Explorations in Economic History 87 (2023) 101476

Contents lists available at ScienceDirect

Explorations in Economic History

journal homepage: www.elsevier.com/locate/eeh

Research Paper

Perks and pitfalls of city directories as a micro-geographic data source☆

Thilo N.H. Albers, Kalle Kappner*

Humboldt-Universität zu Berlin, Germany

ARTICLE INFO

JEL classification:
C8
R1
N9

Keywords:
city directories
data extraction
granular spatial data

ABSTRACT

Historical city directories are rich sources of micro-geographic data. They provide information on the location of households and firms and their occupations and industries, respectively. We develop a generic algorithmic work flow that converts scans of them into geo- and status-referenced household-level data sets. Applying the work flow to our case study, the Berlin 1880 directory, adds idiosyncratic challenges that should make automation less attractive. Yet, employing an administrative benchmark data set on household counts, incomes, and income distributions across more than 200 census tracts, we show that semi-automatic referencing yields results very similar to those from labour-intensive manual referencing. Finally, we discuss how to scale the work flow to other years and cities as well as potential applications in economic history and beyond.

Introduction

Past economic development, urban policies, and shocks continue to affect the present urban geography: The location of chimneys during industrialisation predicts modern-day neighbourhood sorting within UK metropolitan areas, redlining policies implemented in 1930s America have persistent effects on racial segregation, and past shocks such as fires and earthquakes explain the geography of prime locations in contemporary global cities.[1] However, when analysing the persistence and change of urban geography, researchers often face data limitations. Outside of North America and Scandinavia, few countries provide individual-level data sets of full-count historical censuses. Even where such data are available, they are typically not geo-referenced at the address level and may lack the desired frequency.[2] As a second-best, researchers often resort to digitised maps and tabulations that aggregate census (or census-like) material at a lower spatial resolution. In consequence, studies are limited to places where such material had been collected at the required spatio-temporal resolution, preserved, and made available to researchers in digital format.

This paper introduces an alternative micro-geographic data source: city directories. Economic growth and the invention of house numbering spurred their diffusion in the late 18th century, making them virtually ubiquitous by the end of the 19th century as we

☆ We thank Monique Reiske, Kristian Behrens, Nikolaus Wolf, the participants of the 2021 Methodological Advances in the Extraction and Analysis of Historical Data conference, the editor Christian Møller Dahl, and an anonymous referee for comments. Petya Prodanova and Can Aycan provided excellent research assistance. Albers gratefully acknowledges the financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119).

* Corresponding author.
E-mail addresses: alberstn@hu-berlin.de (T.N.H. Albers), kallekappner@googlemail.com (K. Kappner).

[1] See Heblich et al. (2021), Aaronson et al. (2021), and Ahlfeldt et al. (2020).

[2] For example, Berkes et al. (forthcoming) geo-code the full-count US census on a sub-county level, but do not attempt to geo-code it at a within-city level. Knudsen (2021, p. 9) reports that digitised complete-count censuses for the Scandinavian countries are only available for selected years (Denmark: 1845, 1880, 1901; Sweden: 1880, 1890, 1900, 1910; Norway: 1910). See Szołtysek and Gruber (2016) on the lack of other full-count European censuses or the difficulty to obtain them.

https://doi.org/10.1016/j.eeh.2022.101476
Received 15 February 2022; Received in revised form 11 September 2022; Accepted 14 September 2022
Available online 20 September 2022

document in a small meta-study. Early city directories inside and, in particular, outside of the United States often consisted of lists of people ("white pages") and businesses ("yellow pages"). These lists contain spatial information that allow researchers, in principal, to map the social strata and economic configuration of cities at a granular spatial level. Although historians have used city directories as a source throughout the 20th century, resource constraints have so far largely prevented the exploitation of the granularity of their spatial data dimension.[3] To overcome this barrier, we set up an algorithmic work flow that extracts data from scanned directories and converts them into micro-geographic data sets (Albers and Kappner, 2022, accessible via GitHub). We then put it into practice and assess its strength under varying levels of manual labour inputs. This allows us to provide recommendations where such inputs are most productive. The work flow is sufficiently modular to scale it to multiple years and cities as well as to adapt it to the analysis of various listed entities, e.g., firms rather than households.

Our new work flow extracts data from the original directories in four steps: preparation, recognition, structuring, and referencing. For the latter three steps, we highlight a variety of opportunities to improve accuracy. For instance, several optical character recognition (OCR) environments now provide powerful tools by which researchers can re-train existing models on manually generated data specific to their case. The process of structuring OCR output—that is the parsing of recognised text into separate fields for names, occupations, addresses—can be aided through trainable algorithms. Referencing—i.e., assigning a location and occupation-based social status score to a directory entry—can be implemented via semi- or fully-automatic approaches.

We apply the algorithm to a realistic use case, in which census-like individual data were not preserved: Berlin in 1880. The case adds three particular challenges: the source is not in standard script (a challenge for OCR), Berlin's tumultuous history resulted in many changes to the street grid (a challenge for geo-referencing), and existing social status classification systems for historical German occupations are limited in their extent (a challenge for status-referencing). Moreover, using Berlin as a case comes with the advantage of being able to generate a validation data set based on the meticulous work of contemporary Prussian statisticians. In particular, we build a data set on household counts, average incomes, and within-tract income inequality for 200+ municipal census tracts. Aggregating up corresponding measures from our directory data to the tract-level, rank correlations provide a measure of fit to explore two questions. First, how well do city directory data perform in replicating patterns in official tract-level Prussian data given that they may under-report the poorest households? Second, where are manual work hours, e.g. for building training data sets or digitising historical maps, most-efficiently spent?

The under-reporting of poor households appears to only marginally affect the quality of a typical array of economic variables. Rank correlations for household counts vs. directory entries, average income vs. mean social status scores, and income vs. social status score distributions range between .86 and .91. Depending on the precise research questions, the under-reporting bias may still be relevant,[4] but it does not invalidate the use of city directories more generally. With regards to the efficient use of resources, we compare the rank correlations under a set of three different levels of manual labour input for geo-referencing and status-referencing, respectively. Semi-manual referencing approaches achieve results very similar to fully-manual referencing when the variables of interest are aggregated to the tract level. For many research questions, one can thus extract high-quality micro-geographic data from city directories with limited to moderate manual labour inputs.

Besides providing systematic evidence on the validity of city directory as a data source, the second aim of this paper is to provide researchers with a tool that they can scale to other years and cities. As an example, we apply the algorithm to a second cross-section of the Berlin directory and document urban sprawl over time by estimating population density gradients. We also formulate general recommendations for applying the work flow to other cities and discuss the specific example of Boston's 1862 directory.

This paper adds to recent efforts to bridge the gap between computational and social scientists by providing tools that are catered to the latter group's application needs (Abramitzky et al., 2020; Combes et al., 2021; Currie et al., 2020; Dahl et al., 2021b; Gutmann et al., 2018). In particular, we aim to provide urban economists and economic historians with a new tool to exploit the wealth of city directory data that is largely untapped. Heblich and Hanlon (forthcoming, Section 4) provide an excellent survey of the type of studies that can be conducted with historical micro-geographic data and Glaeser (2021) makes the point that, despite all the differences, there is much to learn for cities in developing countries today from looking into the past. Our algorithm equips researchers aiming to exploit within-city variation over time with an adaptable tool for their particular cases.

We are not the very first to highlight the potential usefulness of city directory for social science research. Knights (1969) discusses early applications in economics and Shaw and Coles (1995) make the case for urban history. Naturally, only with recent advances in computing power and recognition software, actual large-scale digitisation has become feasible. This has renewed the interest in using city directories. Two recent collaborations by environmental scientists, computer scientists, and sociologists are closely related to our work. Berenbaum et al. (2019) and Bell et al. (2020) have developed algorithms to extract historic land use patterns and identify potentially hazardous areas of former gas stations, respectively.[5] Compared to their work, our study is focused on individuals rather than firms. It thus incorporates the classification of occupations and individuals into socio-economic status classes. Moreover, we document the precision and value of directory data beyond an expectedly well-documented case such as gas stations.

The remainder of this paper is organised as follows. Section 1 discusses the emergence of city directories, their strength and weaknesses, and their general availability. Section 2 describes the generic workflow of converting these directories into granular

[3] Recent exceptions in economics include Caesmann et al. (2021), Kappner (2022a), and Siodla (2021).

[4] Under-reporting extends to all factors that have been historically correlated with income such as race and gender. It also exists in administrative data (e.g. Chiswick and Robinson, 2021, for gender in US-census data).

[5] In addition, the New York Public Library is involved in a large-scale city directory digitisation project (http://spacetime.nypl.org/) and a digital humanities group appears to work on the large-scale digitisation of the Paris directories (as suggested by this conference report di Lenardo et al., 2019).

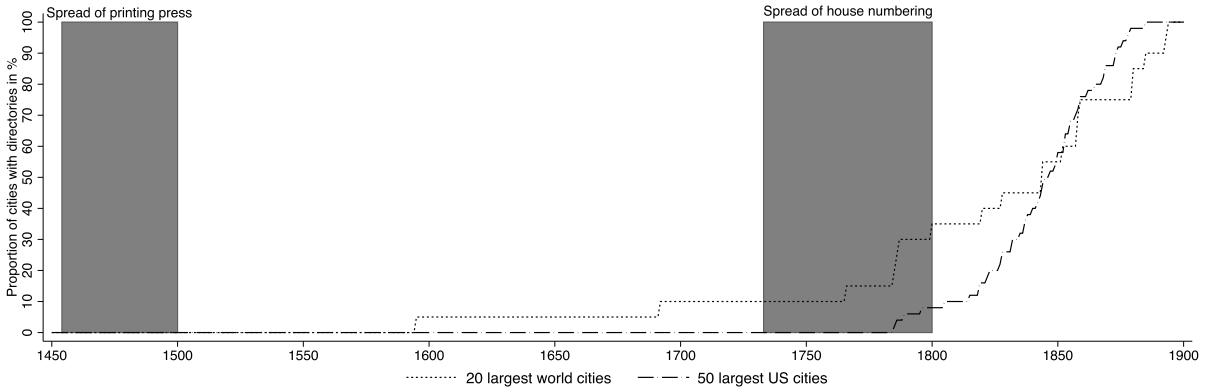


Fig. 1. The spread of city directories. *Note:* The largest cities are defined using the Clio-infra data set on cities and population sizes in 1900 (administered by Buringh, Bosker and Buringh, 2017; Bosker et al., 2013). The data on the first city directories comes largely from Williams (1913) with small hand-collected additions. For dates on the spread of the printing press and house numbering, see text.

spatial data. Section 3 applies the workflow to the Berlin case, assesses trade-offs between precision and manual labour input, and shows how the algorithm can be scaled to other cross-sections and cities. Section 4 discusses potential applications and concludes.

1. City directories as a source

When, how, and why did city directories emerge? What do the answers to these questions imply for their features, strengths, and caveats as a micro-geographic data source?

The spread of city directories. Fig. 1 suggests that the timing of the widespread adoption is itself a tale about the interaction of technology, social innovation, and economic growth. The printing press was invented in 1446 and spread rapidly across Europe over the next 50 years, making the price of books drop significantly compared to the pre-Gutenberg era (Dittmar, 2011). Yet, it took more than two centuries before the first city directory-like book appeared in London (documenting the addresses of the elite), another century until the second one was published in Paris, and yet another hundred years before they became adopted more widely (Williams, 1913). While relatively cheap printing was a necessary condition, it was not sufficient to trigger the emergence of city directories. The growing market economy and a social innovation, house numbering, played an important role.

First, to understand the role of economic growth it is important to note that directories were an innovation coming from the private sector.[6] It thus required substantial private demand for such a product to be marketable. The economic growth that set into motion at the end of the 18th century and beginning of the 19th century in the US and elsewhere correlated with substantial increases in city sizes and inter-urban market integration (for the timing of growth see Gallman and Wallis, 1993). As cross-city exchanges became more common and cities became too large to be known in their entirety to the citizens, directories became a profitable business. Finally, a dynamic economy resulted in the constant movement of firms and individuals. These economic agents aimed to advertise their services and interact with each other (Knights, 1969). This spurred demand for directories, both in the form of "yellow pages" (for businesses) and "white pages" (for individuals).

The second important factor for the emergence of directories was house numbering. Early directories had relied on the description of place names, often with reference to some other locality (Williams, 1913). In the 18th century, however, cities adopted house numbering for a number of different reasons. Prussian cities introduced it to facilitate an easier movement of troops and Madrid did so for tax collection purposes (Rose-Redwood and Tanter, 2012). In Paris and Copenhagen, it was the directory publishers themselves that played a crucial role in introducing house numbering (Rose-Redwood and Tanter 2012, p. 608 and Williams 1913, p. 8). This social innovation greatly improved the usefulness and extent of city directories.

When economic growth and market integration increased and house numbering was more widely adopted, city directories spread rapidly (Fig. 1). The factors that drove their emergence and their widespread adoption shaped the information they included, their strengths, and caveats as a source.

Features, strengths, and caveats. Early city directories around the world often contained two parts: a list of businesses (later dubbed "yellow pages") and a list of the citizens of the city (later dubbed "white pages"). Typically, the "yellow pages" would report firms ordered by industry or product (see, for example, Adams, Sampson, & Company, 1862, for the Boston 1862 directory). The "white pages" would provide information on the home address, name, and occupation of household heads or inhabitants. It is safe to assume that the reporting of industries and occupations originated precisely in the commercial purpose of city directories, with merchants being important customers (see Shaw and Coles 1995, p. 90 and Knights 1969, p. 4). Depending on the country and local context, directories often recorded additional information. In the United States, they sometimes added information on the birth place and

[6] Japan is the case in point. According to Williams (1913, p. 46), its first directory appeared in 1889. Even though it contained the names of those paying direct taxes above a certain level for all larger cities (clearly government data), the publisher appears to be a private entity.

Table 1
Generating socio-economic proxies from city directories: Strengths and caveats

Strengths	Caveats
High frequency	Censoring w.r.t. income, race and gender
Point data	Imprecise occupational titles
Arbitrarily defined spatial units	*Too* precise occupational titles
Income proxies	Little status-differentiation at the top
Distributional information	Structure and content vary across cities
Tracking within-city mobility	
Tracking social mobility	

place of work (Knights, 1969), the latter of which would allow the estimation of mobility matrices for cities. Berlin directories often reported ownership structures, i.e. who owned a given house (see Kappner, 2022a, for more details on these directories). However, even the most basic information allows the exploration of characteristics within a city that would remain unknown in the absence of this source or individual-level census records, the latter of which were not preserved in many countries. Even where census data are available, city directories can be useful by adding data on the locations of firms. Moreover, they were typically published annually and thus at a much higher frequency than censuses were conducted.

The main advantage of city directories, however, derives from the possibility to generate point data where historical census information only survived in aggregated form, e.g. in tract-level cross-tables. Such point data allows the researcher to define their own spatial units independent of historical administrative boundaries. Within these spatial units, names provide information about ethnicity and migration behavior. The count of households in a given district is a good proxy for population density. Occupations can be transformed into social status indicators, for example via HISCO (Historical International Classification of Occupations) and the corresponding social status score system HISCAM (Lambert et al., 2013; Leeuwen et al., 2002) or be combined with wage data. This allows for estimating wage and status distributions. In principle, one could also track individuals through time and space. Directories were kept up-to-date by gathering information from the locals that would then report changes (Knights, 1969). With multiple cross-sections from the frequently published directories, exploring within-city and occupational mobility becomes possible.

The private-sector origin of city directories was important for the inclusion of occupations, but it also leads to important caveats when using them as a source. First, directory data typically under-report low-income households, discriminated groups, and women (Kappner, 2022a; Knights, 1969). Second, occupations recorded in the directories responded to the informational demand of the customers. In consequence, occupational descriptions range from incredibly precise to very imprecise. The Berlin directory provides a good example. Descriptions ranged from very specific such as “the assistant to a lead secretary of the German railway” (*Reichsbahnhauptsekretärgehilfe*) to something as generic as *Rentier*. Third, existing social status scales for occupations may introduce further complications as they offer imperfect differentiation at the top, i.e. many high-status occupations have a very similar score. Table 1 summarises the strengths and caveats of city directories. To gauge the practical relevance of these caveats, we will later compare measures derived from directories with less spatially granular, tract-level administrative data on household counts, income, and income inequality.

Use of directories up until now. At least since the 1930s, local, social, and urban historians and, to some extent, economists have employed city directories as a source (Shaw and Tipper, 2010, p. 1). Knights (1969) discusses some early applications for Philadelphia, Baltimore, and New York, the latter of which were published in the JPE and AER in the 1950s, respectively. They were so popular as a source in the United States that Spear (1961) created an extensive “Bibliography of American directories”. Interest continued outside of the US, where in some cases researchers manually transcribed directories to trace socio-economic development (see e.g. Wiest, 1991, for Munich). Along with a bibliography of British and European directories, Shaw and Coles (1997) and Shaw and Tipper (2010) provide an overview about such applications. Only few recent studies fully exploit the granularity of the spatial data from city directories (see e.g. Caesmann et al., 2021; Kappner, 2022a; Siodla, 2021).

2. Extracting city directory data – the general work flow

How can we transform a given city directory into a geo-referenced data set of socio-economic indicators? This section describes the generic work flow, highlighting the most important steps, obstacles, and tools characterising the extraction process. Its modular structure aims at making adaptions and extensions easier. Complementing this abstract description, Section 3 presents a detailed application to a specific example.

The extraction work flow involves four successive steps (Fig. 2). First, raw scans of the respective sources are obtained and prepared for batch-processing (**i**). Next, the text content and structure of these images are recognised and translated into text strings with positional information (**ii**). Following this, the recognised content is rearranged into a table that appropriately structures the contained information by rows and columns (**iii**). Lastly, the structured information is referenced with respect to space, social status, and other classification systems of interest (**iv**).

While each step of the extraction work flow can be—and traditionally has been—performed manually, there now exists a plethora of tools that support researchers in automating the entire process. An obvious example is step **ii**, where increasingly easy-to-use optical character recognition (OCR) routines can substitute manual transcription efforts, thereby greatly reducing necessary labour input. An integrated, flexible framework for city directory extraction combines these scattered tools, transforming raw scans into

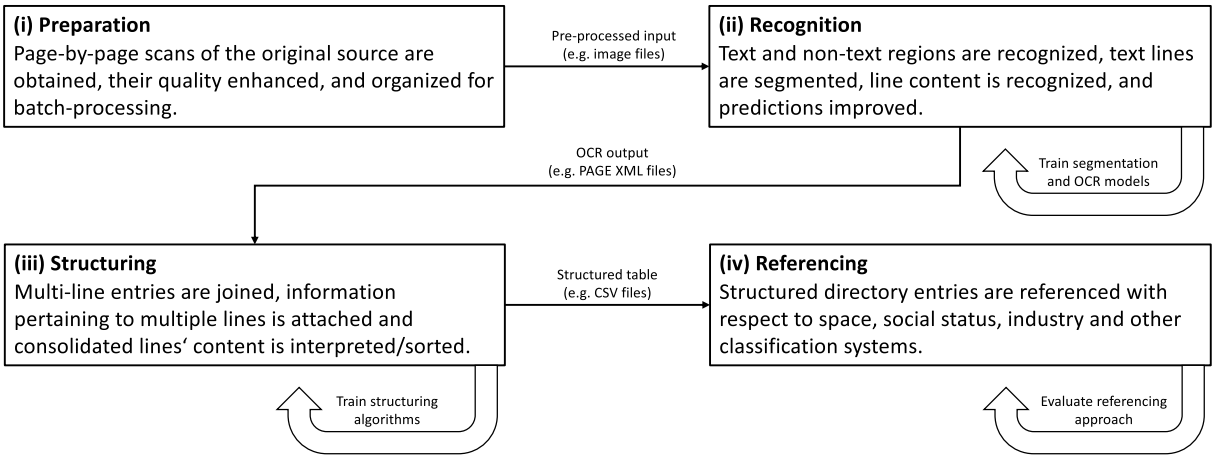


Fig. 2. The generic work flow for city directory extraction

regression-ready data sets in a few steps. Importantly, the serial nature of the extraction process and the possibility to access multiple volumes of a directory series allow to train specialised algorithms along the work flow, as indicated in Fig. 2. In what follows, we highlight the particular challenges at each step and some promising non-commercial tools addressing them.

Preparation (i). The extraction work flow starts with input data in the form of scanned images. Depending on the respective city of interest, researchers may be lucky. In some cases, local libraries, archives, and other public institutions provide high-quality scans of city directories. In other cases, researchers have to scan the directories themselves. Ideally, one should minimise distortions of page position and scan in a high resolution to facilitate the best recognition outcomes in step **ii**. However, even lower-quality scanned material can be enhanced *ex post*, e.g. by deskewing, cropping, binarisation, and other normalisations. Several comprehensive OCR environments like Tesseract (Tesseract contributors, 2021) and OCR4all (Reul et al., 2019) routinely perform these tasks before recognition, providing extensive functionality to adjust them to the particular requirements of the input scans.

Recognition (ii). Once raw scans are appropriately prepared, their content can be recognised via OCR. While different OCR environments apply different prediction algorithms, their common approach is to recognise the spatial structure of a given page by separating non-text and text areas (page or region segmentation), segment the latter into lines (line segmentation), recognise the lines' text content, and possibly enhance the prediction *ex post*. City directories are machine-printed, exert a highly regular layout structure, and contain only few images and non-text structures. They are thus ideal candidates for an OCR-based extraction approach. Importantly, comprehensive software environments—such as the aforementioned OCR4all and Tesseract—equip users with tools that exploit the highly regular structure of city directories to further enhance recognition quality. For instance, existing recognition models can be re-trained on ground truth samples from other volumes of the same series, and they can be conditioned on dictionaries of expected last names, street names and occupations (Dahl et al., 2021a).[7]

Structuring (iii). The output obtained from step **ii** is a collection of text lines with accompanying positional information, e.g. coordinates of their bounding boxes on the original input scan. The goal of step **iii** is to convert this minimally structured content into a database of individuals or firms, suitable for standard querying, sorting, and matching operations. This involves tasks like (1) separating first names, last names, company names, occupations or industries within a given text line, (2) joining multiple text lines referring to the same directory entry, and (3) attaching regularly occurring information related to multiple lines to the respective entries. While conceptually straight-forward, this task is complicated because recognised text lines may exhibit many different formats, both as a result of each entry's particular characteristics, and due to OCR detection errors.[8] A suitable structuring approach thus needs to interpret the format of each line. Solutions to this problem range from manual line-by-line format distinctions, via algorithms that distinguish line formats based on a hard-coded set of regular expressions (e.g. Berenbaum et al. 2019 and Bell et al. 2020), to machine learning-based, trainable parsing algorithms (e.g. Spaan and Balogh, 2021). The latter are particularly promising, as hand-corrected structuring examples serve to gradually enhance the quality of parsing algorithms. Source-specific solutions are likely to perform best. In the past, adapting existing structuring algorithms to a specific case has been demanding in terms of coding skills. However, emerging OCR output structuring tools such as LayoutParser greatly simplify this process (Shen et al., 2021).

[7] These two OCR environments also serve to illustrate the range of user interface approaches. OCR4all features a complete graphical user interface, allowing users to control the whole OCR process without any scripting language or command line inputs. In contrast, Tesseract is geared towards use via common line or application programming interfaces, allowing to embed its capabilities in larger work flows.

[8] For example, one entry might read "*family name, first name, middle name, occupation.*" because this individual has a middle name. Another entry might read "*family name. first name, primary occupation, secondary occupation.*", because this individual has two occupations and the OCR algorithm mistook a comma for a full stop.

Referencing (iv). In the final step of our work flow, each entry of the structured database is referenced with respect to space and other attributes of interest, such as an occupation or industry classification scheme. While contemporary addresses are easily converted into coordinates through various geo-coding packages (e.g. Cambon et al., 2021 for R and Geopy contributers, 2021 for Python), their usefulness for historical city directories depends on the continuity of the urban street grid, street names, and house-numbering practices.[9] Alternatively, historically more sensitive geo-referencing approaches match recognised addresses to spatial information that is extracted manually or automatically from historical maps (Cura et al., 2018; Schlegel, 2021). Next to locating directory entries in space, researchers will often want to classify them with respect to non-spatial dimensions, e.g. social status – what we call status-referencing. At this step, the potential for automation depends on the specific reference system. As a baseline, we recommend matching recognised entries to the Historical International Standard of Classification of Occupations (HISCO) and its extensive list of historical occupations for various languages (Leeuwen et al., 2002).

3. Application, validation, and effort-precision trade-offs

We now employ the generic work flow to a particular case: the 1880 Berlin city directory. Berlin represents a particularly well-suited case to explore the strength of our approach. It adds additional challenges to the processing of the directories that are likely to occur in many (non-US) contexts (non-standard script, changes in the street grid, scarce occupational reference data). Moreover, municipal statisticians recorded aggregate data about households, mean incomes, and within-district income brackets for 200+ census tracts covering our whole study area. Section 3.1 describes our case study, following the four-step process laid out in the previous section. Section 3.2 levies the aforementioned administrative data to carry out extensive validation exercises and to assess where manual labour inputs are most-efficiently spent. Section 3.3 applies the algorithm to an additional cross-section, highlighting and validating the potential to track changes over time. Finally, we briefly discuss the scaling of the work flow to other cities. Code and replication files are accessible in our GitHub repository (Albers and Kappner, 2022).

3.1. Applying the work flow: The case of 19th century Berlin

Berlin's 1880 city directory lists the names and occupations of all "economically self-sufficient inhabitants, excluding journeymen and day laborers" by street and house number, spanning 408 pages with 6 columns each and approximately 100 rows per column (Ludwig, 1880). Exemplifying a typical use case, our application exercise aims to produce a data set of household heads containing information on their location and position on the occupation-based HISCAM social status scale.

Preparation and recognition (i-ii). We obtain raw scans of Berlin's 1880 city directory from a local library (*Zentral- und Landesbibliothek Berlin*). For preparation and recognition, we rely on OCR4all, applying its standard input scan optimisation, region and line segmentation routines, as well as the embedded Calamari OCR engine (Reul et al., 2019; Wick et al., 2020). Our raw data is typeset in Fraktur, a German blackletter script (see Fig. 3a). As Calamari's default recognition model for this script type yields unsatisfactory results in our case, we use OCR4all's ground truth production module to re-train the model. We employ hand-collected ground truth samples from the first 50 pages of the city directory's *1875* volume (Ludwig, 1875). This generates a highly specialised recognition model that can be used for every other volume of the Berlin directory series. Additionally, we correct a small number of line segmentation errors using the graphical user interface of LAREX, OCR4all's segmentation module (Reul et al., 2017). With a character error rate below .75%, the resulting model achieves a high level of recognition accuracy when tested on our ground truth sample.[10] The two upper panels of Fig. 3 show an excerpt of the raw input (left) and the corresponding OCR output (right).

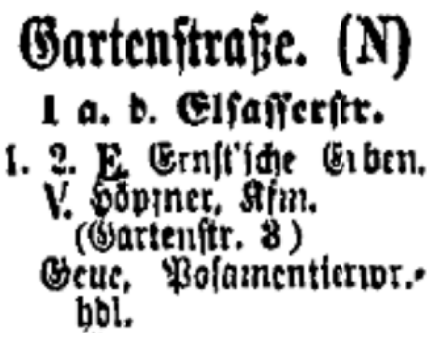
Structuring (iii). As Berlin's city directory exhibits a highly regular layout, we apply a structuring algorithm that transforms the OCR output through a set of case-distinctions based on regular expressions. Consider the variation in font size, relative position to the centre of the column, and indentation in the raw data (Fig. 3a): We exploit the bounding box coordinates of recognised text lines, measured in pixels relative to the origin. For instance, lines, for which the bounding box height surpasses a given threshold value, contain street names, i.e. line `[1]` in the corresponding line-by-line representation of OCR output (Fig. 3b). Similarly, lines whose bounding box' leftmost coordinate are sufficiently shifted to the left of the respective column's centre contain house numbers, e.g. line `[3]`. Tagging these lines as "location lines" and exploiting the directory's ordering of names by address, we attribute the corresponding address to the following lines. Next, we identify indented lines, which continue their preceding lines' content. Their bounding box' leftmost coordinate is sufficiently shifted to the right relative to the preceding line to identify them, e.g. lines `[5]` and `[7]`. Merging accordingly, we arrive at a set of consolidated lines.

The second structuring element is the content *within* each consolidated line, represented in text format in Figure. For example, non-location lines generally contain the last name followed by the occupation, separated by a comma, as illustrated in lines `[4]` and `[6]`. Additionally, they may contain a leading `E.` or `V.` as in `[3]` and `[4]`, and a trailing parenthesised address as in `[5]`. These

[9] Additional limitations arise because few non-commercial geo-coding services allow users to store queried information, and most apply rate limits that undermine their use in mass-querying observations. Furthermore, some geo-coding application interfaces are sensitive to spelling errors resulting from the OCR process.

[10] While generating the ground truth sample bore significant initial investment costs (of around 5 full working days of an experienced Fraktur-reader), the improved OCR model forms the building block for the subsequent processing of directories with a similar typeface. Outside the German-speaking world, city directories were usually printed in modern typefaces and are thus more easily recognised by conventional OCR environments.

a) Raw input



b) OCR output

```
[1]  Gartenstraße.  (N)
[2]  1 a.  d.  Elsasserstr.
[3]  1.  2.  E. Ernst'sche Erben.
[4]  V. Höpfner, Kfm.
[5]  (Gartenstr.  3)
[6]  Geue, Posamentierwr.-
[7]  hdl.
```

c) Structured data

street	number	last name	proprietor status	absentee	absentee address	occupation
Gartenstraße. (N)	1. 2.	Ernst'sche Erben	E.	0		Rentier
Gartenstraße. (N)	1. 2.	Höpfner	V.	1	Gartenstr. 3	Kfm.
Gartenstraße. (N)	1. 2.	Geue		0		Posamentierwr.hdl

d) Referenced data

lat	lon	lot id	inh id	HISCO	HISCAM
52.528	13.394	1	1	-1	99
52.528	13.394	1	2	41020	81.49
52.528	13.394	1	3	41030	59.25

Fig. 3. Berlin Application. *Note:* Stylised example of the output at each step of the city directory extraction work flow. The top left panel is an excerpt of Berlin's 1880 city directory, showing the first few lines of the address Gartenstraße 1/2. The top right panel shows the corresponding string representation obtained through OCR, line by line. The middle panel shows the processed entries, where associated lines are joined, and various information contained in those lines is sorted into separate columns. The lower panel shows the referenced entries, with occupational titles converted into HISCO codes and their associated HISCAM scores. For clarity, this example illustrates the case of an error-free OCR recognition.

indicate (co-)proprietor status and an absentee address, respectively. Using a combination of hard-coded pattern rules and a trainable parsing algorithm, as well as dropping irrelevant information as in [2][11], we arrive at a structured data set of households (Fig. 3c).

Referencing (iv). The final step of the work flow converts the textual information in the structured data to geo-referenced observations with a HISCO code and a corresponding HISCAM score (Fig. 3d). For our validation exercise, we apply three approaches for geo-referencing and status-referencing respectively. Fig. 4 maps them schematically according to required levels of manual effort.

To gauge the importance of precise geo-referencing, we pick three approaches requiring zero, medium, and substantial manual labour inputs. First, we use an online geo-coding application programming interface that matches each address to the most likely candidate among contemporary locations within Berlin (○ symbol in Fig. 4). Second, to increase precision with respect to changes in the street grid, street names and house numbering, we use a historical cadastre map of Berlin to create a shapefile representing each historical street as a linestring. For each street, we identify four house numbers: Those located at the two opposite sides of the beginning of the street, and those located at the two sides of the end of the street. We then interpolate the position of all other integer-valued house numbers at equi-distant segments along the street linestring and match each address in our structured data set to the most similar address in the interpolated shapefile (◇ symbol in Fig. 4). Third, we use a hand-made shapefile containing the exact position of each historical address (△ symbol in Fig. 4).

To reference directory entries on the HISCAM scale, we also pick three approaches requiring varying levels of manual labour efforts. First, we find the best match for each occupation within the list of all occupations currently listed on the HISCO website (⠿

[11] Line [2] describes the street corner at which the house is located. Such information was helpful to visitors unfamiliar with a city, but it is difficult to exploit it for geo-referencing purposes.

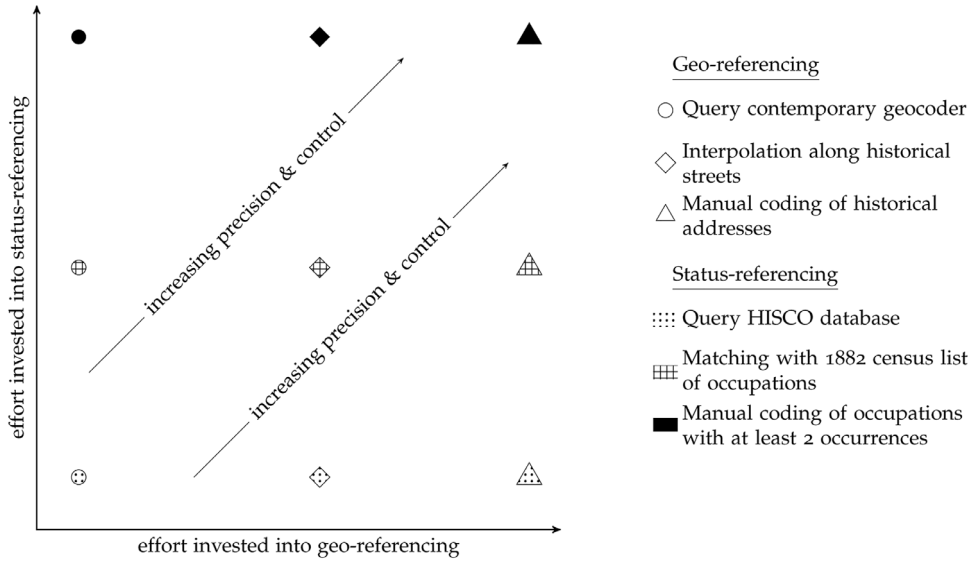


Fig. 4. Effort-precision trade-offs

pattern in Fig. 4).[12] Second, we transcribe a list of 6489 German professions listed in the 1882 Prussian occupational census and assign representative HISCO codes based on their belonging to one of 153 administratively-defined "occupational groups". For the occupations in our structured data set, we then find the best matches among the census-listed occupations (pattern in Fig. 4). Third, we hand-pick an appropriate HISCO code for each occupation occurring at least two times in our structured data set (pattern in Fig. 4). Finally, for all referencing approaches, we convert HISCO codes to HISCAM social status scores (Lambert et al., 2013).[13]

The alternative ways to reference observations in space and on the HISCO status scale require different levels of manual effort, but they also differ with regards to the expected precision and control over the process. The schematic Fig. 4 shows how precision and control increase when moving towards North-East in effort levels. In the next section, we confront actual data with the precision-effort trade-offs.

3.2. Validation and effort-precision trade-offs

How practically relevant are the under-reporting of poor households and the imprecision of occupation-based status scales for typical measures of interest to economic historians and urban economist? Which combination of manual labour and automation is most efficient? To explore these questions, we validate our data set of referenced household heads by comparing various summary statistics of their social status distribution with conceptually related measures derived from reliable, administrative sources. In particular, we group household heads by 216 contiguous census tracts covering all of Berlin. For each tract, we count the number of successfully referenced household heads and compute their average HISCAM value. In addition, we also compute within-tract distributional measures in the form of top-p% status shares: These report the share that the top-status percentile p owns in the sum of HISCAM values of all individuals (i.e. the "HISCAM mass" of the tract). These statistics represent proxies for typical measures of interest in historical urban economics. We then create a corresponding data set of household counts, average income, and income inequality measures from administrative reports, all referring exactly to the year 1880.[14] Importantly, we perform the validation exercise for all possible combinations of the three geo-referencing (GR) and status-referencing (SR) approaches discussed above. This makes the effort-precision trade-offs visible, which researchers face when working with directories.

Household counts. Our first validation focuses on the mass of referenced observations. The central concern is that directory-derived population proxies systematically under-estimate actual population, both because of censoring in the source and deficiencies in the referencing process. While such under-reporting is not a concern if its proportional extent is constant across space, it can be expected to spatially vary in many setups. In turn, directory-derived proxies may compress or inflate population for certain tracts and, hence, mis-represent the population distribution. Fig. 5 allows us to gauge the extent of this problem in our application. For each referencing

[12] Some occupations are listed more than once in the HISCO database. If there are multiple matches, we pick the first of the most frequent HISCO codes. As of August 2022, the HISCO database contained 36,909 entries, among them 1306 entries in German.

[13] See the Web Appendix A.1 for sources.

[14] Importantly—and quite representative for the non-US context—these reports do not contain household-level data (such as the fully-digitised US census) but summarise the respective data for 216 tracts within Berlin. See the Web Appendix A.1 for more details on the sources and procedure.

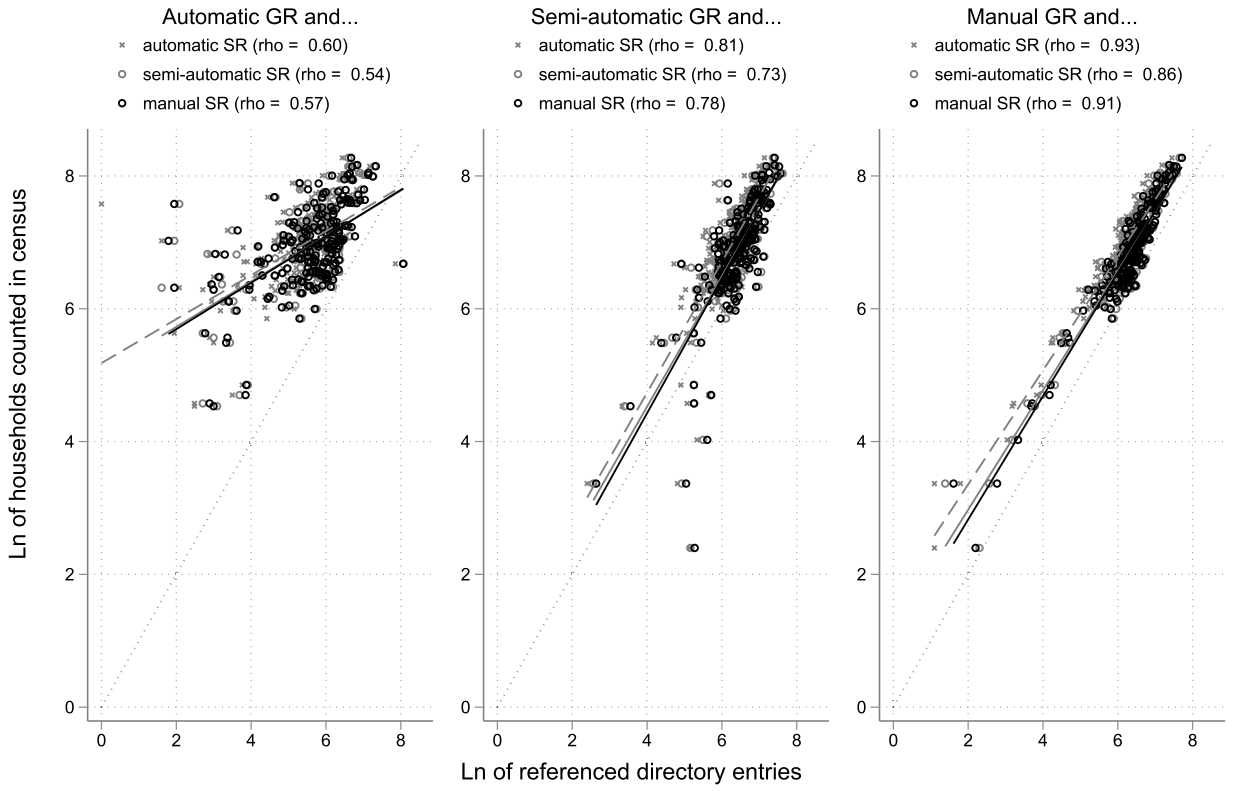


Fig. 5. Census households counts vs. matched HISCAM values. *Note:* Each panel shows a plot of the tract-level (logged) count of households reported in Berlin's 1880 census against the (logged) count of referenced household heads in our city directory data set. The number of tracts is 203 for automatic geo-referencing and 216 for all other graphs/correlations. This difference emerges as automatic referencing sometimes leads to zero households in a tract. The three panels represent different geo-referencing (GR) approaches, while we distinguish by status-referencing (SR) approach within each panel. The dotted 45° lines represent equal counts in both sets, while the other straight lines denote best linear fits. The coefficients reported in the legend refer to the Spearman rank correlation obtained for a specific GR-SR combination. See Section 3.1 for further details on the GR and SR approaches and Web Appendix A.1 for details on the sources.

approach, it plots the (logged) count of referenced observations in our data set against the (logged) count of household heads reported in Berlin's 1880 census, additionally reporting the respective Spearman rank correlation coefficient.

Four central results emerge from Fig. 5. First, while our directory-based data set generally under-counts households, semi-automatic and manual GR approaches reduce the error by a considerable margin.[15] Second, the rank order of observations with respect to the number of households is successfully reproduced by the automatic GR approach, but semi-automatic or manual GR approaches yield substantially higher precision. Third, while the automatic and semi-automatic GR approaches produce a number of outliers (i.e. cases of severe under- or over-reporting of poor households), these are avoided with the manual GR approach. Fourth, both with respect to under-reporting and reproducing the rank order of tracts, automatic, semi-automatic and manual SR techniques perform similarly.

Once population counts (or densities) are used, fully automatic referencing approaches yield satisfactory results. We recommend investing into the GR approach if there are good reasons to expect spatial variation in the extent of under-counting, as in our case.[16]

[15] The 1880 census reports 255,929 resident households in Berlin. While the 1880 Berlin city directory does not report the total of listed household heads, we estimate them to be roughly 195,000, leading to an approximate difference of 25%. Not all of this difference is explained by under-reporting of poor households as official definitions of households and those recorded in the address books likely diverged. Extracting the city directory data, our fully automatic referencing approach yields 62,148 observations, the fully manual referencing approach yields 149,083 observations, with the semi-automatic approach in between (145,929 observations). With the most precise approach, we thus miss around 24% of the city directory-listed household heads. This reflects a) OCR recognition errors that make referencing impossible, b) historical addresses that we could not locate anymore, and c) unique occupations that we did not code into HISCO.

[16] Berlin experienced extensive renaming and renumbering of historical streets in a spatially non-random fashion after World War II. Until 1929, Berlin's streets followed a clockwise or anti-clockwise "horseshoe" numbering system ("Hufeisensystem"). After 1929, new streets had to apply the more common "zig-zag" (or chronological) system with even numbers on one side of the street and odd numbers on the other side. Existing streets

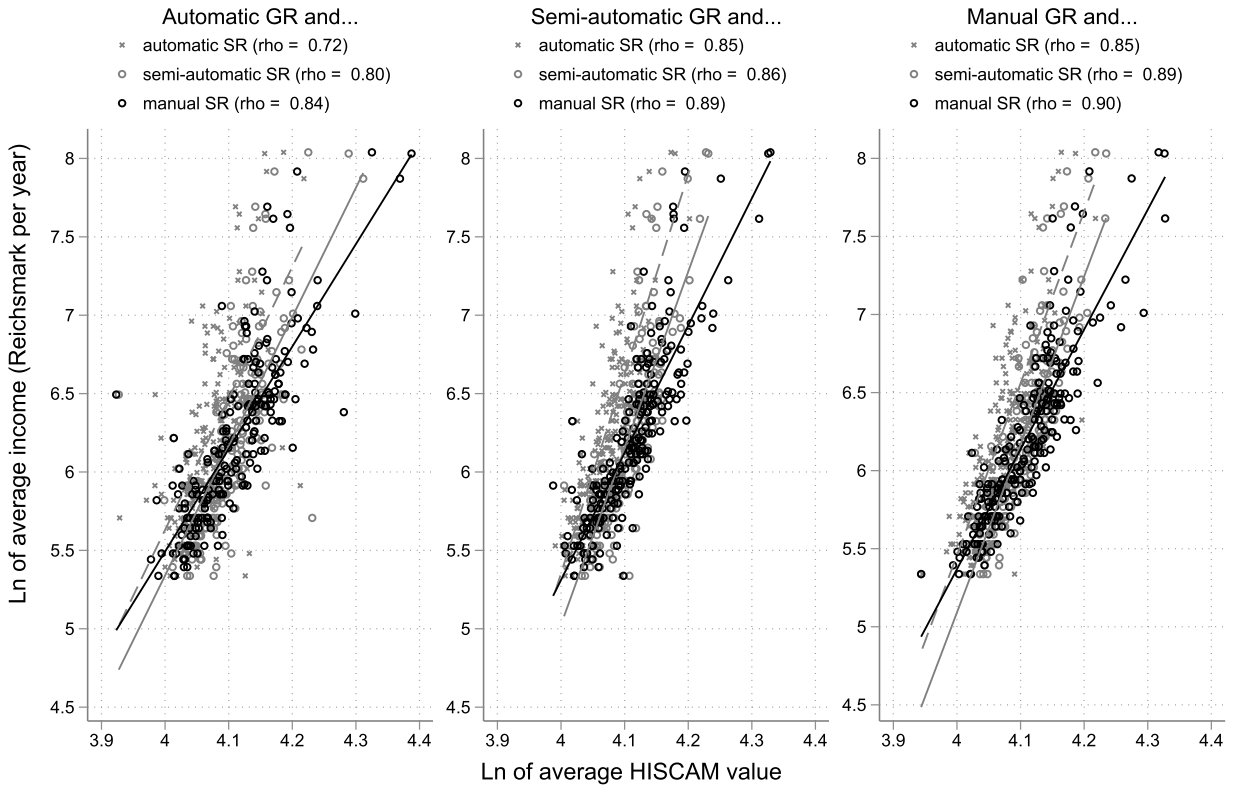


Fig. 6. Mean incomes vs. mean HISCAM values. *Note:* Each panel shows a plot of tract-level (logged) average income against the (logged) average HISCAM score in our city directory data set. The number of tracts is 203 for automatic geo-referencing and 212/213 for all other graphs/correlations. The three panels represent different geo-referencing (GR) approaches, while we distinguish by status-referencing (SR) approach within each panel. The straight lines denote best linear fits. The coefficients reported in the legend refer to the Spearman rank correlation obtained for a specific GR-SR combination. See Section 3.1 for further details on the GR and SR approaches and Web Appendix A.1 for details on the sources.

However, while researchers can expect highly accurate population estimates from manual GR approaches, a semi-automatic GR approach, coupled with an automatic SR approach, will suffice in many scenarios.[17]

Average incomes. Our second validation exercise shifts the focus to the distribution of HISCAM scores across urban space. A typical use case for directory-derived geo-referenced HISCAM scores is the reconstruction of arbitrarily disaggregated social status gradients for historical cities. However, the under-representation of poor households in many directories raises concerns about the accuracy of such estimates. To assess whether our directory-based data set reliably captures status variation across space, we compare (logged) average HISCAM scores to administrative (logged) per-capita income estimates at the tract-level. Fig. 6 plots both measures and reports rank correlation coefficients, separately for each referencing approach. Clearly, our benchmark administrative income measures comprise both wage and non-wage incomes. They are thus conceptually distinct from HISCAM status measures as i) HISCAM scores are an ordinal concept, ii) the differentiation of high-status occupations in terms of HISCAM scores does not necessarily correspond to wage differences, and iii) they do not include capital income (see also Table 1).

Notwithstanding these conceptual caveats, Fig. 6 suggests that the bottom-censored nature of Berlin's 1880 city directory has no discernible impact on the derived average status measures. Average HISCAM scores are excellent proxies for per-capita income. In particular, they accurately reproduce the rank order of tracts with respect to per-capita income, as judged by the invariably high correlation coefficients. While a fully automatic referencing approach already yields decent results, semi-automatic GR and SR approaches substantially raise precision and decrease the incidence of extreme outliers. In contrast, additional gains from thoroughly manual referencing approaches are marginal.

How much effort should be invested into referencing? Our example suggests that there are considerable gains from following a semi-automatic GR approach, possibly coupled with a semi-automatic SR approach. In contrast, there is little to be gained from fully manual referencing approaches. When in doubt, researchers interested in income variation between spatial units should invest

only had to switch to the new system when they were extended, shortened or experienced other significant changes. World War II destruction, denazification and the Berlin Wall provided ample, but spatially non-random, opportunity to apply this law to existing streets.

[17] Additionally, manual GR approaches will be important in the rare situation where researchers are interested in absolute population estimates and cannot inflate directory-based estimates using city-level population totals.

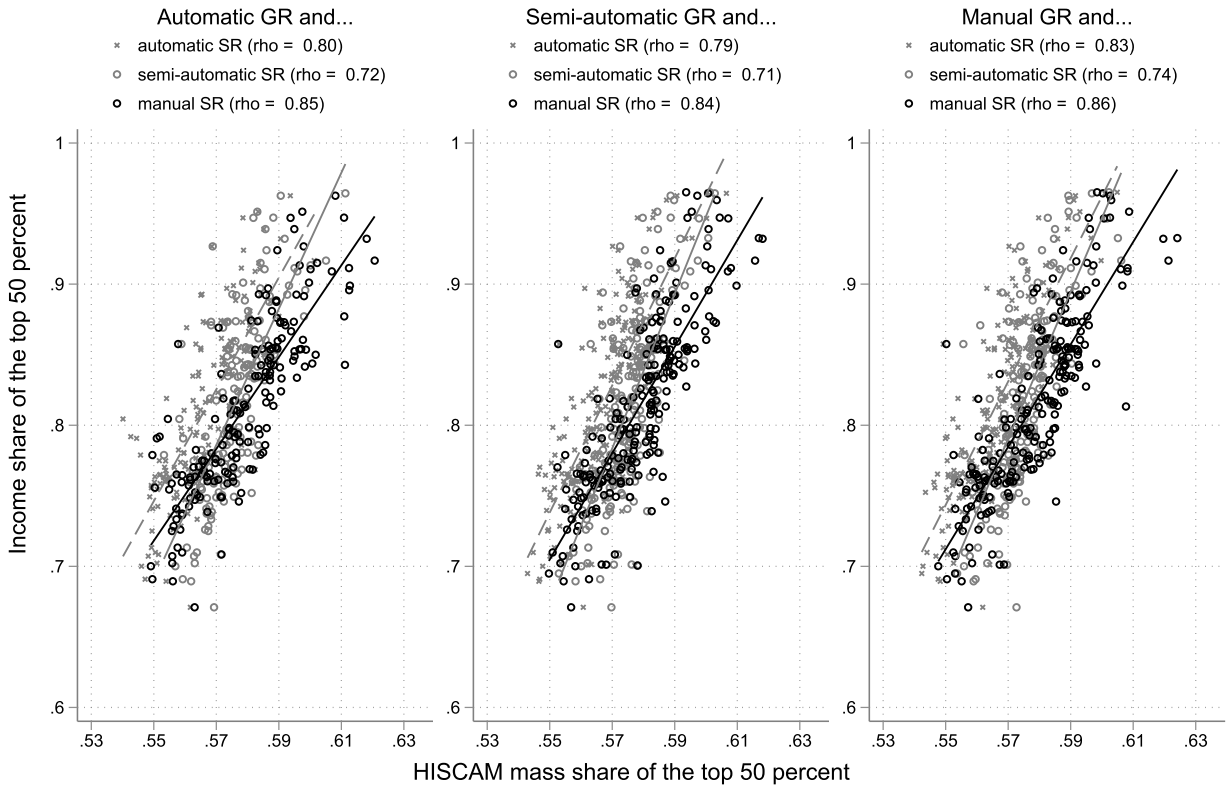


Fig. 7. Top income shares vs. top "HISCAM mass" shares. *Note:* Each panel shows a plot of the tract-level income share of the top 50% against the top 50 %'s share in the "HISCAM mass" in our city directory data set. "HISCAM mass" is the sum of all HISCAM points within a tract and the corresponding top-50 shares report the share of the top half of households (by status) in this mass. The number of tracts is between 168 and 211 depending on the combination of geo- and status-referencing. The exact number of observation varies as we drop those tracts with less than 100 referenced households. The three panels represent different geo-referencing (GR) approaches, while we distinguish by status-referencing (SR) approach within each panel. The straight lines denote best linear fits. The coefficients reported in the legend refer to the Spearman rank correlation obtained for a specific GR-SR combination. See Section 3.1 for further details on the GR and SR approaches and the Web Appendices A.1 and A.2 for details on the sources and the construction of the top-50 % income shares.

resources in improving spatial precision rather than status-referencing. We expect this result to emerge even stronger for cities located in countries with a larger coverage of occupations in the HISCO database.

Top income shares. While the counts and mean incomes focus on the variation *between* census tracts, the possibility of generating distributional measures relying on *within-tract* variation is a particular feature of granular spatial data from city directories. Berlin's city administrators published data that allow us to calculate top-income shares from the 30th percentile upwards employing the generalised Pareto-interpolation (see Blanchet et al., forthcoming, for the method and Web Appendices A.1 and A.2 for more details). Likewise we compute top-shares for the "HISCAM mass". As discussed above, HISCAM scores and incomes are different concepts. One might be particularly worried that HISCAM scores are unable to capture distributional attributes given the importance of high capital incomes at the top of income distributions. Correlating the income share of the top-50% with the respective HISCAM mass of the top-50%, Fig. 7 suggests that these problems do not play an important role for distributional measures that focus on other parts than the top.

Unlike for the previous exercises, different GR approaches do not seem to affect the fit. This is likely due to the focus on *within-tract* rather than *between-tract* variation. In contrast, the results for the different SR approaches indicate the sensitivity of HISCAM distributions to censoring. As discussed above, the semi-automatic status referencing relies on a Prussian census list. This list was heavily biased towards the inclusion of civil servants (*Beamte*), which would typically earn more than ordinary labourers. Compared to the automatic and manual approaches, semi-automatic status-referencing adds more households at the top relative to the bottom. It thus decreases the fit with the administrative income distributions.

Fig. 8 shows the correlation of top-income and top-status shares between the 30th and 99th percentile. From the 30th to the 70th percentile, the HISCAM and income distribution correlate well. When going beyond this level, however, the correlation between HISCAM and income top shares breaks down entirely. That this is true irrespective of the precise status-referencing approach employed precludes precision as an explanation. The underlying reason is the conceptual difference between the measures. Because status differentiation is difficult at the top, HISCAM scores provide relatively little nuance at this part of the distribution: a doctor and

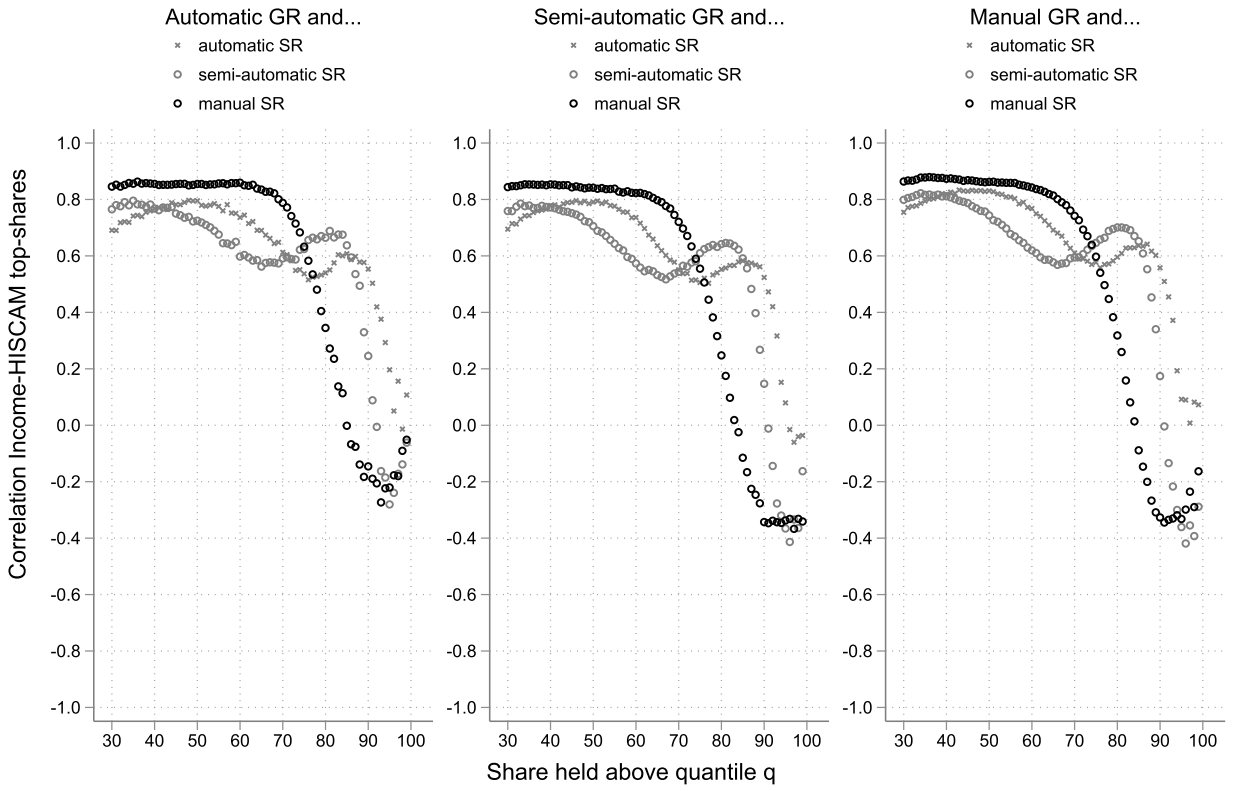


Fig. 8. Correlation of top-income and top-status shares by quantile. *Note:* See text and Web Appendix A.2 for the construction of the top income shares. To obtain sensible comparisons for top-shares, we drop tracts from the sample for which less than 100 households could be referenced automatically.

a bank director both score very highly. It is, however, likely that their salaries, let alone capital incomes, differ. As long as the distributional measure of interest pertains to a larger part of the distribution rather than the nuances among the top-30%, little practical concerns arise when using HISCAM scores for distributional analyses.

Summary: Validation & trade-offs in resource allocation. The preceding exercises benchmark directory-based estimates of population counts, average status and top-shares against their respective administratively measured counterparts. The results suggest that directory data are of high-enough quality to be used as a substitute in the absence of spatially-disaggregated individual-level census data. The under-reporting of low-income groups does not affect the comparison across municipal tracts in a substantial manner. With respect to the imprecision of the HISCAM scores at the top of the distribution, we recommend using distributional measures that focus on the status of the upper 30% relative to the rest (but not beyond that percentile). Top-shares are an obvious choice, not least because the researcher can assess the sensitivity of results to changing the measure along the distribution.

In terms of trade-offs in resource allocation, we recommend the following 3-step procedure: In the first step, assess the quality of text recognition. A simple way to do this is to calculate the number of recognised entries over all entries. As discussed in our particular example, train the algorithm with ground truth data if considered necessary. Akin to the recommendations by Bailey et al. (2020, in the context of record linkage), consider to implement multiple OCR models and identify common matches to reduce Type-I errors.

In a second step, assess the likely trade-offs in geo-referencing. *Ex ante*, assess whether street names in historic and modern maps of the city largely overlap (as is likely in countries with a less tumultuous history than Germany). If the overlap is high, do not invest time in geo-referencing. In cases in which eyeballing suggests that a lot of street names were changed, resort to semi-automatic geo-referencing as described above. *Ex post*, assess the quality of the matching using population data at the level of wards or similar. These data can typically be found in other historical sources such as reports of the city council, newspapers or statistical yearbooks. Sometimes they are even reported in directories themselves (see e.g. Web Appendix A.3 for Boston). In cases, where such data at the ward level do not exist at all for the year of interest, resort to surrounding years. If a particular area has a very low match rate—for example, because street names were changed—consider manual geo-referencing for a subset of the data.

In step three, assess the corpus for status-referencing. Is the corpus too small? Does improving the corpus increase representation at both ends of the distribution, thus avoiding biasing distributional estimates into a certain direction? Finally, would manual status-referencing increase the quality of the measure of interest? There are few application at the *house*-level for which this is likely, for example to assess the effect of mixed-income housing on health outcomes (Kappner, 2022a). However, automatic and semi-automatic status referencing will yield satisfying results for most applications.

Table 2
The change of urban density gradients

	Dependent variable: Population in cell							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Gradient in 1875 (cross-section)				Change in Gradient, 1875 & 1880			
	Georeferenced directory			Census	Georeferenced directory			Census
	A	SA	M		A	SA	M	
ln(Dist)	-0.88***	-0.88***	-0.87***	-0.71***	-	-	-	-
	(0.04)	(0.04)	(0.04)	(0.03)				
ln(Dist) $\times t = 1880$					0.23***	0.25***	0.27***	0.27***
					(0.02)	(0.02)	(0.02)	(0.02)
Cell fixed effects	-	-	-	-	✓	✓	✓	✓
Time fixed effect	-	-	-	-	✓	✓	✓	✓
N (cells)	5135	5135	5135	5135	5135	5135	5135	5135
N (observations)	5135	5135	5135	5135	10270	10270	10270	10270
Pseudo-R^2	0.05	0.07	0.08	0.04	0.87	0.85	0.85	0.87

Note: The table shows the estimated coefficients from $HH_c = \alpha + \beta \ln(Dist_c) + \epsilon_c$ (columns 1–4) and $HH_{ct} = \gamma_c + t_t + \zeta \ln(Dist_c) \times t_t + \epsilon_{ct}$ (columns 5–8) for cell c's household count HH and distance to the city center Dist in year t. All regressions are estimated using the PPML estimator (and results are robust against the separation problem, see Correia et al., 2020). Standard errors are clustered at the cell-level.

3.3. Scaling the algorithm across space and time

The validation exercise in Section 3.2 highlights that our algorithm accurately reproduces density, income, and socio-economic inequality patterns in cross-sectional settings. However, the great potential of the technology lies in scaling it to other years and cities, allowing to create panel data for arbitrary spatial units or even individuals.

Scaling to multiple cross-sections. In the process of scaling the work flow to other years for the same city (i.e. other volumes of the same directory series), three main challenges arise: (i) The directory's layout changes (e.g. from lists ordered by address to lists ordered by name, or from a two-column to a four-column layout); (ii) city administrators modify street names and house numbering systems (e.g. from "horseshoe numbering" to chronological numbering); (iii) the meaning and status of occupations, industries and products evolves.[18] While these changes pose a serious challenge when researchers aim to create panel data from directories over multiple decades or even centuries, they are unlikely to hamper data creation within shorter time horizons—which form the majority of use cases.

To demonstrate the potential that lies in scaling, we apply our work flow to recognise and reference the 1875 volume of the Berlin directory (Ludwig, 1875). In generating this cross-section, we use the same OCR model, structuring techniques and geo-referencing data as for the 1880 volume. We then employ our two cross-sections to explore urban sprawl between 1875 and 1880, following the common practice to calculate population density gradients relative to the central business district (see McDonald, 1989, for an extensive discussion). We partition our study area into grid cells with a size of around 12,000 square meters. For these cells, we calculate the number of households identified by each of our three geo-referencing approaches in each year (1875 & 1880). In addition, we derive the "true" household count for each grid cell by matching building-level data documented in the 1875 and 1880 census reports (Böckh, 1878, 1883). Rather than just demonstrating that our work flow is easily scaled to generate additional cross-sections, the comparison with census-derived counts allows us to judge the reliability of our algorithm in capturing changes over time for fixed spatial units.

The first four columns of Table 2 report the cross-sectional estimates of the population density gradient in 1875. The three geo-referencing approaches (A: automatic; SA: semi-automatic; M: manual) yield virtually identical results. The gradient based on official census counts (Column 4) is flatter, indicating that our directory data under-samples in the urban outskirts. Columns 5–8 show the change of the gradient between 1875 and 1880. The positive coefficients suggest that the gradient has become flatter over time, i.e. that the city sprawled.[19] Strikingly, the estimates based on manually-referenced directory data and census data are the same. Moreover, even the estimate based on automatically geo-referenced directory data leads to virtually the same conclusion concerning the degree of urban sprawl.

[18] The Berlin directory series, published since 1704, is a case in point for these problems. Over the past 300 years, new occupations emerged, the directory's layout structure and typeface was changed and many streets changed their names and numbering systems.

[19] This flattening mirrors that of other global cities towards the end of the 19th century (Clark, 1951) and is typically attributed to the advent of cheap transportation, massive population influx into the outskirts, and the conversion of central residential into commercial neighbourhoods (McDonald, 1989). Indeed, Berlin experienced both an extension of public transport systems and high levels of in-migration in the second half of the 19th century, and the 1870s in particular. Between 1875 and 1880, the circle line (*Ringbahn*) was completed, multiple expansions of the horse-driven tram networks (*Pferdebahn*) were finished, and the central metro line (*Stadtbahn*) was being built (Prussian Ministry of Public Works, 1896). Population grew from 964,539 to 1,123,849 people, with annual gross in-migration rates of up to 140 immigrants per 1000 inhabitants (Hirschberg, 1904, p. 3 & 118).

Scaling to other cities. While city directories across the world share important features, they differ with respect to page layouts and information content. Moreover, the availability of corresponding reference information such as occupational dictionaries and historic maps varies. The usefulness of our algorithm for other researchers depends on how easily the work flow can be adapted to other cities' directories. In Web Appendix A.3, we focus on an example from a North American city, the Boston 1862 directory. We discuss to what extent our work flow can be applied off-the-shelf and where it would need adaptions and extensions. We also highlight additional tools and sources that help researchers interested in an application to US cities. In particular, we identify two substantial, but solvable challenges in applying the work flow to Boston: (i) Street names are often abbreviated and thus potentially misinterpreted and (ii) entries are sometimes continued in a right-justified position on the text line *above*, necessitating an adjustment to our algorithm's entry consolidation approach. While other cities' directories will introduce similar challenges to the algorithm, we conclude that adaptions can be carried out with moderate effort.

4. Conclusion and future applications

This paper presents an algorithmic work flow to extract micro-geographic data at the household or firm level from widely available historical city directories. Under-reporting of the poor is a common feature of this source, but it does not seem to seriously affect between-tract comparisons for typical variables of interest such as average social status, a close proxy for average income. Additional manual labour inputs increase precision, but the size of these gains when going from semi-automatic to fully-manual approaches does not seem to justify the cost. With relatively little manual work, researchers can thus create high-quality geo-referenced household-level data sets in the absence of census data. Once the work flow is successfully implemented for one year, scaling to other years comes at a comparatively low cost, allowing to analyse urban change over time. We chose Berlin as a case study as it encapsulates many additional idiosyncratic challenges and thus strongly biases our analysis against finding huge potential for automation. Having found such potential nonetheless, we are confident that our insights generalise to applications for many other cities.

Perhaps most obviously, the first set of such applications pertains to a literature that exploits shocks to cities such as fires, wars, and earthquakes to test theories in urban economics (Ahlfeldt et al., 2015; Hornbeck and Keniston, 2017; Siodla, 2017). Similarly, natural experiments in the pandemic-prone 19th century can improve our understanding of health economics, e.g. by testing the effects of social diversity on health outcomes (Kappner, 2022a). So far this literature has been restricted to cities (and shocks) where administrative before-after shock data exist. Our algorithm lifts this restriction, since city directories were published annually in most medium-sized and large cities.

A second set of applications could improve our understanding of how different transport modes changed the structure of cities and to what degrees those persisted. In the beginning of the 19th century, cities were typically monocentric (Anas et al., 1998). Early rapid public transport networks established towards the turn of the century appear to have cemented concentration where they appeared (Ahlfeldt et al., 2020), whereas individualised mobility shaped the structure of cities through the required build-up of highways (Brinkman and Lin, 2019). That some American directories contain information on the location of living *and* workplace (Knights, 1969) would greatly help the analysis of the interaction of transport modes and the urban spatial structure. However, most importantly, micro-geographic data based on city directories from around the globe will allow us to understand the evolution of the economic structure of cities from a comparative perspective.

A third set of potential applications pertains to urban inequality. For example, recent studies explore within-urban health inequality and unequal access to sanitary infrastructure in historical cities (Beach et al., 2022; Costa and Kahn, 2015; Kappner, 2022b). Our application suggests that city directories can be used to derive inequality measures for consistent spatial units over time, and potentially track the sorting of households in response to epidemic shocks and sanitary investments.

In addition to these thematic fields of applications, future methodological work could provide additional validation exercises, improve the existing work flow, and extend it with new features. First, researchers could explore additional potential pitfalls when scaling to multiple years. For instance, household and firms might select into providing updated information for a given directory's next-year volume. While our analysis of changing population density gradients for Berlin does not suggest that this is a major problem, other setups might be more prone to such bias. Second, researchers could further improve the work flow presented here. One such improvement would be implementing multiple OCR algorithms, to identify common matches across them, and thereby to reduce Type-1 errors (such as in Bailey et al., 2020). Finally, we see two potential extensions to the algorithm that would greatly improve its applicability across academic fields. One could use city directories for tracking households over time by combining our algorithm with the recent advances in record-linkage algorithms (Clark and Cummins, 2015). Moreover, a natural extension would be to adapt the algorithm to firm directories ("yellow pages"). For instance, linking recognised firms to SIC codes would allow researchers to analyse the evolution of within-city industrial clusters over time.

Appendix A. Web Appendix

A1. Sources

A1.1. Berlin case study data

The Berlin city directory. For the city of Berlin, city directories (*Adressbücher*) were published between 1704 and 1970, with an almost annual frequency since 1820. Before their official incorporation into the Berlin municipality in 1920, a growing number of

suburbs were also included in the directories. After WW2, the directory only referred to the Western part of the city.[20] In our study, we use the 1880 directory (Ludwig, 1880), whose spatial coverage extends to what is now central Berlin, i.e. *Mitte* and parts of neighbouring districts. We also use the 1875 directory (Ludwig, 1875) to train our OCR model and to estimate the change in the city-wide population density gradient between 1875 and 1880 in Section 3.3.

Geo-referencing. To reference directory entries in space, we rely on several historical sources. Most importantly, the first true-to-scale cadastral map for the city was published in 1910 (Straube, 1910). It shows the extent of every plot and its house number. We geo-reference this map, draw polygons for every plot and compute their centroids to localise addresses. For our semi-automatic geo-referencing approach, we also use the Straube map to draw linestrings for every street and transcribe the house numbers at both ends. To account for changes in the street grid, street names and house numbers (e.g. due to plot consolidation) between 1880 and 1910, we supplement our data with information reported in the 1880 city directory (Ludwig, 1880) and a not-to-scale map from around 1880 (Straube, 1883).

Status-referencing. To map household heads' occupations to the 1675 distinct occupation codes contained in the Historical International Classification of Occupations (HISCO, Leeuwen et al., 2002), we employ several sources. For our fully-automatic status-referencing approach, we simply use the online HISCO database, containing 33,620 entries of which 1297 refer to German occupations. For our semi-automatic status-referencing approach we rely on a list of 6489 German occupations published in the 1882 Prussian occupational census (Königliches Statistisches Amt, 1884, (30)–(67)). Importantly, this list also indicates each occupation's belonging to one of 153 "occupational groups" (e.g. industries or sectors of the government bureaucracy). For each of these groups, we manually pick a HISCO code we deem most representative. Finally, for the manual referencing approach, we individually code each occupation appearing at least twice in our directory data set. This involved substantial research based on historical encyclopedias (see Kappner, 2022a, p. 59, for further details). To map HISCO codes to the HISCAM social status scale, we obtained a cross-walk from the HISCAM website, applying the most-robust "men-only universal scale" as recommended by the authors (Lambert et al., 2013).

A1.2. Validation data

In our validation exercise (Section 3.2), we employ high-quality administrative data reported on the level of 216 census tracts (*Stadtbezirke*).[21] We geo-reference them using a historical map of tract boundaries (Straube, 1883). We get the number of households per tract from the 1880 census, reported in Böckh, 1883, pp. 66–69). Data allowing us to reconstruct average income per tract and the within-tract income distribution comes from Berlin's statistical yearbook (Böckh, 1881, pp. 229–236) and the magistrate's administrative report (Magistrat zu Berlin, 1881, pp.18–25). For more information on the income and tax data, see Web Appendix A.2. In Section 3.3, we use building-level household counts for 1875 and 1880 reported in Böckh (1878) and Böckh (1883).

A1.3. Meta study

We derive our samples of "largest cities" in the US and world in 1900 respectively from the Clio-infra project (administed by Buringh, Bosker and Buringh, 2017; Bosker et al., 2013). These data exclude Chinese cities, but for the general point this omission has little relevance. We then collect the date of the first city directory from Williams (1913). For very few cities, this "directory of directories" does not contain the date of the first city directory. In these cases, we collect it from other sources (available upon request).

A2. Estimating top-shares

A2.1. Administrative income data

It is important to note that the Prussian tax data for this period is very detailed and covers a large share of the population. In 1880, only 167,306 inhabitants did live in households paying no tax, whereas 775,342 lived in households paying the so-called "class tax" (an income tax for those with small incomes)[22] and 82,062 lived in households paying the income tax proper. This allows us to compute top-shares from the 17th percentile upwards.

The statistical yearbook for the city of Berlin (Böckh, 1881) and the magistrate's annual report (Magistrat zu Berlin, 1881) provide the following information at the tract level:

- The number of people paying (i) no taxes because they do not earn above a minimum threshold, (ii) the number of people and tax units in households paying "class tax", classified by income bracket, (iii) the number of people in households paying income tax (Magistrat zu Berlin, 1881).
- the (i) average income (150 Mark) of those not paying any form of tax and (ii) the assumed average income for each bracket of the "class tax" (Böckh, 1881, p. 231).
- the average income per capita for each tract (Böckh, 1881, p. 234).

[20] Scans are available from the *Zentral- und Landesbibliothek Berlin*. See Heegewaldt and Rohrlach (1990) and von Gebhardt (1930) on the history of the Berlin city directories.

[21] In 1880, these tracts had an average population of 5000, rendering them comparable to modern US census tracts, whose target size is 4000 inhabitants.

[22] The *Klassensteuer* underwent many reforms through the 19th century. At this point in time, it was based on the income of the citizens and thus comparable to an income tax.

Additionally, Böckh (1881), p. 233) contains the city-wide ratio of tax units paying the income tax relative to the number of people living in such households ($\frac{25{,}200}{82{,}062}$).

Number of tax units. In the first step, we calculate for each tract the number of tax units. We assume that household sizes do not vary across tracts for the small portion of the population that pays the income tax. This allows us to calculate the number of tax units paying the income tax by applying the above city-wide ratio. Additionally, we estimate the number of tax units paying no taxes by multiplying the tract-specific ratio $\frac{\text{tax units paying the "class tax"}}{\text{people in households paying the "class tax"}}$ with the number of people paying no taxes in a given tract.

Total income. For each tract t we calculate the total income as:

$$Y_t = POP_t \times PCINC_t \tag{A.1}$$

where POP and $PCINC$ are the population and the per-capita income as derived from the sources.

Income of those paying income tax. Since the source specifies the average income of the tax units not paying taxes and those paying "class taxes", we can estimate the income of those paying income taxes by calculating:

$$Y_t^{\text{paying income tax}} = Y_t - Y_t^{\text{paying no taxes}} - Y_t^{\text{paying class tax}} \tag{A.2}$$

Calculation of top-shares. We now have tabulated data on income and tax units for fourteen bins:

1. those not paying taxes: below 420 Marks
2. those paying class tax: 12 classes, ranging from [1] 420-600 Marks to [12] 2700-3000 Marks
3. those paying income tax: earning above 3000 Marks

We follow the industry standard by applying the generalised Pareto interpolation suggested by Blanchet et al. (forthcoming) to estimate the top-shares.

A2.2. HISCAM scores

For the top-shares in the "HISCAM mass", we do not require the Pareto interpolation as we have individual-level data. We simply rank all households in a given tract, then calculate the overall mass, and the mass above the percentile p of interest. By dividing the mass above percentile p by the total mass, we arrive at the top-share.

A3. Adapting the algorithm to other cities

Our algorithm has a high degree of flexibility such that it can be applied to other cities. As stated in Section 3, the Berlin directory adds three additional challenges that may not be present for other cities, in particular those located in the US, to the same degree: non-standard script, changes in the street grid, and scarce occupational reference data. In the following we discuss likely changes that would have to be made when applying the algorithm to another case. To facilitate an easier discussion of potential changes we chose an example city: Boston in 1862. We demonstrate some important changes that would have to be made to the algorithm.

Preparation. The Boston Public Library provides links to scans of the city's directories, spanning issues from the late 18th century to the late 20th century. For example, the 1862 issue can be accessed via the Internet Archive/archive.org. On pages 15 to 444, it contains an alphabetical list of household heads with occupation and address ("white pages"). First, one would need to extract these pages from the downloaded file and store them as a new PDF or a series of pages-as-images, depending on the employed OCR software. Second, the quality of these input files should be enhanced, in particular by dewarping and sharpening the scans, and removing noise. Several non-commercial software packages are available, for instance the Python-based OpenCV framework (OpenCV contributors, 2016) or OCR4all (Reul et al., 2019).

Recognition. The Boston directory is machine-printed and features a highly regular layout with clearly delineated columns and without images or other irregular elements. In contrast to our Berlin example, the typeface is a modern serif font. Hence, a standard OCR framework like Tesseract (Tesseract contributors, 2021) or LayoutParser (Shen et al., 2021) is likely to yield satisfactory results off-the-shelf (both on the character recognition and segmentation level), i.e., without labor-intensive training of a specialised recognition model. Researchers who are less familiar with command line interface-based tools can use the OCR capabilities of commercial software with graphical user interfaces, e.g. Adobe Acrobat or Abbyy Finereader. While our algorithm accepts any OCR output structured in the PAGE XML format (Pletschacher and Antonacopoulos, 2010), OCR output in other formats needs to be converted first.[23]

Structuring. Our excerpt of the 1880 Berlin city directory reports household heads by street address. In contrast, the list of household heads in the 1862 Boston directory is ordered alphabetically, reporting the residency address of every single entry separately. Thus, our algorithmic approach to infer each entry's street and house number from dedicated "location lines" can be wholly skipped. Similar to our use case, entries spanning multiple lines are clearly indented and thus easily joinable following our procedure. The resulting consolidated lines for each entry have a highly regular structure, e.g. "*last name first name(s), occupation, address, other information.*".[24] Such regular structures lend themselves to the trainable conditional random field (CRD) based parsing algorithm incorporated in our work flow (Spaan and Balogh, 2021).

[23] A useful overview of the respective tools is available at the OCR conversion GitHub repository https://github.com/cneud/ocr-conversion.

[24] A unique challenge in the Boston case results from the occasional continuation of an entry line in the right-justified position on the line *above*, e.g. in line 2 on page 433. To correctly attribute this information, our algorithm would need an extension that identifies strings starting with a square bracket *[* and appends them at the and of the next line.

Referencing I: Status-referencing. To sort the identified occupations by social status, researchers can build on a range of recent classification efforts for 19th century English-language occupations. Given the more complete coverage of the English-language HISCO database, we expect the automatic referencing approach to perform better than in the German case.[25] However, a semi-automatic referencing approach based on a mapping of occupations to HISCO (or any other classification scheme) is also feasible within our work flow. For instance, the IPUMS USA census digitisation project contains a classification scheme for historical American occupations.[26] Alternatively, Roberts et al. (2003) coded historical occupation data for the United States, Canada, and United Kingdom within the North Atlantic Population Project (Ruggles et al., 2011) and classified them with a modified version of HISCO. These occupational dictionaries should cover close to the the universe of the occupations listed in Boston's 1862 directory. Once they are transformed into a CSV, JSON or similar database file format, such occupational dictionaries are easily fed into our script.

Referencing II: Geo-referencing. While both our (semi-)automatic and manual referencing approaches can be applied off-the-shelf to address information parsed from Boston's 1862 directory, there are two reasons why the matching results may be less reliable than in the Berlin 1880 case. First, because the Boston directory is ordered by family name rather than address, there is no correlation between the street names and house numbers of two adjacent entries. Thus, in contrast to the Berlin case, entries with faulty address information (indicated, for instance, by a low match certainty or no match at all for a given geo-referencing approach) cannot derive information from neighbouring entries. Second, the Boston directory abbreviates street names such that finding a match within a geo-referencing approach optimised for non-abbreviated street names becomes less likely. For instance, Google Maps fails to identify "*88 Com'l*" as "*88 Commercial St.*". While our algorithm does not account for this issue, researchers can decode abbreviated street names automatically as shown by Spaan (2017) for the New York city directory. To pursue a semi-automatic or manual geo-referencing approach, one could employ the collection of historical, geo-referenced maps ("Boston Atlas") provided by the Boston Planning & Development Agency.

Validation with aggregates at a lower spatial resolution. City directories sometimes also reported statistics about the city that were collected at a more aggregate level. As such the 1864 issue of the Boston city directory (Adams, Sampson, & Company, 1864) documents the population, number of families, and dwellings for each of Boston's twelve wards as of 1860. These data could be used to test whether some areas of the cities are more affected by under-reporting than others. Alternatively, it would also be possible to calculate summary statistics of the US census at the places-level (based on the work by Berkes et al., forthcoming). It is important to note, however, that these data, like the ward-level data, are not at the street level either.

References

Aaronson, D., Hartley, D., Mazumder, B., 2021. The effects of the 1930s HOLC "redlining" maps. American Economic Journal: Economic Policy 13 (4), 355–392. doi:10.1257/pol.20190414. https://www.aeaweb.org/articles?id=10.1257/pol.20190414

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., Pérez, S., 2020. Automated linking of historical data. NBER Working Paper No. 25825.

Adams, Sampson, & Company (Ed.), 1862. The Boston City Directory. Embracing the city record, a general directory of citizens, and a business directory, for the year commencing July 1, 1862. Adams, Sampson, & Company. Boston.

Adams, Sampson, & Company (Ed.), 1864. The Boston City Directory. Embracing the city record, a general directory of citizens, and a business directory, for the year commencing July 1, 1864. Adams, Sampson, & Company. Boston.

Ahlfeldt, G.M., Albers, T.N.H., Behrens, K., 2020. Prime locations. CEPR Discussion Paper No. 15470.

Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2015. The economics of density: Evidence from the Berlin wall. Econometrica 83 (6), 2127–2189.

Albers, T. N. H., Kappner, K., 2022. City directory extraction repository. https://github.com/kkappner/berlin-city-directory.

Anas, A., Arnott, R., Small, K.A., 1998. Urban spatial structure. Journal of Economic Literature 36 (3), 1426–1464.

Bailey, M.J., Cole, C., Henderson, M., Massey, C., 2020. How well do automated linking methods perform? lessons from us historical data. Journal of Economic Literature 58 (4), 997–1044. doi:10.1257/jel.20191526. https://www.aeaweb.org/articles?id=10.1257/jel.20191526

Beach, B., Parman, J., Saavedra, M., 2022. Segregation and the initial provision of water in the united states. American Economic Review: Papers and Proceedings 112.

Bell, S., Marlow, T., Wombacher, K., Hitt, A., Parikh, N., Zsom, A., Frickel, S., 2020. Automated data extraction from historical city directories: The rise and fall of mid-century gas stations in providence, ri. PLoS One 15 (8), e0220219. doi:10.1371/journal.pone.0220219.

Berenbaum, D., Deighan, D., Marlow, T., Lee, A., Frickel, S., Howison, M., et al, 2019. Mining spatio-temporal data on industrialization from historical registries. Journal of Environmental Informatics 34 (1), 28–34. doi:10.3808/jei.201700381.

Berkes, E., Karger, E., Nencka, P., forthcoming. The census place project: A method for geolocating unstructured place names. Explorations in Economic History.

Blanchet, T., Fournier, J., Piketty, T., forthcoming. Generalized pareto curves: Theory and applications. Review of Income and Wealth10.1111/roiw.12510

Böckh, R., 1878. Die Bevölkerungs-, Gewerbe- und Wohnungs-Aufnahme vom 1. December 1880 in der Stadt Berlin. Erstes Heft. Commissions-Verlag von Leonhard Simion, Berlin.

Böckh, R., 1881. Statistisches Jahrbuch der Stadt Berlin. Siebenter Jahrgang. Statistik des Jahres 1879. Verlag von Leonhard Simion, Berlin.

Böckh, R., 1883. Die Bevölkerungs- und Wohnungs-Aufnahme vom 1. December 1880 in der Stadt Berlin. Erstes Heft. Commissions-Verlag von Leonhard Simion, Berlin.

Bosker, M., Buringh, E., 2017. City seeds: Geography and the origins of the european city system. Journal of Urban Economics 98, 139–157. doi:10.1016/j.jue.2015.09.003. https://www.sciencedirect.com/science/article/pii/S0094119015000613

Bosker, M., Buringh, E., Van Zanden, J.L., 2013. From Baghdad to London: Unraveling urban development in Europe, the Middle East, and North Africa, 800–1800. Review of Economics and Statistics 95 (4), 1418–1437.

Brinkman, J., Lin, J., 2019. Freeway revolts!. Federal Reserve Bank of Philadelphia Working Papers No. 19-29.

Caesmann, M., Caprettini, B., Voth, H.-J., Yanagizawa-Drott, D., 2021. Going viral: Nazi marches and the spread of extremism. Mimeo.

Cambon, J., Hernangómez, D., Belanger, C., Possenriede, D., 2021. tidygeocoder: An R package for geocoding. Journal of Open Source Software 6 (65), 3544. doi:10.21105/joss.03544.

Chiswick, B.R., Robinson, R.H., 2021. Women at work in the united states since 1860: An analysis of unreported family workers. Explorations in Economic History 82, 101406. doi:10.1016/j.eeh.2021.101406. https://www.sciencedirect.com/science/article/pii/S0014498321000243

Clark, C., 1951. Urban population densities. Journal of the Royal Statistical Society. Series A (General) 114 (4), 490–496.

[25] As of August 2022, the HISCO database contained 2299 English-language occupations.

[26] This classification scheme can also be mapped to HISCO. See https://doi.org/10.17026/dans-zap-qxmc for a cross-walk.

Clark, G., Cummins, N., 2015. Intergenerational wealth mobility in england, 1858-2012: surnames and social mobility. The Economic Journal 125 (582), 61–85. doi:10.1111/ecoj.12165. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12165
Combes, P.-P., Gobillon, L., Zylberberg, Y., 2021. Urban economics in a historical perspective: Recovering data with machine learning. Regional Science and Urban Economics 103711. doi:10.1016/j.regsciurbeco.2021.103711. https://www.sciencedirect.com/science/article/pii/S0166046221000715
Correia, S., Guimarães, P., Zylkin, T., 2020. Fast poisson estimation with high-dimensional fixed effects. The Stata Journal 20 (1), 95–115.
Costa, D.L., Kahn, M.E., 2015. Declining mortality inequality within cities during the health transition. American Economic Review 105 (5), 564–569. doi:10.1257/aer.p20151070.
Cura, R., Dumenieu, B., Abadie, N., Costes, B., Perret, J., Gribaudi, M., 2018. Historical collaborative geocoding. ISPRS International Journal of Geo-Information 7 (7). doi:10.3390/ijgi7070262.
Currie, J., Kleven, H., Zwiers, E., 2020. Technology and big data are changing economics: Mining text to track methods. AEA Papers and Proceedings 110, 42–48. doi:10.1257/pandp.20201058.
Dahl, C.M., Johansen, T., Sørensen, E.N., Wittrock, S., 2021. HANA: A handwritten name database for offline handwritten text recognition. CoRR abs/2101.10862. https://arxiv.org/abs/2101.10862
Dahl, C.M., Johansen, T.S.D., Sørensen, E.N., Westermann, C.E., Wittrock, S.F., 2021. Applications of machine learning in document digitisation. CoRR abs/2102.03239. https://arxiv.org/abs/2102.03239
Dittmar, J.E., 2011. Information technology and economic Change: The impact of the printing press. The Quarterly Journal of Economics 126 (3), 1133–1172. doi:10.1093/qje/qjr035.
Gallman, R.E., Wallis, J.J., 1993. American Economic Growth and Standards of Living before the Civil War: National Bureau of Economic Research Conference Report. University of Chicago Press, Chicago.
von Gebhardt, P., 1930. Die Anfänge des Berliner Adressbuches: ein bibliographischer Versuch. Selbstverlag, Berlin.
Geopy contributers, 2021. geopy. https://github.com/geopy/geopy.
Glaeser, E.L., 2021. What can developing cities today learn from the urban past? NBER Working Paper No. 28814 doi:10.3386/w28814. http://www.nber.org/papers/w28814
Gutmann, M.P., Merchant, E.K., Roberts, E., 2018. 'Big Data' in Economic History. The Journal of Economic History 78 (1), 268–299. doi:10.1017/S0022050718000177.
Heblich, S., Hanlon, W., forthcoming. History and urban economics. Regional Science and Urban Economics.
Heblich, S., Trew, A., Zylberberg, Y., 2021. East-side story: Historical pollution and persistent neighborhood sorting. Journal of Political Economy 129 (5), 1508–1552.
Heegewaldt, W., Rohrlach, P.R., 1990. Berliner Adressbücher und Adressenverzeichnisse 1704–1945: eine annotierte Bibliographie mit Standortnachweis für die "ungeteilte" Stadt. Helmut Scherer, Berlin.
Hirschberg, E., 1904. Statistisches Jahrbuch der Stadt Berlin. 28. Jahrgang. Enthaltend die Statistik des Jahres 1903. Verlag von P. Stankiewiez, Berlin.
Hornbeck, R., Keniston, D., 2017. Creative destruction: Barriers to urban growth and the great boston fire of 1872. American Economic Review 107 (6), 1365–1398.
Kappner, K., 2022. Dense, diverse and healthy? mixed-income housing and the spread of urban epidemics. Mimeo.
Kappner, K., 2022. Sanitation, externalities and the urban mortality transition. Mimeo.
Knights, P.R., 1969. City directories as aids to ante-bellum urban studies: A research note. Historical Methods Newsletter 2 (4), 1–10. doi:10.1080/00182494.1969.10593895.
Knudsen, A. S. B., 2021. Those who stayed: Selection and cultural change in the age of mass migration.
Königliches Statistisches Amt, 1884. Berufsstatistik nach der allgemeinen Berufszählung vom 5. Juni 1882. 3. Berufsstatistik der Staaten und gräeren Verwaltungsbezirke. Erster Theil. Statistik des Deutschen Reichs. Neue Folge. Band 4. Erstes Drittel. Verlag von Puttkammer & Mühlbrecht, Berlin.
Lambert, P.S., Zijdeman, R.L., Leeuwen, M.H.D.V., Maas, I., Prandy, K., 2013. The construction of hiscam: A stratification scale based on social interactions for historical comparative research. Historical Methods: A Journal of Quantitative and Interdisciplinary History 46 (2), 77–89. doi:10.1080/01615440.2012.715569.
Leeuwen, M.H.v., Maas, I., Miles, A., 2002. HISCO: Historical international standard classification of occupations. Leuven University Press, Leuven.
di Lenardo, I., Barman, R., Descombes, A.B., Kaplan, F., 2019. Repopulating Paris: Massive extraction of 4 million addresses from city directories between 1839 and 1922. Abstracts and Posters from the Digital Humanities 2019 conference. Utrecht University.
Ludwig, A., 1875. Berliner Adreß-Buch für das Jahr 1875. Unter Benutzung amtlicher Quellen. VII. Jahrgang. Druck und Verlag der Societät der Berliner Bürger-Zeitung, Berlin.
Ludwig, A., 1880. Berliner Adreß-Buch für das Jahr 1880. Unter Benutzung amtlicher Quellen. XII. Jahrgang. W & S Loewenthal, Berlin.
Magistrat zu Berlin, 1881. Verwaltungs-Bericht des Magistrats zu Berlin pro 1880. Beilagen zu Nr. 33. Julius Sittenfeld, Berlin.
McDonald, J.F., 1989. Econometric studies of urban population density: A survey. Journal of Urban Economics 26 (3), 361–385. doi:10.1016/0094-1190(89)90009-0. https://www.sciencedirect.com/science/article/pii/0094119089900090
OpenCV contributors, 2016. opencv-python. https://github.com/opencv/opencv-python.
Pletschacher, S., Antonacopoulos, A., 2010. The PAGE (page analysis and ground-truth elements) format framework. Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010) 257–260. doi:10.1109/ICPR.2010.72.
Prussian Ministry of Public Works, 1896. Berlin und seine Eisenbahnen 1846-1896, Vol. 1. Springer.
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 9 (22), 4853. doi:10.3390/app9224853.
Reul, C., Springmann, U., Puppe, F., 2017. Larex: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, pp. 137–142.
Roberts, E., Woollard, M., Ronnander, C., Dillon, L.Y., Thorvaldsen, G., 2003. Occupational classification in the north atlantic population project. Historical Methods: A Journal of Quantitative and Interdisciplinary History 36 (2), 89–96.
Rose-Redwood, R., Tanter, A., 2012. Introduction: Governmentality, house numbering and the spatial history of the modern city. Urban History 39 (4), 607–613. doi:10.1017/S0963926812000405.
Ruggles, S., Roberts, E., Sarkar, S., Sobek, M., 2011. The north atlantic population project. progress and prospects. Historical Methods 44 (1), 1–6. doi:10.1080/01615440.2010.515377.
Schlegel, I., 2021. Automated extraction of labels from large-scale historical maps. AGILE: GIScience Series 2, 12. doi:10.5194/agile-giss-2-12-2021. https://agile-giss.copernicus.org/articles/2/12/2021/
Shaw, G., Coles, T., 1995. European directories: a universal source for urban historians. Urban History 22 (1), 85–102. doi:10.1017/S0963926800011391.
Shaw, G., Coles, T., 1997. A Guide to European Town Directories. Ashgate. https://books.google.de/books?id=O1tAwAEACAAJ
Shaw, G., Tipper, A., 2010. British directories, 2nd Bloomsbury Publishing. https://books.google.de/books?id=ROneBAAAQBAJ
Shen, Z., Zhang, R., Dell, M., Lee, B., Carlson, J., Li, W., 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. International Conference on Document Analysis and Recognition.
Siodla, J., 2017. Clean slate: Land-use changes in san francisco after the 1906 disaster. Explorations in Economic History 65, 1–16. doi:10.1016/j.eeh.2017.04.001. https://www.sciencedirect.com/science/article/pii/S001449831730089X
Siodla, J., 2021. Firms, fires, and firebreaks: The impact of the 1906 san francisco disaster on business agglomeration. Regional Science and Urban Economics 88, 103659. doi:10.1016/j.regsciurbeco.2021.103659. https://www.sciencedirect.com/science/article/pii/S0166046221000193
Spaan, B., 2017. nyc-street-normalizer. https://github.com/nypl-spacetime/nyc-street-normalizer.
Spaan, B., Balogh, S., 2021. city-directory-entry-parser. https://github.com/nypl-spacetime/city-directory-entry-parser.
Spear, D.N., 1961. Bibliography of American directories through 1860. American Antiquarian Society Worcester, Mass.
Straube, J., 1883. Plan von Berlin mit Angabe der Sterblichkeitsziffer und graphischer Darstellung der Bevölkerungsdichtigkeit. Map.
Szołtysek, M., Gruber, S., 2016. Mosaic: recovering surviving census records and reconstructing the familial history of europe. The History of the Family 21 (1), 38–60.

Straube, J., 1910. Übersichtsplan von Berlin in 44 Blättern. Map.
Tesseract contributors, 2021. Tesseract open source ocr engine. https://github.com/tesseract-ocr/tesseract.
Wick, C., Reul, C., Puppe, F., 2020. Calamari - A high-performance tensorflow-based deep learning package for optical character recognition. Digital Humanities Quarterly 14 (1).
Wiest, E., 1991. Gesellschaft und Wirtschaft in München, 1830-1920: die sozioökonomische Entwicklung der Stadt dargestellt anhand historischer Adressbücher, Vol. 3. Centaurus-Verlagsgesellschaft.
Williams, A., 1913. The development and growth of city directories. Williams directory Company.