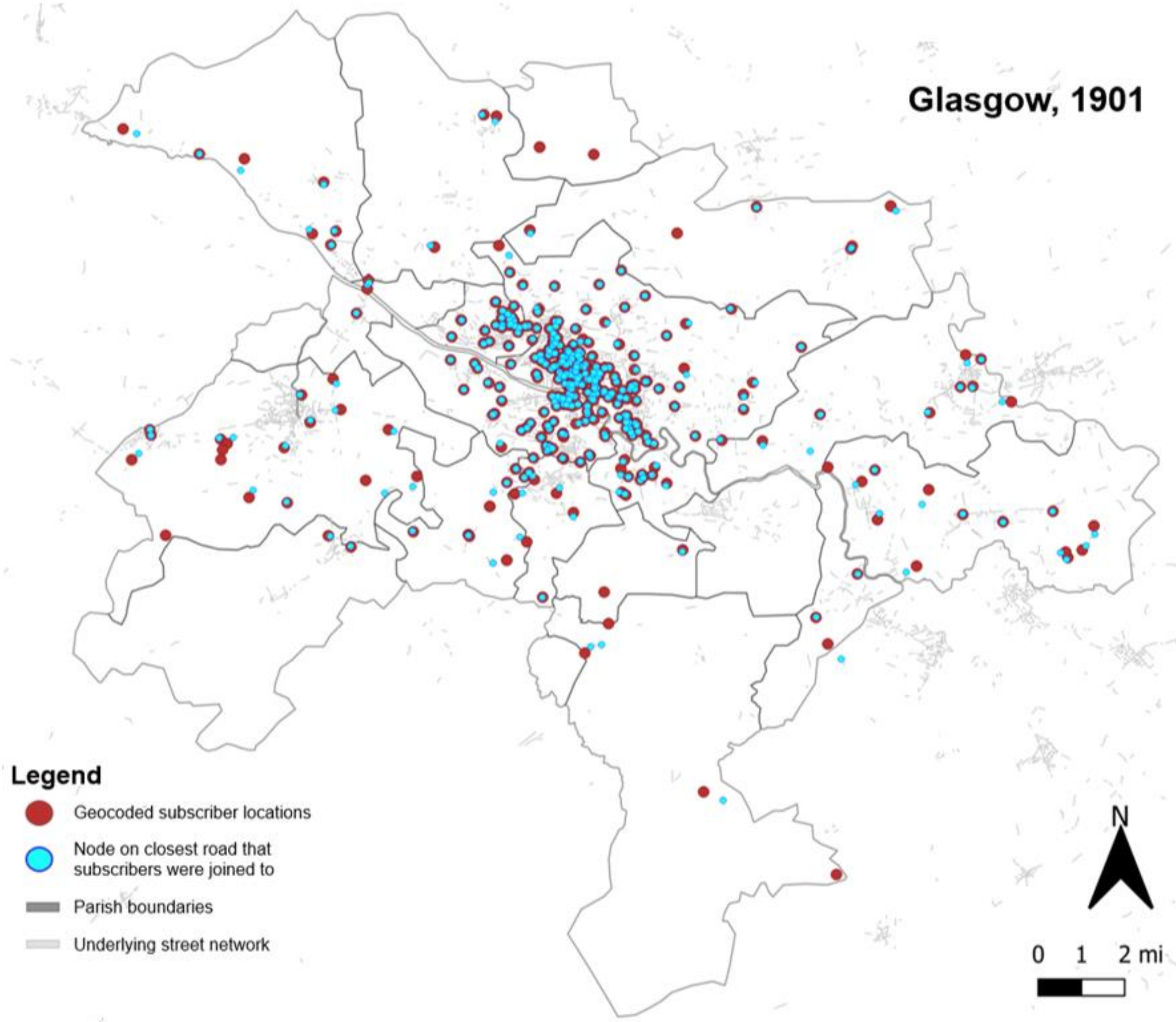


Figure 22 shows variations in the accuracy of spatial joins done on the telephone subscriber records across London, Manchester and Glasgow. While the areas of concern are assuredly non-rural and in theory should have excellent coverage within spatial administrative datasets, street networks have probably undergone marked changes in the century that has passed since 1901. In London, a noticeable North-South divide, marked by the River Thames, is perceptible: street networks terminate abruptly both North and South of this geographical feature running through the settlement centre and there are no points corresponding either to telephone subscribers or nodes on street segments which overlap this unbroken line. The density of telephone subscription is also highest in the districts immediately North of the Thames including Strand, City of London and Holborn. In opposition, the coverage of modern street networks is sparsest in the largest historical districts South of the River, namely Wandsworth, Lewisham and Woolwich. The implication of this variance in street network density is that telephone subscribers in these districts are more likely to be matched to streets that are further away and therefore less likely to be the streets on which they actually reside. This outcome arises from the aggregation of telephone subscribers to street segments having been done in a two-stage process: subscriber records were first attributed geographical coordinates through matching their address strings against a list of already geocoded addresses from the Census of 1881. Subscribers were then aggregated to the street level by means of spatial joins, not string matching, to allow for exploratory analysis at this scale. The modernity of the street networks used for the aggregation is likely why some subscribers in Figure 22 are shown to be located somewhere not precisely on a street.

A pattern similar to that in London is also observed in Glasgow, where telephone subscription concentrates around historical Scottish Parishes around the River Clyde that flows through the city, notably with the inclusion of the Parish of Glasgow. Both street network coverage and the density of residential telephone subscriptions becomes much sparser in the Northernmost Parishes, like Old Kilpatrick and New Kilpatrick, and Southernmost Parishes such as Neilston and East Kilbride. For Glasgow, these patterns may also be indicative of an underlying issue of starker inconsistencies between past and present civil and telephone exchange geographies that implicates on the processes of geocoding and then cascades to affect the spatial join operation with the street networks. This issue will be discussed at greater length shortly below.

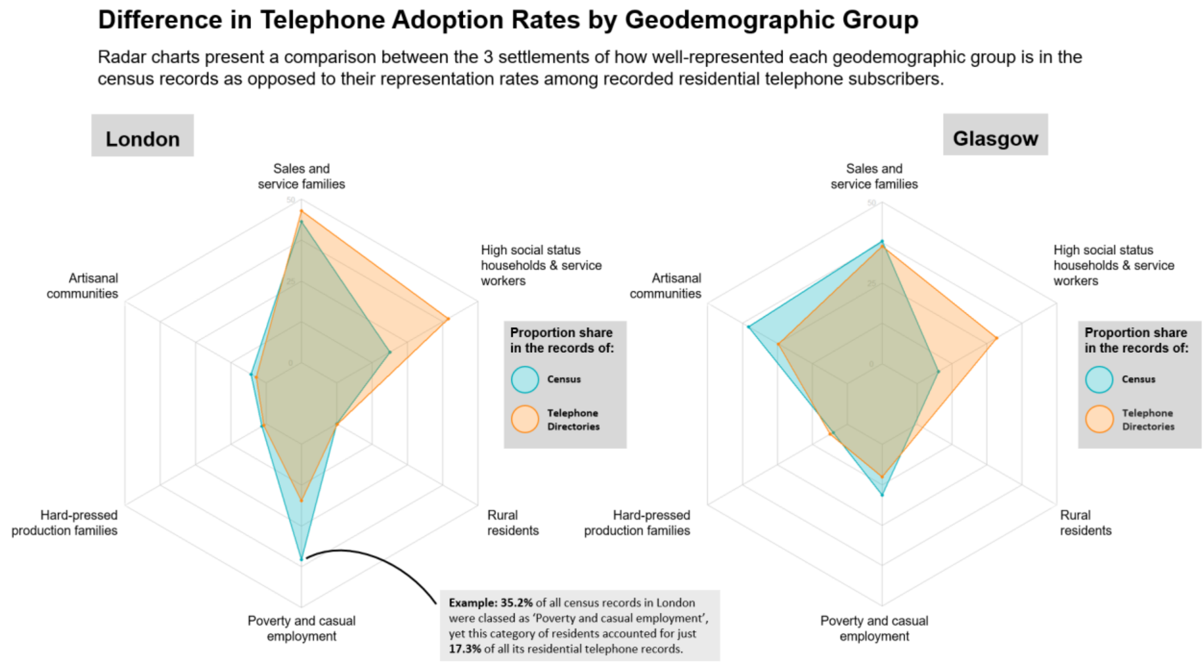
4.2.4 Settlement-wide Geodemographic Profiles of Telephone Ownership

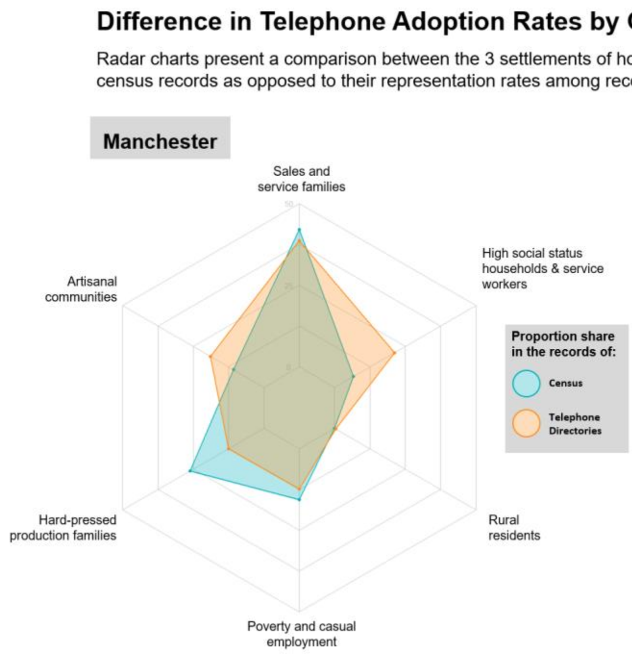
Having explained the process of attributing residential telephone subscribers with geodemographic characteristics above, this following sub-section presents results of the initial analyses undertaken with regard to uncovering how use-cases for the telephone differed across the three settlements of London, Manchester and Glasgow. The aim of this series of visualisations is to provide a starting point for the exploration of how the patterns of adoption of fixed-line telephones could be characterised beyond the numerical summaries that had been presented in the chapter prior, and beyond simply where its physical infrastructure has spread to, since telephone networks were and are commonly described by way of the number of exchanges and processor units they encompass, or in terms of the extents of its cable wiring (British Telecommunications, 2006).

The visualisations present an alternative illustration of the spatiality of telephone adoption by focusing on a differentiation of the characteristics of its users, uncovering some biases that are inherent to what who had been represented in the list of telephone subscribers, and in which periods relative to the lifespan of the telephone. This would be meaningful given the highly selective the coverage of the telephone directories data, given as well that merely 35% of British households had access to a household telephone in 1970, just 14 years before the last data point in this dataset. (Office for National Statistics, 2019). As such, this research found it important to conduct baseline analyses on which strata of the population would be overrepresented, and others underrepresented, owing to reasons of particularities in their consumption patterns. A seemingly obvious facet of this unevenness in representation, for one, would be that those in the higher social classes, regardless of location, were likely to be early adopters (Casson, 1910). Moreover, connectivity on the telephone network was also hugely shaped by preexisting social relations (Hamill, 2010), as those who had an adopter of the technology within their social sphere were more likely to obtain one themselves, leading to a possible scenario whereby the early adopters of the telephone can exert considerable influence on who subsequently adopts the telephone. On this note, it would be potentially interesting to examine if there existed strong spatial clustering of telephone subscribers belonging to the same geodemographic groups, on the basis that these people were more likely to be in each other's social circles, and thus exert some influence on whether others in the circle would adopt the telephone.

To begin this exploration, Figure 23 depicts radar charts which compare, for each settlement, the representation of individuals in each geodemographic grouping with their collective representation in the telephone directory records of 1901. Representation of each geodemographic group in each dataset (Census or telephone directories) is defined by the percentage of all individual records in that dataset which is accounted for by individuals belonging to that group. A geodemographic group is considered overrepresented in a settlement if its representation in the telephone directories exceeds that in the Census; it is considered underrepresented if the reverse is true. As a note on how calculations had been done, Census records were demarcated as belonging to each settlement if the historical street they had registered as a residential address fell within the spatial extent of these settlements, as defined through the contiguity of historical street networks by Lan and Longley (2021). Referring to the figure, for instance, the geodemographic group labelled 'Poverty and casual employment' in London is underrepresented in the telephone subscriber dataset as it accounts for 35.2% of all Census records, but only 17.3% of records in the digitised telephone directories.

Figure 23: Charts comparing the demographic make-up of each settlement in the telephone directories, as compared to the national Censuses, in 1901.





Comparing the Census record make-up of the three settlements (the blue graphs), one constant is the high proportion of sales and service families in each settlement, with this figure being lowest in Glasgow, but in which this geodemographic group already accounts for a considerable 36.3% of all its residents. The representation of this group amongst telephone adopters also remains rather proportionate to its representation in the Census, as shown by the fact that the points on the blue (Census) and orange (Telephone Directories) graphs for are all relatively close to one another for each settlement. This suggests that such subsets of the population were rather avid adopters of the telephone in its early days, which seems to align with a report on the National Telephone Company in Britain from 1904, which suggested that telephones were in demand mainly from the wealthy and from businesses (The Economist, 1904); it would be sensible to assume that these populations could justify subscription to the telephone given that the increased ease of communication brought benefits to them in their line of work. However, a key difference between the settlements lies in which group, other than the sales and service families, are most highly represented in the Census records. In London, this group is those in poverty and casual employment, in Manchester the hard-pressed production families and in Glasgow, the artisanal communities. Of these respective geodemographic groups in their respective settlements, the artisanal communities of Glasgow seem to be the keenest subscribers to the telephone and account for just under a quarter of the city's telephone subscription in 1901.

When comparing the geodemographic profiles of telephone adoption (orange charts in Figure 23, London stands out from among the three in that its telephone user base is extremely skewed to just two groups: sales and service families, and high social status household and service workers (Figure 23), which together account for 78.0% of its telephone subscribers in 1901. The telephone adoption profiles of Manchester and Glasgow are similar to each other, with a more even spread of telephone subscribers between three groups, that are, the same two groups that were dominant in London with the addition of the artisanal communities. There is notably less concentration of telephone adopting circles in these settlements than in London: while two groups in London alone accounted for 78% of the settlement's subscribers, in Manchester the three largest groups accounted for just 74.6% of the total, while in Glasgow they together accounted for a comparable 83.8% of all its subscribers. These differences seem to imply that, from its onset, access to the fixed-line telephone was more democratised in the smaller cities, which were also later to get the telephone; in London its adoption seemed very concentrated among the well-to-do, who could afford to take up the service to convenience them in the most menial tasks (Dutton, 1992). This profile is also in line with what Casson (1910), in writing about the early history of the telephone, notes about a peculiarity of telephone adoption in the British, as opposed to the American context: there, early adoption was especially led by the aristocracy and by those who lived in, or had dealings in, the City of London. The radar charts, illustrating the relative composition in each settlement of the records in the Censuses and the telephone directories, therefore serve as a starting point for understanding how the diffusion of the telephone took place differently in these settlements alone, with probably yet larger differences to be uncovered if this analysis had been extended to other large settlements.

As a further supplement to Figure 23, Table 17 summarises these patterns, expressing them in the form of indices that compare the representation of each geodemographic group in the telephone directories, as opposed to their representation in the Census records of each settlement.

Table 17: Indices of proportion representation of each group in the telephone directories relative to their representation in the Census in 1901.

Geodemographic Group	Glasgow	London	Manchester
Sales and service families	0.96	1.08	0.92
Poverty and casual employment	0.64	0.49	0.8
High social status households and service workers	3.7	2.1	3.18
Artisanal communities	0.7	0.67	1.75
Hard-pressed production families	1.22	0.63	0.48
Rural residents	*	*	*

** These records accounted for 0.01% or less of all Census records for the settlement and thus were excluded*

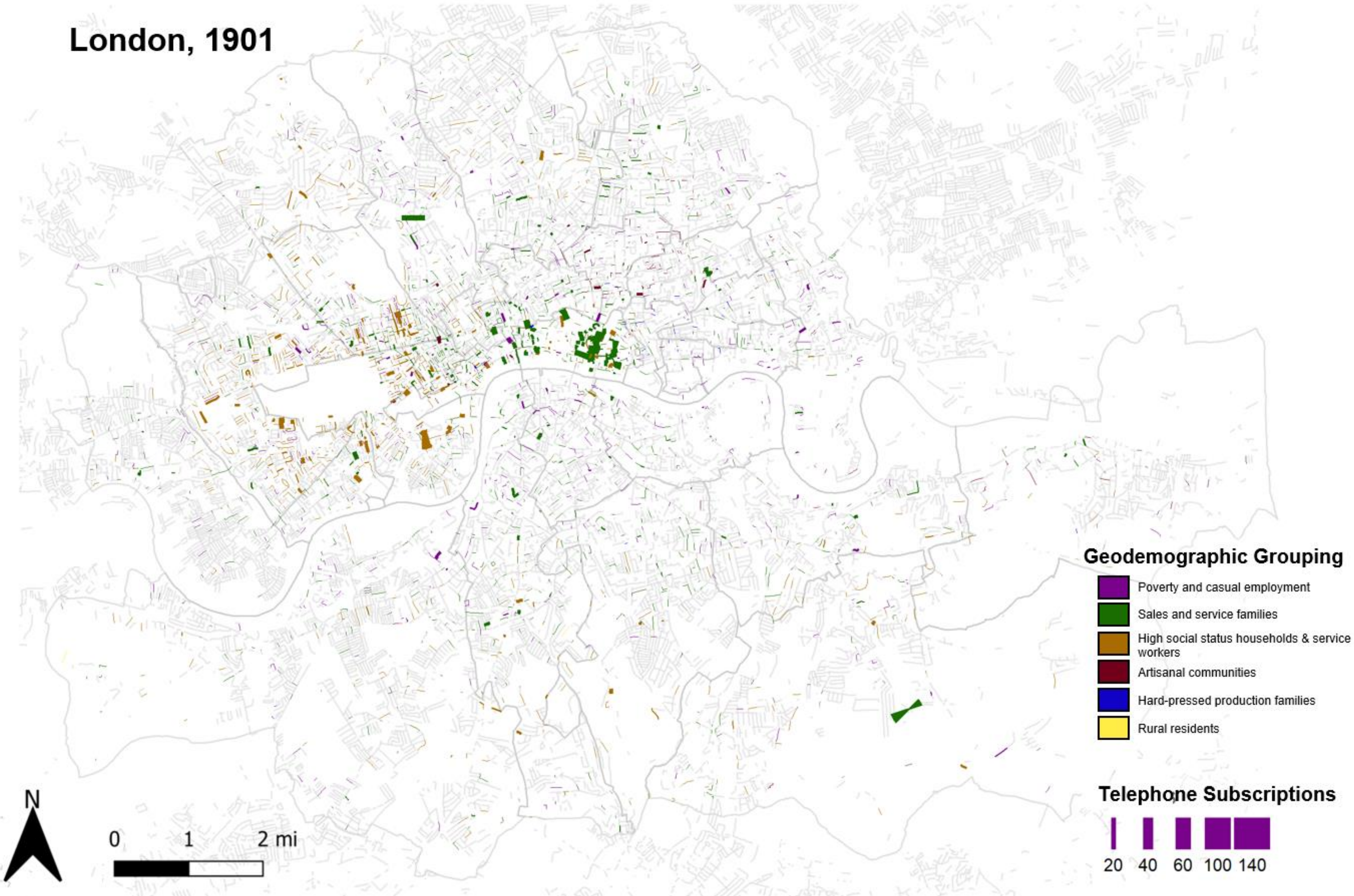
Common also to all settlements is that the representation of this geodemographic grouping among registered subscribers is extremely proportionate to their representation in the settlement's population as a whole: the ratios between these two figures fall within a narrow band ranging from 0.92 in Manchester to 1.08 in London. In comparison, the ratios of difference in representation in the two datasets vary much more widely for other geodemographic groups, in particular for the artisanal communities where, in Manchester, they are 1.75x overrepresented in the telephone directories but in London are 0.67x underrepresented.

Still, broad patterns that can be surmised from the different chart, including the fact that the high social status households and service workers are consistently overrepresented in the telephone subscription records – by as much as 3.7x in Glasgow. Although the telephone was thought of initially as almost exclusively a luxury good, the fact that service workers also began subscribing to the telephone in large numbers may correspond the beginning of a shift of the role of the telephone to a necessity, or even a social requisite (McGuire, 2019), at least for this subset of people. In other words, as telephone subscriptions became more financially viable for larger portions of the British population to rent, so did the use-cases for one also begin to diversify (Perry, 1977). Meanwhile, those in poverty and casual employment remain underrepresented across every settlement. The hard-pressed production families make up the smallest proportion of the recorded population of Glasgow in 1901 but yet are slightly overrepresented in the settlement's telephone directories. Given that it was unlikely that hard-pressed production families would be above average in their telephone adoption rates as opposed to other geodemographic groupings, this may have been a result of a combination of the small number of residential telephone subscribers in Glasgow that year and the troubles with harmonising the settlement's different geographies, as discussed earlier.

Moving on from the summaries of the geodemographic profiles of the residents of each settlement that were the radar charts, the following figures seek to depict the spatial distribution of the residential telephone subscribers, by their geodemographic grouping; Figure 24 to 26 are the preliminary results of attempts to encode records of telephone subscribers with the geodemographic information of the streets on which they resided, achieved by means of the aforementioned spatial join operations. Each figure, underlain by the modern street networks of that settlement and the boundaries of its Registration Districts (London and Manchester) and Scottish Parishes (Glasgow) in grey, depict the extent of telephone adoption in different street segments. These street segments vary both by width – the wider a segment, the more residential telephone subscribers it had – and by colour, which indicate which geodemographic grouping the residents of that street segment were classified under. Admittedly, it is not possible to examine patterns as closely or as quantitatively as was done with the indices of representation presented before, but these figures were aimed to be exploratory visualisations of the distribution of subscribers by geodemography (and thus implicitly by their different use-cases for the telephone) from which broad patterns can be gleaned.

In London, the Registration Districts with individual streets that hold the most residential telephone subscribers are concentrated to the North of the River Thames, in the city centre (Figure 24), mirroring what had been seen in Figure 22. The concentration of subscribers in the centre, other than being due to demand-side factors, could also be explained by supply-side factors: after 1901, the market for telephones in London was starkly divided, with the National Telephone Company predominating in the Eastern half, and the Post Office being more strongly present in the West, resulting in fierce competition to increase their respective market shares in the central districts (Perry, 1977). Another observation that can be made about the density of subscribers in the central districts North of the Thames is that there appears to be loose zoning of telephone subscription patterns in London by geodemographics. Here, subscribers who belong to high social status households and service workers appear to predominate in the western districts of Marylebone, Kensington and Chelsea while in the eastern central districts of Strand and City of London, subscribers are overwhelmingly sales and service families. These two geodemographic groups also correspond to the groups which, in London, not only accounted for the vast majority of all its telephone subscribers, but also were the only groups that were over-represented in the directories as compared to the Census, with indices of 2.1 and 1.08, respectively (Table 17).

Figure 24: Map of telephone adoption volumes and geodemographic groupings of street segments in London



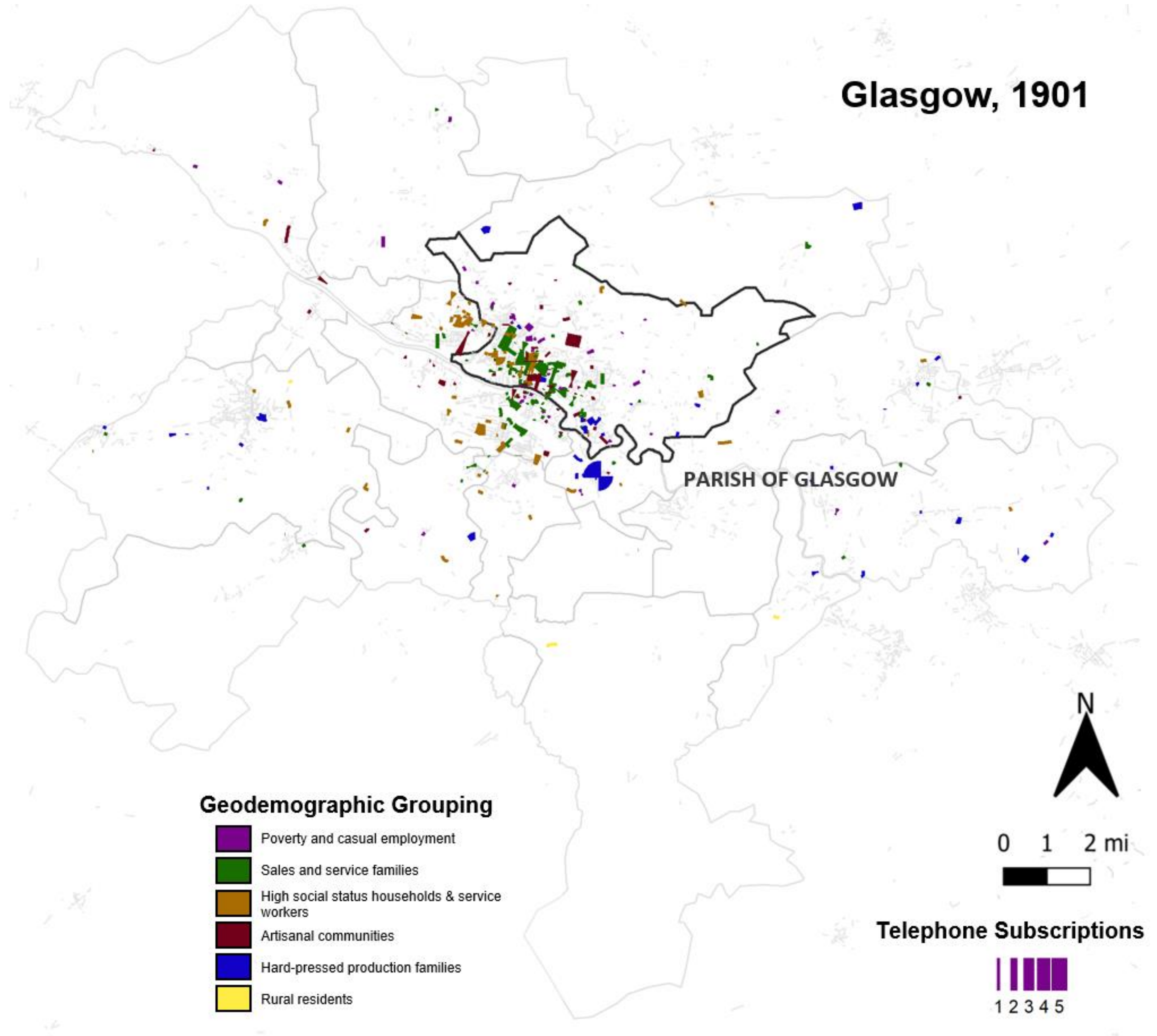
Meanwhile, in Registration Districts South of the Thames, on top of sparser telephone subscription patterns as a whole, it is more difficult to determine a dominant type of telephone subscriber as streets with residents that are in poverty and casual employment are interspersed with streets housing sales and service families.

The visualisation above provides an avenue for an exploration of the geographical patterning of telephone subscribers of different geodemography in London (as do the following two figures for Glasgow and Manchester). However, its usefulness beyond this, for in-depth analysis, is acknowledged to be limited. For one, this research also faced difficulties in illustrating the aforementioned variations on a map because of the vast differences between the concentration of telephone subscribers in street segments in the more central Districts, as opposed to those in the outer Districts. It was thus a delicate balancing act in choosing an appropriate scale on which to vary the widths of street segments, based on the volume of telephone subscription within them: on one hand, the bars representing streets with high telephone subscriptions would be too thick and obscure the bars representing adjacent street segments, while on the other hand, the bars representing streets with sparse telephone subscription rates would be made barely visible. Furthermore, regarding the precision of geocoding subscribers, it is difficult to definitively conclude if telephone subscribers in the areas which are, on the whole, sparsely populated with subscribers, were geocoded with a precision similar to that for subscribers in areas of concentration of telephone subscription. This is because in parts of the settlement with less dense street networks, where an exact match between a telephone subscriber's coordinates and the nearest street segment is not found, the next nearest segment is likely to be further away than if the subscriber were in an area with denser street networks. Nonetheless, because London subscribers had been geocoded in a way that took specific account of how its telephone exchange and civil geographies related to each other – by means of overlaying maps containing this information on top of one another – it can be said with confidence that its telephone network did extend considerably to the outskirts of the settlement as shown in Figure 24.

Similarly, Figure 25 depicts the spread of telephone subscribers by streets and their geodemographic classifications in Glasgow in 1901. Of the three settlements, the spread of Glaswegian telephone subscribers across its street segments is the most uniform of all, meaning that there was much less overt clustering of telephone subscribers in a select few streets. It is as such that the larger scale of representation of street segments in Figure 25 may make it seem that streets in the centre of Glasgow are comparably densely packed to those in London's centre. Highlighted also are the outlines of the historic Parish of Glasgow, which alone accounts for over 60% of telephone subscribers in the city in 1901, with most subscribers being matched to streets in its Southwestern corner. There are also notably next to no telephone subscribers being matched to streets on the city's Northern and Southern fringes because modern street networks very sparsely occur in these parts of Glasgow. The telephone subscribers joined to streets on the outermost fringes of the settlement are, unlike in London, more probably to be mismatched addresses as the telephone network of Glasgow, in one of its first years of service in 1901, was unlikely to extend far out of the settlement centre. This is on top of the fact that there were no maps of the telephone exchanges of Glasgow in 1901 or 1911 with which these extents could be ascertained. Relative to London, there is less clear zoning of telephone subscribers by geodemographic type but most of the centre block of subscribers are either on streets with sales and service families or high social status families and service workers, with some streets of artisanal communities also in the mix. Interestingly, some subscribers were matched to less central streets that had been typified as having hard-pressed production families as residents.

One plausible factor that contributed to these oddities in encoding telephone subscriber records for Glasgow would be the unusually prominent lack of harmonisation of civil geographical and telephone exchange geographies for the region surrounding Glasgow. It proved a challenge to find a consistent definition of Glasgow city in the early 20th century as, although the Parish of Glasgow existed, it was very much connected to neighbouring Parishes, especially to Govan, which it borders on its West and Southwest sides, as seen not least from the density of modern street networks that still connect the two historical Scottish Parishes. Furthermore, brief descriptions on the information pages preceding the telephone book sections for Glasgow in the raw data made clear indication that its coverage went beyond the local area to even, in most directory entries, include the neighbouring settlement of Paisley – in the 1920s constituting part of the Parish "Paisley with Abbey" and whose area overlaps with the boundaries of today's Glasgow City Council.

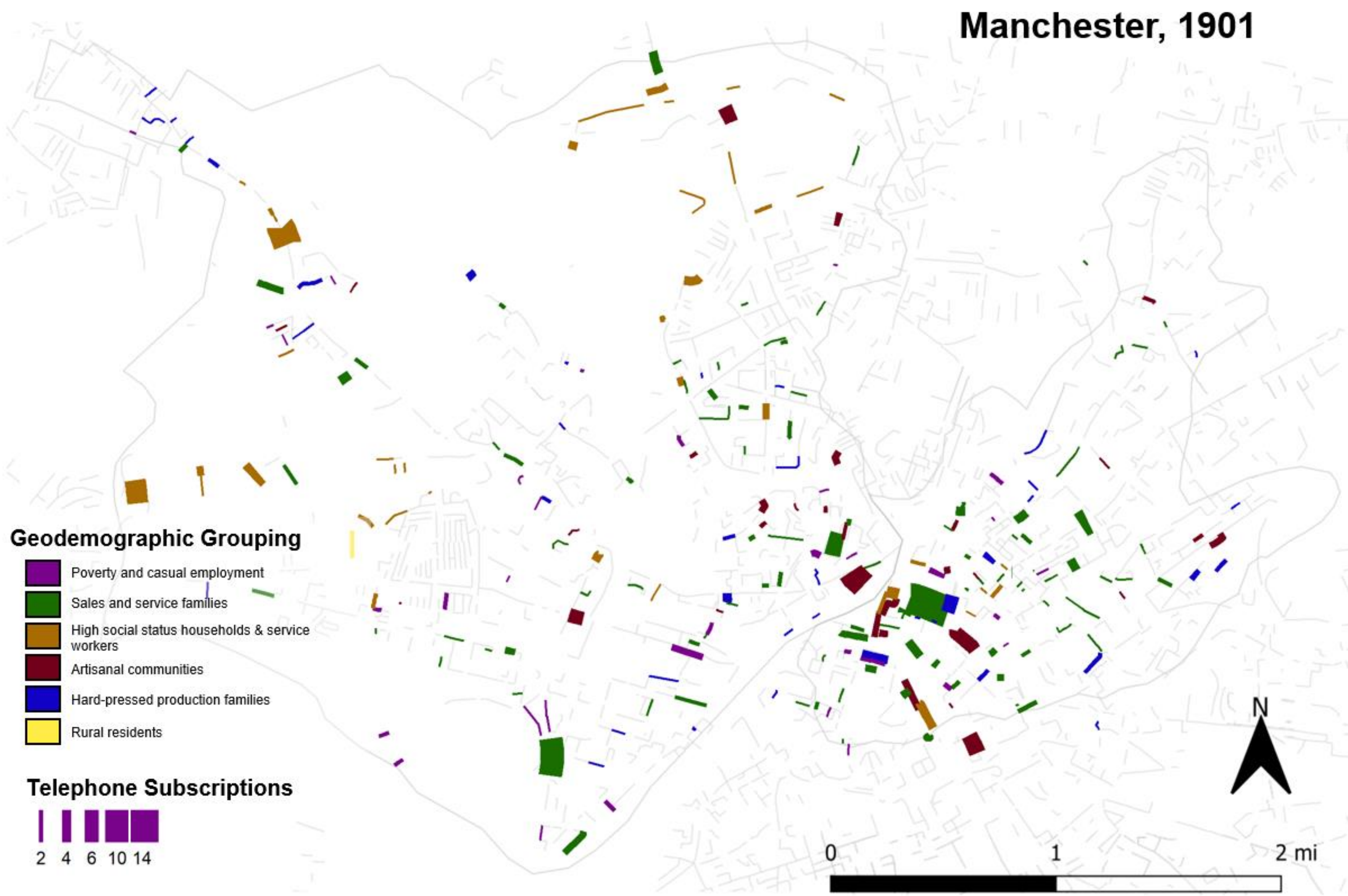
Figure 25: Map of telephone adoption volumes and geodemographic groupings of street segments in Glasgow



Finally, Figure 26 shows the distribution of telephone subscribers in Manchester, which was taken in this analysis to comprise both the Registration Districts of Manchester and Salford, as had been indicated on the information pages of the telephone directories. As in Glasgow, there seems to be little zoning of subscribers from different geodemographic groups but most subscribers appear to reside in the South of the city. Streets that hold more telephone subscribers in this Southern agglomeration are usually streets containing sales and service families, or artisanal communities, the latter of which is much better represented relative to other geodemographic groups as opposed to their counterparts in the other settlements. Compared to the telephone subscribers of London and Glasgow, the spread of telephone subscribers (and therefore, by extension, of the telephone network there) in Manchester is much more confined to a smaller area; this is evidenced by the much larger scale of representation in Figure 26 than in the previous figures representing the other two settlements. This may point to how the size of telephone exchange areas in Northern and Central English settlements were intentionally designed to be more restricted, in order to preserve the economic interests of its competitor, the state-owned telegraph system (Perry, 1977).

Altogether, the exploratory analysis presented in the latter half of this chapter has demonstrated possibilities in combining telephone subscription data with external sources of information about historical populations, derived from the Census, to paint a more vivid picture of the different types of telephone adopters, how they are spatially patterned on more local scales and how their overall make-up differs from settlement to settlement. When this analysis is extended to other settlements beyond the three largest cities, comparisons of patterns of telephone subscription can be made in more detail, perhaps moving beyond characterising individual subscribers to typify settlements in how they interact with this new technology, in the veins of research already conducted on residential differentiation across Great Britain.

Figure 26: Map of telephone adoption volumes and geodemographic groupings of street segments in Manchester



4.3 Conclusion

This chapter has presented an overview of the investigation which was carried out into possible applications of the newly digitised telephone directories dataset. Containing information on telephone subscribers that span several decades and with coverage for most of the major urban settlements of Britain, the dataset could be used either on its own or as a supplement to other related data to study questions at various spatial and temporal scales. It could serve as a resource that captures the patterns of diffusion of an emerging technology, as well as providing important information about historical populations. Given that it holds encoded records that exist at the disaggregate level of individual persons or households, the dataset enables analysis that will mitigate against the ecological fallacy, which is something that is hard to achieve because of the general lack of availability of disaggregated data for studying historical phenomena.

Leveraging on the contemporaneity of granular Census records available until 1911, this research explored how these data could be used to study the geodemographic provenance of the recorded telephone subscribers in the three largest settlements in Britain. A comparison between the geodemographic composition of the Census records and of the telephone directory records gave more contextual clues as to the dominant nature of telephone subscriptions in each settlement, for instance as to whether they were a mere luxury item that was affordable only by large swathes of high social status households, or if they were more practical investments for which professionals, artisans and other workers found uses. The probabilistic differentiation between commercial and residential subscribers provides yet another avenue for subsequent research to characterise telephone usage in different settlements, as while this exploratory analysis has focused on analysing residential patterns, the telephone served as an equally important breakthrough for commercial users. While the analysis in this section had also been limited to the largest urban settlements, it could be extended to the numerous other British settlements to allow for a wider scope of comparison.

Beyond 1911, in the absence of a granular Census-like dataset serving as a basis for comparison, this analysis explored how the telephone subscription data could be contextualised in more self-contained manners. The research looked into creating a measure of inequality of telephone access through cumulative frequency charts that depict how residential telephone subscribers were distributed across the street networks of a settlement. An understanding of innovation diffusion might suggest that telephone uptake should become more spread out across the street networks as time passes, but initial findings showed that while this may be the case, the growth of adoption rates in areas that were early adopters might well outpace those in the areas containing later adopters. Maps of location quotients are also plotted to show how intra-urban inequalities of residential telephone uptake are rather persistent over time in the early period of telephone service. When telephones become more widespread in the latter half of the 20th century, there is potential for this analysis to be expanded to compare variations across geographical units not just within one settlement, but across settlements, in order to facilitate the identification of forerunner or laggard areas on a national scale.

5 Conclusion

The main limiting factor that has been imposed on the analyses of historical phenomena has remained, for a long period, the time available for researchers to pour into the examination of materials. Familiarity with and subject expertise on the topic being researched would indubitably reduce the immensity of this effort, but the comprehensiveness of these studies usually hinges at least partially on the volume and variety of sources manually perused, meaning that large time investments are unavoidable. Beginning in the 1960s, what was dubbed the quantitative revolution in geography would set geographic research on the path of employing an ever-increasing range of quantitative techniques that are now considered commonplace, among which are statistics, spatial analysis, and modelling and machine learning. More recently, in the age of big data, the accompanying increase in the amount of data available for social science research has been drastic owing to the opening up of 'alternative' datasets, originating not from official sources but rather from business and open-source platforms. Amidst all these developments, the study of historical geographies has been left mostly unturned because of its non-conformity to new data structures and perceived incompatibility with newer quantitative methods, with much of it still resting upon the shoulders of traditional, tried-and-tested and almost always qualitative methods. This thesis argues and demonstrates that this long-standing pairing need not remain the only viable route, with the opening up of telephone subscription data as an alternative source of data on the historical populations of Britain that is also compatible with quantitative analysis methods used on contemporary big datasets.

Summary of Findings

This research has shown that the coming together of recent technologies of image text capture, data manipulation and organisation of large databases has enabled the collection of archived telephone directories procured to be digitised and brought in line with the format of more recent large datasets. Comprising numerous steps and having undergone a process of iterative refinement, the processing pipeline developed with the help of colleagues is semi-automated and robust in its operation on telephone directory data with layouts that constantly change through editions. First, the pages holding key information – that of telephone subscribers – are identified and systematically separated from pages with other content. The text embedded in these pages are then extracted through OCR to give unstructured blocks of text. The next series of steps restructures the text data into a tabular format, with each row containing information of one subscriber and each column a different type of information about them. This processing is tailored to each unique directory layout that is identified among a select sample of all directories with the help of a library of modularised functions that was created, 'teldiR', which also constitutes an output of this research that is accessible online. Finally, geocoding of subscriber records in select urban settlements aligns them with more precise geographical coordinates, making them fit for use with further spatial analysis techniques. Uncertainty is invariably present along each step of this pipeline, and the thesis makes note of these as comprehensively so possible, so that end-users are aware of their impact on the provenance of the data. This research argues that this dataset, while with its issues, offer a real improvement over the current dearth of granular population data especially for Britain for most of the 20th century.

The thesis then proceeds to examine the utility of the digitised data for studying both innovation diffusion of fixed-line telephones and the geodemographic characteristics of early adopters of the telephone. In attempt to conduct exploratory analysis with a wider geographical coverage of Britain, it attributes a geography at a more macro-level, but to much more subscribers, than had been done for select records at the end of the pipeline. This process produced figures that enabled a step-by-step comparison of telephone subscription volumes and adoption rates per unit population in the 53 largest urban settlements of Britain in 1901. From this, strong regional divides in telephone adoption were observed, with results also showing that the time at which a settlement first receives telephone services does not strongly correlate with how adoption rates stand by 1951. Another interesting inference made was that, although telephone adoption volumes were highest in the largest British settlements including London, Manchester and Glasgow, these same places did not rank nearly as highly in terms of adoption rates, implying that while telephone subscriptions may have increased more rapidly here, the gap between adopters and non-adopters remained relatively high compared to smaller settlements. This calls into question as to whether the innovation diffusion of telephones that took place followed a strict logic of urban hierarchy, passing down from larger settlements to smaller ones, or if other, possibly more local factors exert stronger influence on the process, such as personal circumstance.

Rounding off the exploratory analysis, the thesis explores possibilities for analysis of the dataset in tandem with other major sources of quantitative historical information, the sole accessible one being the digitised Census records. Using the surnames and street names of telephone subscribers' addresses, a method to fuzzy match telephone subscriber records to their contemporaneous Census records is devised for subscribers in the three largest settlements. Such matching at the individual level would serve to enrich understandings of the demographic profile of people that were among the earliest telephone subscribers in Britain. The results of this matching was far from perfect, and so an alternative approach was to conduct linkage at the more aggregate scale of street segments, rather than individuals. The results illustrated that the mix of geodemographic profiles that dominated each settlement's telephone subscribers was highly distinct from each other, as was their spatial distribution within the settlement. This exercise showcased how nuance can be added to understandings of the contents of the telephone directories dataset by corroborating it with other sources of contemporaneous information about the same population and suggests that this means of analysis can be further explored in future research.

Further Work

While the analysis done by this research into the applications of the telephone directories dataset has only been exploratory, it hopes to lay the groundwork for future studies in the many disciplines that are concerned with examining various aspects of society in British history. The first possible avenue of extending this work, as mentioned before, is for the dataset or subsets thereof to be used to study other research questions while bearing in mind the provenance of the dataset and the fact that it far from represents the entire population. The dataset that was created itself provides a resource equally about the persons who are represented in it as about the expansion of the service for which it was created. Just as Census data had been used to augment an analysis of the types of people that tended to be subscribers, overlaying onto the analysis other related data could facilitate the study of factors that may have affected either the provision or consumption of telephone services.

Another pathway for future work is by extending data capture from the telephone directory archives. The ready-built data processing pipeline coupled with 'teldiR' – the library of functions made available online – enables interested parties to expand its coverage or improve on the data capture that has already been done within a narrower frame of geographical and/or temporal focus. Where a researcher's particular interest is focused on periods or locations for which data has yet to be digitised, they would be able to adapt code already written to replicate the digitisation process for those data; where they feel that the digitisation already done could be improved, they could tailor the code to specific samples of records, thereby improving the quality of data capture.

Finally, the very actualisation of the concept of a pipeline which extracts and assembles structured information from an initially unstructured collection of digital archives is noteworthy. The constituent parts of this pipeline are automated and, where this was not possible, data processing was modularised into smaller functions to facilitate troubleshooting on the part of the end-user. With the thesis showing that semi-automated data capture could be workable on a raw dataset whose production spanned over a century and in which there existed immense variability, this insight might signal the possibility for similar methods to be applied on other kinds of massive historical data that have yet to be digitised, without the need for intensive human labour. It is hoped that this approach would prompt more social science researchers to reopen inquiry into old research questions with the toolkit of new quantitative methods in hand.

References

British Telecommunications (2006) 'Annual Report 2006'. Available at: https://www.bt.com/content/dam/bt-plc/assets/documents/investors/financial-reporting-and-news/annual-reports/2006/bt-annual-report-form-20-f-2006.pdf (accessed April 2024).

British Telecommunications (2007) 'BT an predecessors - a corporate timeline'. Available at: https://www.bt.com/bt-plc/assets/documents/about-bt/our-history/bt-archives/information-sheets-and-timelines/bt-and-predecessors.pdf (accessed April 2024).

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

Calvo, A. (2006) 'The shaping of urban telephone networks in Europe, 1877-1926', *Urban History*, Cambridge University Press, 33, 411–434.

Campesato, O. (2019) *Regular Expressions: Pocket Primer*, Dulles, VA: Mercury Learning and Information.

Casson, H.N. (1910) *The History of the Telephone*, Chicago: McClurg.

Christaller, W. (1966) *Central Places in Southern Germany*, Prentice-Hall.

Connected Earth (2006) 'Shaping our lives', *Connected Earth (Archived site)*. Available at: https://web.archive.org/web/20061108143348/http://www.connectedearth.com/Galleries/Shapingourlives/Livingwiththetelephone/index.htm (accessed April 2024).

Crutchley, E.T. (1938) *GPO*, Cambridge: Cambridge University Press.

Dennis, R. (1991) 'History, Geography, and Historical Geography', *Social Science History*, Cambridge University Press, 15, 265–288.

'Dial B for Britain: The Story of the Landline' (2017) *Timeshift*, BBC Four.

Dowle, M. & Srinivasan, A. (2023) 'data.table: Extension of \`data.frame\`'. Available at: https://r-datatable.com, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.table. (accessed September 2023).

Dueire Lins, R., Guimarães Neto, M., França Neto, L. & Galdino Rosa, L. (1994) 'An environment for processing images of historical documents', *Microprocessing and Microprogramming*, 40, 939–942.

Dutton, W.H. (1992) 'The social impact of emerging telephone services', *Telecommunications Policy*, 16, 377–387.

Flegg, A.T., Webber, C.D. & Elliott, M.V. (1995) 'On the Appropriate Use of Location Quotients in Generating Regional Input–Output Tables', *Regional Studies*, Routledge, 29, 547–561.

Freshwater, R. (2024) 'History of the United Telephone Company', *British Telephones*. Available at: https://www.britishtelephones.com/histutc.htm (accessed April 2024).

Great Britain Historical GIS Project (2017) 'Great Britain Historical GIS Project', University of Portsmouth.

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

Gurr, D. & Hunt, J. (1998) *The Cotton Mills of Oldham*, Oldham: Oldham Education & Leisure Services.

Hägerstrand, T. (1965) 'A Monte Carlo Approach to Diffusion', *European Journal of Sociology*, 6, 43–67.

Hägerstrand, T. (1967) *Innovation Diffusion as a Spatial Process*, Chicago: The University of Chicago Press.

Hamill, L. (2010) 'The Social Shaping of British Communications Networks prior to the First World War', *Historical Social Research / Historische Sozialforschung*, GESIS - Leibniz-Institute for the Social Sciences, Center for Historical Social Research, 35, 260–286.

Hemmeon, J.C. (1912) 'The Post Office and the Telephone Companies', in *The History of the British Post Office*, Cambridge: Harvard University Press, 219–236.

Holcombe, A.N. (1906) 'The Telephone in Great Britain', *The Quarterly Journal of Economics*, Oxford University Press, 21, 96–135.

Jahangir, W. & Zia-ul-Haq (2023) 'Integrating Technology Acceptance Model, Theory of Diffusion of Innovations and Theory of Planned Behaviour to Study the Adoption of Facebook Marketplace', *NMIMS Management Review*, SAGE Publications, 31, 214–222.

Kay, M.A. (2014) *Inventing telephone usage: Debating ownership, entitlement and purpose in early British telephony*, phd. University of Leeds.

Kieve, J.L. (1973) *The electric telegraph a social and economic history*, Newton Abbot: David and Charles.

King, G. (2011) 'Ensuring the data-rich future of the social sciences', *Science (New York, N.Y.)*, 331, 719–721.

Kingsbury, J.E. (1915) *The Telephone and Telephone Exchanges: Their Invention and Development*, London: Longmans, Green, and Co.

Lan, T., van Dijk, J. & Longley, P. (2021) 'Family names, city size distributions and residential differentiation in Great Britain, 1881–1901', *Urban Studies*, SAGE Publications Ltd, 00420980211025721.

Lan, T. & 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, 320.

Lan, T. & Longley, P.A. (2021) 'Urban Morphology and Residential Differentiation across Great Britain, 1881–1901', *Annals of the American Association of Geographers*, Taylor & Francis, 111, 1796–1815.

Lansley, G., Li, W. & Longley, P.A. (2019) 'Creating a linked consumer register for granular demographic analysis', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182, 1587–1605.

Library of Congress (n.d.) 'Telephone and Multiple Telegraph', *Alexander Graham Bell Family Papers at the Library of Congress*. Available at: https://www.loc.gov/collections/alexander-graham-bell-papers/articles-and-essays/telephone-and-multiple-telegraph/ (accessed March 2024).

Mahler, A. & Rogers, E.M. (1999) 'The diffusion of interactive communication innovations and the critical mass: the adoption of telecommunications services by German banks', *Telecommunications Policy*, 23, 719–740.

Marsden, B. & Smith, C. (2008) *Engineering Empires: A Cultural History of Technology in Nineteenth-Century Britain*, Basingstoke: Palgrave Macmillan.

McGeevor, S. (2014) 'How well did the nineteenth century census record women's "regular" employment in England and Wales? A case study of Hertfordshire in 1851', *The History of the Family*, Routledge, 19, 489–512.

McGuire, C.A. (2019) 'The categorisation of hearing loss through telephony in inter-war Britain', *History and Technology*, Routledge, 35, 138–155.

Miller, H.J. & Goodchild, M.F. (2015) 'Data-driven geography', *GeoJournal*, 80, 449–461.

Mitchell, B.R. (2011) *British Historical Statistics*. Reissue Edition, Cambridge: Cambridge University Press.

Navarro, G. (2001) 'A guided tour to approximate string matching', *ACM Computing Surveys*, 33, 31–88.

Office for National Statistics (2011) 'Confidentiality'. Available at: https://www.ons.gov.uk/census/2011census/confidentiality (accessed October 2023).

Office for National Statistics (2019) 'Percentage of households with durable goods'.

Office for National Statistics (2021) 'National life tables – life expectancy in the UK: 2018 to 2020'.

Perry, C.R. (1977) 'The British Experience', in Pool, I. de S. (ed.) *The Social Impact of the Telephone*, Cambridge: The MIT Press, 69–96.

Pool, I. de S. (1977) *The Social impact of the telephone*, Cambridge, Mass. : MIT Press.

Post Office Telecommunications (1970) 'The Story of the Telephone: A Booklet for Students', Post Office Telecommunications.

Reffle, U. & Ringlstetter, C. (2013) 'Unsupervised profiling of OCRed historical documents', *Pattern Recognition*, 46, 1346–1357.

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A. & Puppe, F. (2019) 'OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings', *Applied Sciences*, Multidisciplinary Digital Publishing Institute, 9, 4853.

Robson, B.T. (2006) *Urban Growth: An Approach*, London: Routledge.

Ryan, B. & Gross, N. (1943) 'Acceptance and Diffusion of Hybrid Corn Seed in Two Iowa Communities', *Rural Sociology*, 8, 15–24.

Saha, S., Basu, S. & Nasipuri, M. (2014) 'iLPR: An Indian license plate recognition system', *Multimedia Tools and Applications*, 74.

Sawyer, R.K. (2005) *Social Emergence: Societies As Complex Systems*, Cambridge: Cambridge University Press.

Schurer, K. & Higgs, E. (2023) 'I-CeM Integrated Census Microdata (I-CeM), 1851-1911', UK Data Service.

Singh, S. & Grewal, S.K. (2012) 'Text Extraction and Character Recognition form Image using Mathematical Morphology and OCR Technique', *International Journal of Science and Research*, 3, 952–955.

Sleight, P. (1997) *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. 2nd edition, Henley-on-Thames: NTC Publications.

Stein, J.L. (1996) *Ideology and the telephone: the social reception of a technology, London 1876-1920.*, *Doctoral thesis, University of London.*, Doctoral. University of London.

Tesseract-OCR (2021) 'Improving the quality of the output', *Tesseract Documentation*. Available at: https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html (accessed March 2021).

The Economist (1904) 'The Government and the National Telephone Company', *The Economist, Vol. 62, Issue 3161, Pg. 523-524*. Available at: http://archive.org/details/sim_economist_1904-03-26_62_3161 (accessed May 2024).

The Postal Museum (2021) 'Postcodes', *The Postal Museum*. Available at: https://www.postalmuseum.org/discover/collections/postcodes/ (accessed February 2021).

Van Dijk, J., Lansley, G. & Longley, P. (2021) 'Using Linked Consumer Registers to Estimate Residential Moves in the United Kingdom', *Journal of the Royal Statistical Society*, 31–60.

Victoria County History (1961) 'Modern York: Economy, 1839-1900', *A History of the County of York: the City of York (Digitised by British History Online)*. Available at: https://www.british-history.ac.uk/vch/yorks/city-of-york/pp269-275 (accessed April 2024).

Wagner, R.A. & Fischer, M.J. (1974) 'The String-to-String Correction Problem', *Journal of the ACM*, 21, 168–173.

Wickham, H. (2023) 'stringr: Simple, Consistent Wrappers for Common String Operations'. Available at: https://github.com/tidyverse/stringr, https://stringr.tidyverse.org (accessed June 2023).

Williams, S., Newton, G. & Satchell, M. (2020) 'Workhouse populations, 1851-1911', UK Data Service. Available at: https://reshare.ukdataservice.ac.uk/853999/ (accessed April 2024).

You, X. (2020) 'Working With Husband? "Occupation's Wife" And Married Women's Employment in the Censuses in England and Wales Between 1851 And 1911', Cambridge University Press.

APPENDIX

Appendix Table 1: Derivation of multipliers, based on population statistics, used to moderate counts of telephone subscriptions per urban settlement (detailed in Section 3.1.2)

Year	Settlement	Multiplier	Population Statistics Used	Relevant Directory
1891	Birmingham	0.921	1901 of: Birmingham; West Bromwich	bt_900524_box62_1891-1892_apr_001
1891	Liverpool	0.865	1901 of: Liverpool; Birkenhead	bt_900524_box62_1891-1892_apr_001
1891	Sheffield	0.883	1901 of: Sheffield; Rotherham	bt_900524_box62_1891-1892_apr_001
1891	Birkenhead	0.135	1901 of: Birkenhead; Liverpool	bt_900524_box62_1891-1892_apr_001
1891	West Bromwich	0.079	1901 of: West Bromwich; Birmingham	bt_900524_box62_1891-1892_apr_001
1901	Birmingham	0.921	1901 of: Birmingham; West Bromwich	bt_900544_box65_1901_jan_001
1901	West Bromwich	0.079	1901 of: West Bromwich; Birmingham	bt_900544_box65_1901_jan_001
1911	Birmingham	0.925	1911 of: Birmingham; West Bromwich	bt_900595_box75_1911_jul_003
1911	Liverpool	0.852	1911 of: Liverpool; Birkenhead	bt_900597_box76_1911_jan_001
1911	Sheffield	0.88	1911 of: Sheffield; Rotherham	bt_900595_box75_1911_jul_003

1911	Birkenhead	0.148	1911 of: Birkenhead; Liverpool	bt_900597_box76_1911_jan_001
1911	Birkenhead	0.148	1911 of: Birkenhead; Liverpool	bt_900599_box76_1911_jul_001
1911	Southport	0.1	1911 of: Southport; Ormskirk	bt_900599_box76_1911_jul_001
1911	Stoke-on-Trent	0.1	1911 of: Stoke-on-Trent; Potteries District	bt_900595_box75_1911_jul_003
1911	West Bromwich	0.075	1911 of: West Bromwich; Birmingham	bt_900595_box75_1911_jul_003
1911	Liverpool	0.852	1911 of: Liverpool; Birkenhead	bt_900599_box76_1911_jul_001
1921	Birmingham	0.926	1921 of: Birmingham; West Bromwich	bt_900611_box78_1921_may_001
1921	Liverpool	0.847	1921 of: Liverpool; Birkenhead	bt_900611_box78_1921_may_001
1921	Birkenhead	0.153	1921 of: Birkenhead; Liverpool	bt_900611_box78_1921_may_001
1921	West Bromwich	0.074	1921 of: West Bromwich; Birmingham	bt_900611_box78_1921_may_001
1931	Birmingham	0.926	1921 of: Birmingham; West Bromwich	bt_900696_box89_1931_apr_001
1931	Bristol	0.846	1921 of: Bristol; Bath	bt_900694_box89_1931_mar_001
1931	Glasgow	0.903	1921 of: Glasgow; Paisley	bt_900700_box90_1931_jan_001
1931	Glasgow	0.903	1921 of: Glasgow; Paisley	bt_900701_box90_1931_jul_001
1931	Liverpool	0.847	1921 of: Liverpool; Birkenhead	bt_900698_box90_1931_jun_001
1931	Liverpool	0.847	1921 of: Liverpool; Birkenhead	bt_900699_box90_1931_dec_001
1931	Manchester	0.67	1921 of: Manchester; Oldham; Rochdale; Stockport	bt_900698_box90_1931_jun_001
1931	Manchester	0.67	1921 of: Manchester; Oldham; Rochdale; Stockport	bt_900699_box90_1931_dec_001
1931	Norwich	0.671	1921 of: Norwich; Cambridge	bt_900695_box89_1931_sep_001
1931	Nottingham	0.799	1921 of: Nottingham; Lincoln; Peterborough	bt_900696_box89_1931_apr_001
1931	Nottingham	0.799	1921 of: Nottingham; Lincoln; Peterborough	bt_900697_box89_1931_oct_001
1931	Portsmouth	0.495	1921 of: Portsmouth; Southampton; Bournemouth	bt_900695_box89_1931_sep_001
1931	Reading	0.618	1921 of: Reading; Oxford	bt_900694_box89_1931_mar_001
1931	Sheffield	0.632	1921 of: Sheffield; Bradford	bt_900696_box89_1931_apr_001
1931	Stoke-on-Trent	0.855	1921 of: Stoke-on-Trent; Chester	bt_900696_box89_1931_apr_001
1931	Stoke-on-Trent	0.855	1921 of: Stoke-on-Trent; Chester	bt_900697_box89_1931_oct_001

1931	Ashton-under-Lyne	0.043	1921 of: Ashton-under-Lyne; Manchester; Oldham; Rochdale	bt_900698_box90_1931_jun_001
1931	Birkenhead	0.153	1921 of: Birkenhead; Liverpool	bt_900698_box90_1931_jun_001
1931	Bradford	0.368	1921 of: Bradford; Sheffield	bt_900696_box89_1931_apr_001
1931	Dewsbury	0.078	1921 of: Dewsbury; York; Halifax; Leeds	bt_900698_box90_1931_jun_001
1931	Halifax	0.143	1921 of: Halifax; York; Dewsbury; Leeds	bt_900698_box90_1931_jun_001
1931	Leeds	0.659	1921 of: Leeds; York; Dewsbury; Halifax	bt_900698_box90_1931_jun_001
1931	Oldham	0.133	1921 of: Oldham; Manchester; Rochdale; Stockport	bt_900698_box90_1931_jun_001
1931	Paisley	0.097	1921 of: Paisley; Glasgow	bt_900700_box90_1931_jan_001
1931	Rochdale	0.083	1921 of: Rochdale; Manchester; Oldham; Stockport	bt_900698_box90_1931_jun_001
1931	Southampton	0.637	1921 of: Southampton; Southampton; Bournemouth	bt_900695_box89_1931_sep_001
1931	Stockport	0.113	1921 of: Stockport; Manchester; Rochdale; Oldham	bt_900698_box90_1931_jun_001
1931	Sunderland	0.366	1921 of: Sunderland; Newcastle upon Tyne	bt_900698_box90_1931_jun_001
1931	West Bromwich	0.074	1921 of: West Bromwich; Birmingham	bt_900696_box89_1931_apr_001
1931	York	0.121	1921 of: York; Dewsbury; Halifax; Leeds	bt_900698_box90_1931_jun_001
1941	Birmingham	0.771	1921 of: Birmingham; Walsall; West Bromwich; Wolverhampton	bt_900786_box104_1941_mar-apr_001
1941	Coventry	0.585	1921 of: Coventry; Northampton	bt_900786_box104_1941_mar-apr_001
1941	Glasgow	0.903	1921 of: Glasgow; Paisley	bt_900789_box105_1941_oct_001
1941	Liverpool	0.847	1921 of: Liverpool; Birkenhead	bt_900788_box104_1941_jul_001
1941	Manchester	0.645	1921 of: Manchester; Ashton-under-Lyne; Oldham; Rochdale; Stockport	bt_900788_box104_1941_jul_001
1941	Nottingham	0.669	1921 of: Nottingham; Derby	bt_900786_box104_1941_mar-apr_001
1941	Preston	0.414	1921 of: Preston; Wigan; Southport	bt_900788_box104_1941_jul_001
1941	Blackburn	0.31	1921 of: Blackburn; Bolton; Burnley	bt_900788_box104_1941_jul_001

1941	Ashton-under-Lyne	0.038	1921 of: Ashton-under-Lyne; Manchester; Oldham; Rochdale; Stockport	bt_900788_box104_1941_jul_001
1941	Birkenhead	0.153	1921 of: Birkenhead; Liverpool	bt_900788_box104_1941_jul_001
1941	Bolton	0.437	1921 of: Bolton; Blackburn; Burnley	bt_900788_box104_1941_jul_001
1941	Bradford	0.743	1921 of: Bradford; Halifax	bt_900787_box104_1941_sep_001
1941	Burnley	0.253	1921 of: Burnley; Bolton; Blackburn	bt_900788_box104_1941_jul_001
1941	Chatham	0.639	1921 of: Chatham; Canterbury	bt_900784_box104_1941_jan_001
1941	Cheltenham	0.485	1921 of: Cheltenham; Gloucester	bt_900785_box104_1941_feb_001
1941	Derby	0.331	1921 of: Derby; Nottingham	bt_900786_box104_1941_mar-apr_001
1941	Dewsbury	0.106	1921 of: Dewsbury; Leeds	bt_900787_box104_1941_sep_001
1941	Greenock	0.1	1921 of: Greenock; "Scotland West"	bt_900789_box105_1941_oct_001
1941	Halifax	0.257	1921 of: Halifax; Bradford	bt_900787_box104_1941_sep_001
1941	Ipswich	0.647	1921 of: Ipswich; Colchester	bt_900784_box104_1941_jan_001
1941	Leeds	0.894	1921 of: Leeds; Dewsbury	bt_900787_box104_1941_sep_001
1941	Newcastle upon Tyne	0.634	1921 of: Newcastle upon Tyne; Sunderland	bt_900787_box104_1941_sep_001
1941	Northampton	0.415	1921 of: Northampton; Coventry	bt_900786_box104_1941_mar-apr_001
1941	Oldham	0.128	1921 of: Oldham; Ashton-under-Lyne; Manchester; Rochdale; Stockport	bt_900788_box104_1941_jul_001
1941	Paisley	0.097	1921 of: Paisley; Glasgow	bt_900789_box105_1941_oct_001
1941	Rochdale	0.08	1921 of: Rochdale; Ashton-under-Lyne; Oldham; Manchester; Stockport	bt_900788_box104_1941_jul_001
1941	Southport	0.27	1921 of: Southport; Preston; Wigan	bt_900788_box104_1941_jul_001
1941	Stockport	0.109	1921 of: Stockport; Ashton-under-Lyne; Oldham; Rochdale; Manchester	bt_900788_box104_1941_jul_001
1941	Sunderland	0.366	1921 of: Sunderland; Newcastle upon Tyne	bt_900787_box104_1941_sep_001
1941	Walsall	0.081	1921 of: Walsall; Birmingham; West Bromwich; Wolverhampton	bt_900786_box104_1941_mar-apr_001
1941	West Bromwich	0.062	1921 of: West Bromwich; Walsall; Birmingham; Wolverhampton	bt_900786_box104_1941_mar-apr_001

1941	Wigan	0.315	1921 of: Wigan; Preston; Southport	bt_900788_box104_1941_jul_001
1941	Wolverhampton	0.086	1921 of: Wolverhampton; Walsall; West Bromwich; Birmingham	bt_900786_box104_1941_mar-apr_001
1951	Birmingham	0.771	1921 of: Birmingham; Walsall; West Bromwich; Wolverhampton	bt_900857_box114_1951_jun_001
1951	Blackburn	0.31	1921 of: Blackburn; Bolton; Burnley	bt_900859_box115_1951_nov_001
1951	Bradford	0.743	1921 of: Bradford; Halifax	bt_900858_box115_1951_dec_001
1951	Coventry	0.585	1921 of: Coventry; Northampton	bt_900857_box114_1951_jun_001
1951	Leeds	0.894	1921 of: Leeds; Dewsbury	bt_900858_box115_1951_dec_001
1951	Liverpool	0.847	1921 of: Liverpool; Birkenhead	bt_900859_box115_1951_nov_001
1951	Manchester	0.645	1921 of: Manchester; Ashton-under-Lyne; Oldham; Rochdale; Stockport	bt_900859_box115_1951_nov_001
1951	Nottingham	0.669	1921 of: Nottingham; Derby	bt_900857_box114_1951_jun_001
1951	Preston	0.414	1921 of: Preston; Southport; Wigan	bt_900859_box115_1951_nov_001
1951	Cardiff	0.684	1921 of: Cardiff; Newport	bt_900855_box114_1951_aug_001
1951	Newcastle upon Tyne	0.634	1921 of: Newcastle upon Tyne; Sunderland	bt_900858_box115_1951_dec_001
1951	Ashton-under-Lyne	0.038	1921 of: Ashton-under-Lyne; Manchester; Oldham; Rochdale; Stockport	bt_900859_box115_1951_nov_001
1951	Birkenhead	0.153	1921 of: Birkenhead; Liverpool	bt_900859_box115_1951_nov_001
1951	Bolton	0.437	1921 of: Bolton; Blackburn; Burnley	bt_900859_box115_1951_nov_001
1951	Burnley	0.253	1921 of: Burnley; Bolton; Blackburn	bt_900859_box115_1951_nov_001
1951	Chatham	0.639	1921 of: Chatham; Canterbury	bt_900854_box114_1951_mar_001
1951	Cheltenham	0.485	1921 of: Cheltenham; Gloucester	bt_900855_box114_1951_aug_001
1951	Derby	0.331	1921 of: Derby; Nottingham	bt_900857_box114_1951_jun_001
1951	Dewsbury	0.106	1921 of: Dewsbury; Leeds	bt_900858_box115_1951_dec_001
1951	Glasgow	0.903	1921 of: Glasgow; Paisley	bt_900853_box114_1950_nov_001
1951	Greenock	0.1	1921 of: Greenock; Scotland West	bt_900853_box114_1950_nov_001
1951	Halifax	0.257	1921 of: Halifax; Bradford	bt_900858_box115_1951_dec_001
1951	Ipswich	0.647	1921 of: Ipswich; Colchester	bt_900854_box114_1951_mar_001
1951	Newport	0.316	1921 of: Newport; Cardiff	bt_900855_box114_1951_aug_001

1951	Northampton	0.415	1921 of: Northampton; Coventry	bt_900857_box114_1951_jun_001
1951	Oldham	0.128	1921 of: Oldham; Ashton-under-Lyne; Manchester; Rochdale; Stockport	bt_900859_box115_1951_nov_001
1951	Paisley	0.097	1921 of: Paisley; Glasgow	bt_900853_box114_1950_nov_001
1951	Rochdale	0.08	1921 of: Rochdale; Ashton-under-Lyne; Oldham; Manchester; Stockport	bt_900859_box115_1951_nov_001
1951	Southport	0.27	1921 of: Southport; Preston; Wigan	bt_900859_box115_1951_nov_001
1951	Stockport	0.109	1921 of: Stockport; Ashton-under-Lyne; Oldham; Rochdale; Manchester	bt_900859_box115_1951_nov_001
1951	Sunderland	0.366	1921 of: Sunderland; Newcastle upon Tyne	bt_900858_box115_1951_dec_001
1951	Walsall	0.081	1921 of: Walsall; Birmingham; West Bromwich; Wolverhampton	bt_900857_box114_1951_jun_001
1951	West Bromwich	0.062	1921 of: West Bromwich; Walsall; Birmingham; Wolverhampton	bt_900857_box114_1951_jun_001
1951	Wigan	0.315	1921 of: Wigan; Preston; Southport	bt_900859_box115_1951_nov_001
1951	Wolverhampton	0.086	1921 of: Wolverhampton; Walsall; West Bromwich; Birmingham	bt_900857_box114_1951_jun_001