Perks and Pitfalls of City Directories as a Micro-Geographic Data Source

Thilo N. H. Albers (HU Berlin)
Kalle Kappner (HU Berlin)

Discussion Paper No. 315

January 29, 2022

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Perks and pitfalls of city directories as a micro-geographic data source[*]

Thilo N. H. Albers[†] Kalle Kappner[‡]

January 28, 2022

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 potential applications in economic history and beyond.

Keywords: city directories, data extraction, granular spatial data
JEL classification: C8, R1, N9

[*] We thank Monique Reiske, Kristian Behrens, Nikolaus Wolf and the participants of the 2021 Methodological Advances in the Extraction and Analysis of Historical Data conference for comments. Petya Prodanova provided excellent research assistance. Albers gratefully acknowledges the financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119).

[†] Humboldt University Berlin; alberstn@hu-berlin.de.

[‡] Humboldt University Berlin; kallekappner@googlemail.de.

Introduction

Throughout history, individuals have shaped the appearance of cities through their decisions on where to live, work, and consume. Typically, these choices are based on the prevailing first and second nature geography, and do not seldom change in response to economic shocks such as fires, earthquakes, pandemics, and wars. Such shocks provide a source of exogenous variation, allowing us to test theoretical propositions in urban economics and beyond.[1] While history provides many shocks, we often lack the appropriate granular spatial data to exploit them as natural experiments. Existing work typically relies on administrative data or digitized maps. Naturally, this confines studies to urban places where such material had been collected at the required spatio-temporal resolution. In many cases, administrative data do not exist at all.

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 document in a small meta-study. They typically contain information on the occupation (for individuals) and industry (for businesses), allowing 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 constrains have so far largely prevented the exploitation of the granularity of their spatial data dimension.[2] 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.[3] We then take the work flow to a use case and assess its strength under varying levels of manual labour inputs. This allows us to provide recommendations where such inputs are most productive.

The 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 for the city directory they are interested in. The process of structuring OCR output—that is the parsing of recognized 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 (HISCAM) 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

[1] See Siodla (2017), Hornbeck and Keniston (2017), and Heblich et al. (2020) for recent examples.

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

[3] Code and replication files are at `https://github.com/kkappner/berlin-city-directory`.

to the street grid (a challenge for geo-referencing), and social status classification systems for historical German occupations are limited in their extent (a challenge for status-referencing). Moreover, using Berlin data 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 same spatial units, rank correlations provide a measure of fit to explore two questions. First, how well do city directory data perform in replicating patterns in administrative data given that they may under-report the poorest households? Second, where are manual work hours, e.g. for building training data sets or digitizing 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 HISCAM scores, and income vs. HISCAM 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 efforts in referencing 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.

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; Gutmann et al., 2018; Combes et al., 2021; Currie et al., 2020; Dahl et al., 2021b). 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. Finally, we add to an older literature that has used address books in the past and a nascent one that exploits their granular dimension.

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 spatial data. Section 3 applies the workflow to the Berlin case and assesses trade-offs between precision and manual labour input. Section 4 discusses potential applications and concludes.

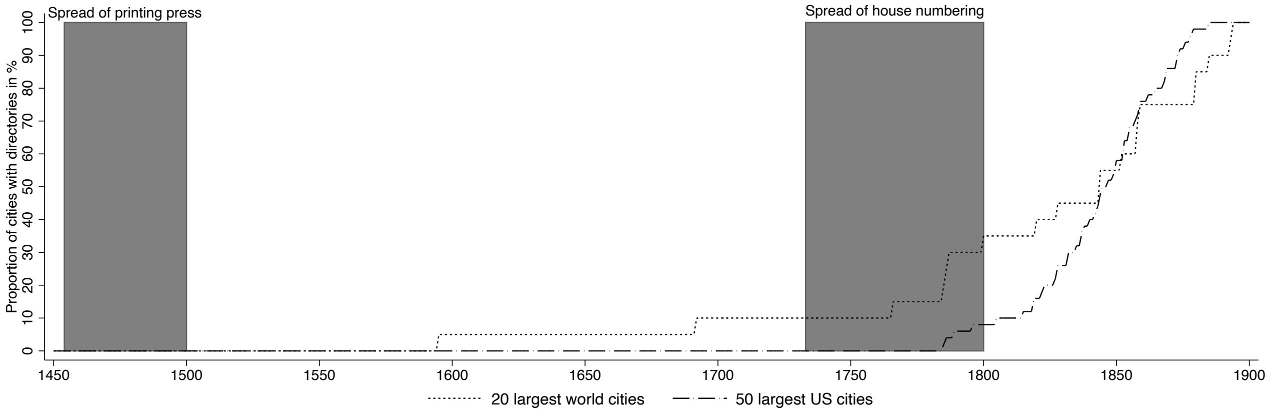
[4] Under-reporting extends to all factors that have been historically correlated with income such as race and gender. It also exists in administrative data (e.g. Chiswick and Robinson, 2021, for gender in US-census data).

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 Figure 1 suggests that the timing of the widespread adoption is itself a tale about the interaction of technology proper, social technologies, 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 technology, house numbering, played an important role.

Figure 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 (administed by Buringh, Bosker et al., 2013; Bosker and Buringh, 2017). 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.

First, to understand the role of economic growth it is important to note that directories were an innovation coming from the private sector.[5] 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 larger cities and substantial increases in inter-urban market integration (for the timing

[5] 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.

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, making directories a necessity for firms in larger cities simply in order to advertise their services (Knights, 1969).

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 technology greatly improved the usefulness and extent of city directories.

When economic growth and market integration increased and the social technology of house numbering was implemented, city directories spread rapidly (Figure 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 City directories around the world typically contained three pieces of information for a given household: home address, name, and occupation. Depending on the country and local context, directories often recorded additional information. In the United States, they sometimes added information on birth places and the 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, 2021, 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 household-level census records, the latter of which were not preserved in many countries. Even where census data are available, city directories can be useful. They can complement census data with the locations of firms, which they usually contained. Moreover, they were typically published annually and thus at a much higher frequency than censuses were conducted.

The main advantage of directory data, however, derives from its nature as point data. This allows the researcher to arbitrarily define spatial units. 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 the Historical International Classification of Occupations (Leeuwen et al., 2002; Lambert et al., 2013) or be combined with wage data. This allows for estimating wage and status distributions. In principle, one could also

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	

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 (Knights, 1969; Kappner, 2021). 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 administrative data.

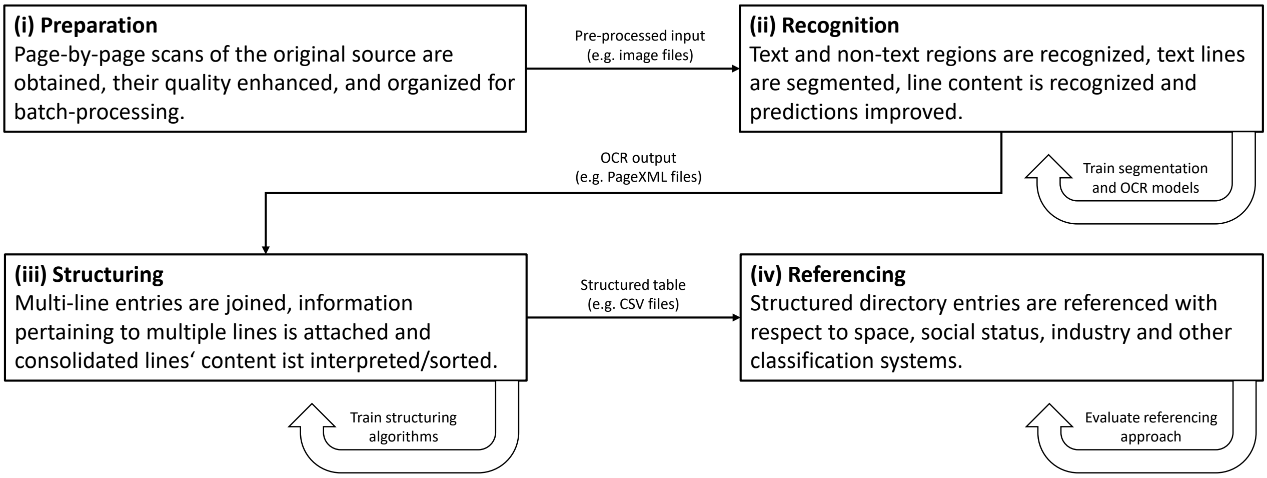
Use of directories up until now At least since the 1930s, local, social, and urban historians and, to some extent, economists have employed city directories as a source (Shaw and Tipper, 2010, p. 1). Knights (1969) discusses some early applications for Philadelphia, Baltimore, and New York, the latter 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, 2021; 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 a generic work flow, highlighting the most important steps, obstacles, and tools characterizing the extraction process. In section 3, we present a detailed application of this work flow to a specific example and validate the results.

The extraction work flow involves four successive steps (Figure 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 recognized and translated into text strings with positional information (**ii**). Following this, the recognized 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**).

Figure 2: A generic work flow for city directory extraction



(i) Preparation
Page-by-page scans of the original source are obtained, their quality enhanced, and organized for batch-processing.

(ii) Recognition
Text and non-text regions are recognized, text lines are segmented, line content is recognized and predictions improved.

(iii) Structuring
Multi-line entries are joined, information pertaining to multiple lines is attached and consolidated lines‘ content ist interpreted/sorted.

(iv) Referencing
Structured directory entries are referenced with respect to space, social status, industry and other classification systems.

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, greatly reducing necessary labour input. An integrated, flexible framework for city directory extraction combines these scattered tools, transforming raw scans into 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 specialized algorithms along the work flow, as indicated in Figure 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 typically starts with input data in the form of scanned images. Only rarely will researchers need to scan original sources themselves as high-quality scans of many city directories are readily provided by local libraries, archives, and other public institutions. In order to prepare these scans for automatic recognition in step **ii**, they usually require substantial enhancement, e.g. deskewing, cropping, binarization, and other normalizations. 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 recognized via OCR. While different OCR environments apply different prediction algorithms, their common approach is to recognize 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), recognize the lines' text content, and possibly enhance the prediction *ex post*. City directories are machine-printed, exert a highly regular layout structure, and contain 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 examples 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).[6]

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 recognized text lines may exhibit many different formats, both as a result of each entry's particular characteristics, and due to OCR detection errors.[7] A suitable structuring approach thus needs to interpret the format of each line. Solutions to this problem range

[6] 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.

[7] For example, one entry might read "*family name, first name, middle name, occupation.*" because this individual has a middle name. Another entry might right "*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.

from manual line-by-line format distinctions, via algorithms that distinguish line formats based on a hard-coded set of regular expressions (e.g. Bell et al. 2020 and Berenbaum et al. 2019), 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).

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.[8] Alternative, historically more sensitive geo-referencing approaches match recognized 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. For example, recognized occupations can be matched to the Historical International Standard of Classification of Occupations (Leeuwen et al., 2002) and its extensive list of historical occupations for various languages.

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

Taking the extraction work flow to the data, we conduct a case study of Berlin's 1880 city directory. Berlin represents a particularly well-suited case to explore the strength of our approach. First, 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). Section 3.1 discusses how these challenges can be overcome not only for our case but more generally. Second, municipal statisticians recorded aggregate data about households, mean incomes, and within-district income tabulations for 200+ census tracts covering our whole study area. This allows us to carry out extensive validation exercises and to assess where manual labour inputs are most-efficiently spent (Section 3.2).

[8] 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.

3.1 Applying the work flow: The case of 19$^{\text{th}}$ 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.

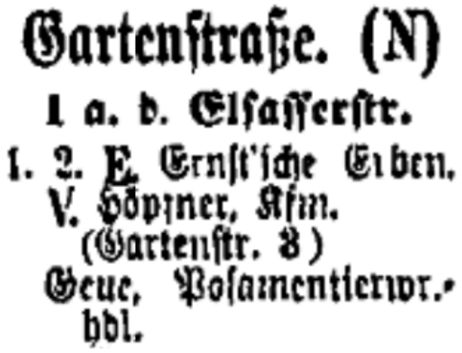
Preparation and recognition (i-ii) We obtain raw scans of Berlin's 1880 city directory from a local library (*Zentral- und Landesbibliothek Berlin*). For preparation and recognition, we rely on OCR4all, applying its standard input scan optimization, region and line segmentation routines, as well as the embedded Calamari OCR engine (Reul et al., 2019; Wick et al., 2020). Our raw data is typeset in Fraktur, a German blackletter script (see Figure 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 specialized 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). The resulting model achieves a satisfactory recognition quality.[9] The two upper panels of Figure 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 center of the column, and indentation in the raw data (Figure 3a): We exploit the bounding box coordinates of recognized 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 (Figure 3b). Similarly, lines whose bounding box' leftmost coordinate are sufficiently shifted to the left of the respective column's center 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

[9] While our procedure bore significant initial investment costs, 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 recognized by conventional OCR environments.

Figure 3: Berlin Application

a) Raw input



Gartenſtraße. (N)
1 a. d. Elſaſſerſtr.
1. 2. E. Ernſt'ſche Erben.
V. Höpfner, Kfm.
(Gartenſtr. 3)
Geue, Poſamentierwr.-
hdl.

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

Note: Stylized 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 examples illustrates the case of an error-free OCR recognition.

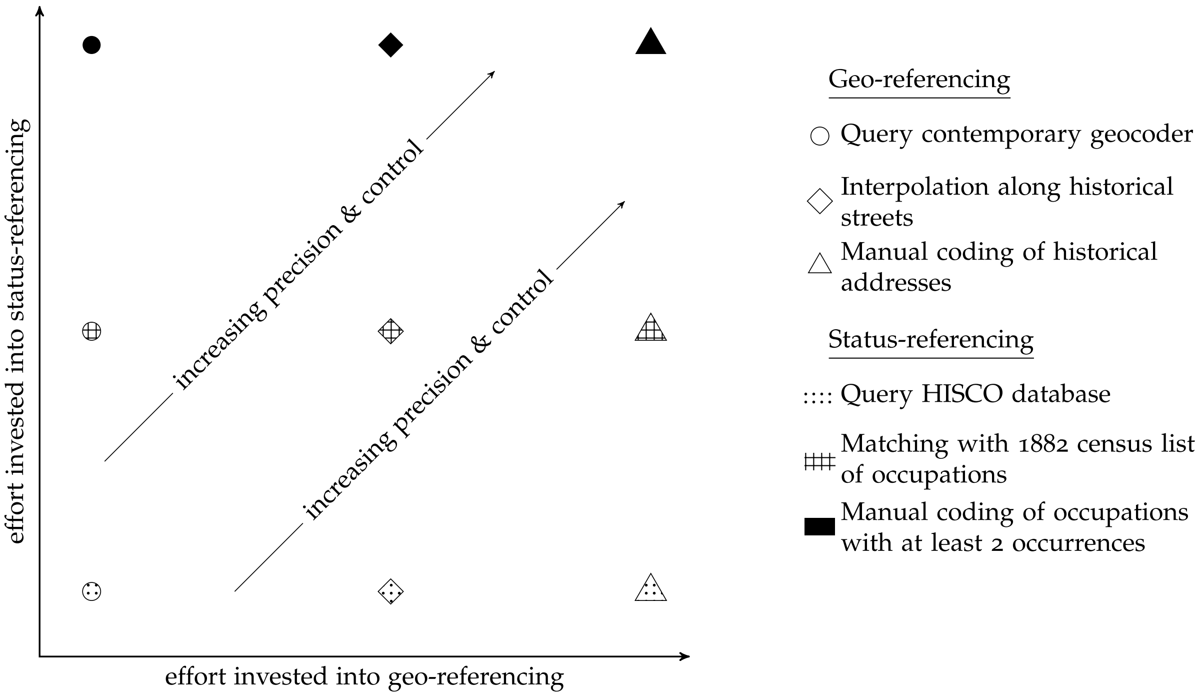
sufficiently shifted to the right relative to the preceding line to identify them, e.g. lines [5] and [7]. Merging accordingly, we arrive at a set of consolidated lines.

The second structuring element is the content *within* each consolidated line, represented in text format in Figure 3b. 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 parenthesized address as in [5]. These indicate (co-)proprietor status and an absentee address, respectively. Using these and similar pattern rules as well as dropping irrelevant information as in [2][10], we arrive at a structured data set of households (Figure 3c).

[10] 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 for geo-referencing purposes.

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 (Figure 3d). For our validation exercise, we apply three approaches for geo-referencing and status-referencing respectively. Figure 4 maps them schematically according to required levels of manual effort.

Figure 4: Effort-precision trade-offs


effort invested into status-referencing
increasing precision & control
increasing precision & control
effort invested into geo-referencing
Geo-referencing
○ Query contemporary geocoder
◇ Interpolation along historical streets
△ Manual coding of historical addresses
Status-referencing
:::: Query HISCO database
⌗ Matching with 1882 census list of occupations
▬ Manual coding of occupations with at least 2 occurrences

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 Figure 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 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 Figure 4). Third, we use a hand-made shapefile containing the exact position of each historical address (△ symbol in Figure 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 (:::: pattern in

Figure 4).[11] 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 Figure 4). Third, we hand-pick an appropriate HISCO code for each occupation occurring at least two times in our structured data set (■ pattern in Figure 4). Finally, for all referencing approaches, we convert HISCO codes to HISCAM social status scores (Lambert et al., 2013).[12]

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 Figure 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, their mean HISCAM value, and the share of the highest-ranked 50% in the total "HISCAM mass". These statistics represent proxies for typical measures of interest in historical urban economics. We create a corresponding data set of population counts, average income, and income inequality measures from administrative reports, all referring exactly to the year 1880.[13] Importantly, we perform the validation exercise for all possible combination of the three geo-referencing (GR) and status-referencing (SR) approaches discussed above. This makes the effort-precision trade-offs visible that working with city directories entail.

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

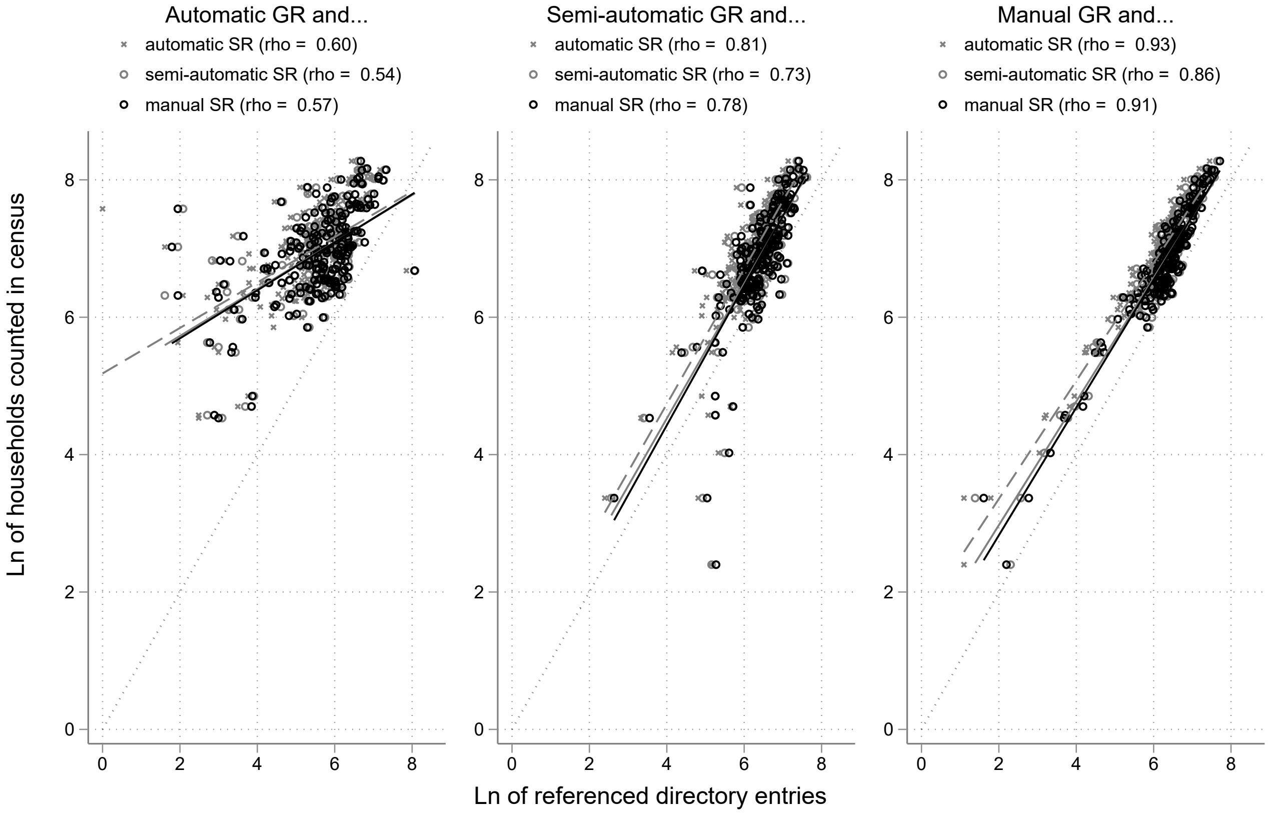
[11] Some occupations are listed more than once in the HISCO database. If there are multiple matches, we pick the first of the most frequent HISCO codes. As of October 2021, the HISCO database contained 33,620 entries.

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

[13] See the Web Appendix A.1 for more details on the sources and procedure.

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 the scaling of population estimates relative to the actual distribution. Figure 5 allows us to gauge the extent of this problem in our application. For each referencing 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.

Figure 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.

Four central results emerge from Figure 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.[14] Second, the rank order of observations with respect to

[14] 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

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 thus 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.[15] 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.[16]

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. Figure 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).

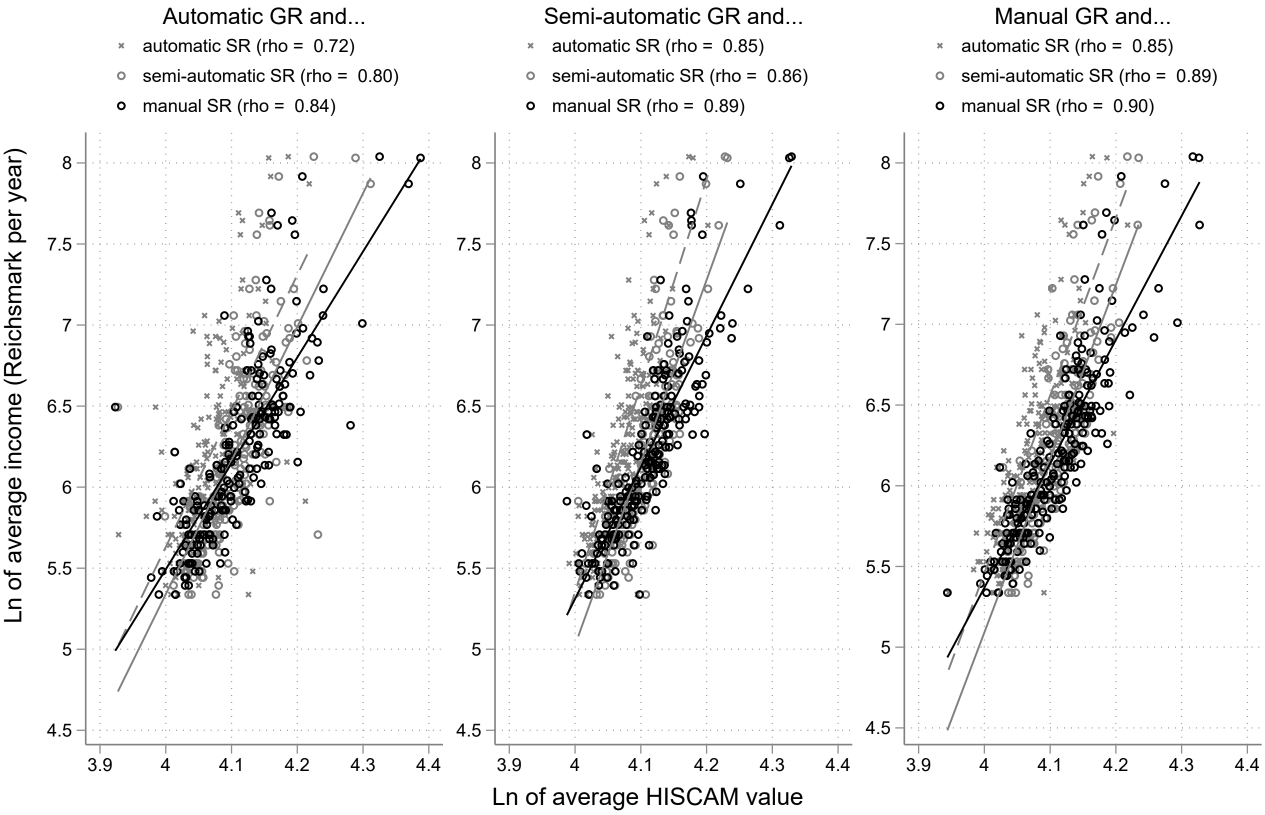
Notwithstanding these conceptual caveats, Figure 6 suggests that the bottom-censored nature of Berlin's 1880 city directory has no discernible impact on the derived average status

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.

[15] 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" system ("Zick-Zack-System") with even numbers on one side of the street and odd numbers on the other side. Existing streets 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.

[16] Additionally, manual GR aproaches 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.

Figure 6: Mean incomes vs. mean HISCAM values

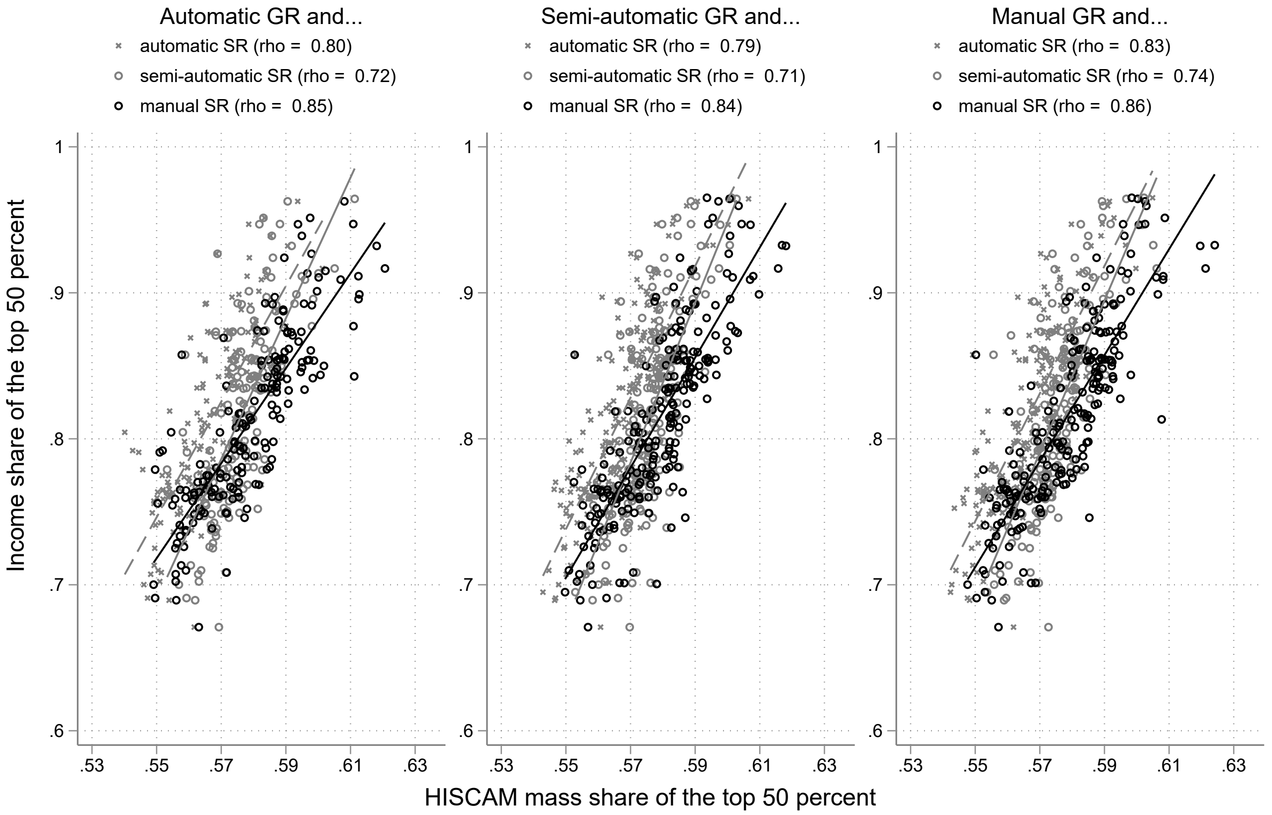


Note: Each panel shows a plot of tract-level (logged) average income against the (logged) average HISCAM score in our city directory data set. The number of tracts is 203 for automatic geo-referencing and 212/213 for all other graphs/correlations. The three panels represent different geo-referencing (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.

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 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 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.

Figure 7: Top income shares vs. top "HISCAM mass" shares



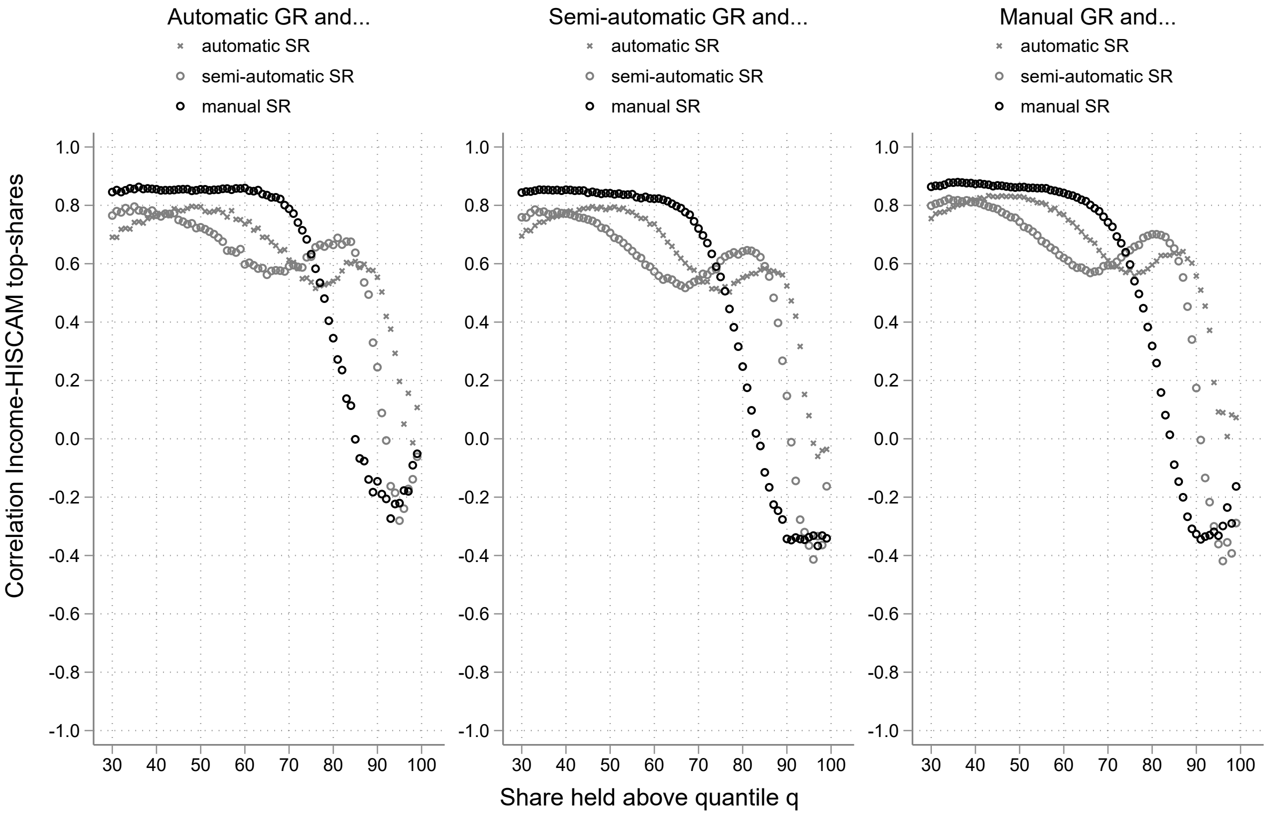
Note: Each panel shows a plot of the tract-level income share of the top 50 % against the top 50 %'s share in the "HISCAM mass" in our city directory data set. The number of tracts is between 168 and 211 depending on the combination of geo- and status-referencing. The exact number of observation varies 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.

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%, Figure 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 laborers. 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.

Figure 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.

Figure 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 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, top-shares against their respective administrative counterparts. The results suggest that directory data are of high-enough quality to be used as a substitute in the absence of spatially-disaggregated administrative 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 recognized entries over all entries. As discussed in our particular example, train the algorithm with ground truth data if considered necessary. In a second step, assess the likely trade-offs in geo-referencing. If the street names in historic maps of the city and modern maps overlap (as is likely in countries with a less tumultuous history than Germany), 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. Additionally, assess the likelihood of the low-income under-reporting bias' relevance with respect to your measure. It is most relevant for count and thus density data, but better geo-referencing only marginally improves comparisons of average incomes and within-tract distributions. 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, 2021). However, automatic and semi-automatic status referencing will yield satisfying results in most cases.

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 labor 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. 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 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 such as fires, wars, and earthquakes to cities 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, i.e. by testing the effects of social diversity on health outcomes (Kappner, 2021). 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 larger cities.

A second set of applications could improve our understanding of how different mobility 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 individualized 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. Yet, most importantly, micro-geographic data from city directories will allow us to understand the evolution of the economic structure of cities from an international comparative perspective.

A third set of potential applications pertains to urban upward mobility and inequality in the 19th century. City directories contain names and occupations. Because they are frequently published, one could identify rare surnames in order to track social mobility over time as is done elsewhere with probate records (Clark and Cummins, 2015). Our application also suggests that distributional measures derived from status data correspond to the income distribution in a reasonable manner. This could allow researchers to track urban inequality through time.

In addition to these fields of applications, future work should provide additional validation for the methodological findings of this study. Moreover, we discussed new tools that could further improve the work flow presented here.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez, "Automated linking of historical data," *NBER Working Paper No. 25825*, 2020.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf, "The economics of density: Evidence from the Berlin Wall," *Econometrica*, 2015, *83* (6), 2127–2189.

_ , Thilo N. H. Albers, and Kristian Behrens, "Prime locations," *CEPR Discussion Paper No. 15470*, 2020.

Anas, Alex, Richard Arnott, and Kenneth A. Small, "Urban spatial structure," *Journal of Economic Literature*, 1998, *36* (3), 1426–1464.

Bell, Samuel, Thomas Marlow, Kai Wombacher, Anina Hitt, Neev Parikh, Andras Zsom, and Scott Frickel, "Automated data extraction from historical city directories: The rise and fall of mid-century gas stations in Providence, RI," *PLoS One*, 2020, *15* (8), e0220219.

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

Blanchet, Thomas, Juliette Fournier, and Thomas Piketty, "Generalized pareto curves: Theory and applications," *Review of Income and Wealth*, forthcoming.

Bosker, Maarten and Eltjo Buringh, "City seeds: Geography and the origins of the European city system," *Journal of Urban Economics*, 2017, *98*, 139–157.

_ , _ , and Jan Luiten Van Zanden, "From Baghdad to London: Unraveling urban development in Europe, the Middle east, and North Africa, 800–1800," *Review of Economics and Statistics*, 2013, *95* (4), 1418–1437.

Brinkman, Jeffrey and Jeffrey Lin, "Freeway revolts!," *Federal Reserve Bank of Philadelphia Working Papers No. 19-29*, 2019.

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

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

Caesmann, Marcel, Bruno Caprettini, Hans-Joachim Voth, and David Yanagizawa-Drott, "Going viral: Nazi marches and the spread of extremism," *Mimeo*, 2021.

Cambon, Jesse, Diego Hernangómez, Christopher Belanger, and Daniel Possenriede, "tidygeocoder: An R package for geocoding," *Journal of Open Source Software*, 2021, *6* (65), 3544.

Chiswick, Barry R. and RaeAnn Halenda Robinson, "Women at work in the United States since 1860: An analysis of unreported family workers," *Explorations in Economic History*, 2021, *82*, 101406.

Clark, Gregory and Neil Cummins, "Intergenerational wealth mobility in England, 1858–2012: surnames and social mobility," *The Economic Journal*, 2015, *125* (582), 61–85.

Combes, Pierre-Philippe, Laurent Gobillon, and Yanos Zylberberg, "Urban economics in a historical perspective: Recovering data with machine learning," *Regional Science and Urban Economics*, 2021, p. 103711.

Cura, Rémi, Bertrand Dumenieu, Nathalie Abadie, Benoit Costes, Julien Perret, and Mau-

rizio Gribaudi, "Historical collaborative geocoding," *ISPRS International Journal of Geo-Information*, 2018, *7* (7).

Currie, Janet, Henrik Kleven, and Esmée Zwiers, "Technology and big data are changing economics: Mining text to track methods," *AEA Papers and Proceedings*, May 2020, *110*, 42–48.

Dahl, Christian M., Torben Johansen, Emil N. Sørensen, and Simon Wittrock, "HANA: A HAndwritten NAme database for offline handwritten text recognition," *CoRR*, 2021, *abs/2101.10862.*

_, Torben S. D. Johansen, Emil N. Sørensen, Christian E. Westermann, and Simon F. Wittrock, "Applications of machine learning in document digitisation," *CoRR*, 2021, *abs/2102.03239.*

Dittmar, Jeremiah E., " Information technology and economic Change: The impact of the printing press," *The Quarterly Journal of Economics*, 2011, *126* (3), 1133–1172.

Gallman, Robert E and John Joseph Wallis, *American Economic Growth and Standards of Living before the Civil War: National Bureau of Economic Research Conference Report*, Chicago: University of Chicago Press, 1993.

Geopy contributers, "geopy," https://github.com/geopy/geopy 2021.

Glaeser, Edward L., "What can developing cities today learn from the urban past?," *NBER Working Paper No. 28814*, 2021.

Gutmann, Myron P., Emily Klancher Merchant, and Evan Roberts, "'Big Data' in economic history," *The Journal of Economic History*, 2018, *78* (1), 268–299.

Heblich, Stephan and Walker Hanlon, "History and urban economics," *Regional Science and Urban Economics*, forthcoming.

_, Stephen J Redding, and Daniel M Sturm, "The making of the modern metropolis: Evidence from London," *The Quarterly Journal of Economics*, 2020, *135* (4), 2059–2133.

Heegewaldt, Werner and Peter R. Rohrlach, *Berliner Adressbücher und Adressenverzeichnisse 1704–1945: eine annotierte Bibliographie mit Standortnachweis für die "ungeteilte" Stadt*, Berlin: Helmut Scherer, 1990.

Hornbeck, Richard and Daniel Keniston, "Creative destruction: Barriers to urban growth and the Great Boston Fire of 1872," *American Economic Review*, 2017, *107* (6), 1365–98.

Kappner, Kalle, "Dense, diverse and healthy? Mixed-income housing and the spread of urban epidemics," *Mimeo*, 2021.

Knights, Peter R., "City directories as aids to ante-bellum urban studies: A research note," *Historical Methods Newsletter*, 1969, *2* (4), 1–10.

Königliches Statistisches Amt, *Berufsstatistik nach der allgemeinen Berufszählung vom 5. Juni 1882. 3. Berufsstatistik der Staaten und größeren Verwaltungsbezirke. Erster Theil. Statistik des Deutschen Reichs. Neue Folge. Band 4. Erstes Drittel*, Berlin: Verlag von Puttkammer & Mühlbrecht, 1884.

Lambert, Paul S., Richard L. Zijdeman, Marco H. D. Van Leeuwen, Ineke Maas, and Kenneth Prandy, "The construction of HISCAM: A stratification scale based on social interactions for historical comparative research," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2013, *46* (2), 77–89.

Ludwig, A., *Berliner Adreß-Buch für das Jahr 1875. Unter Benutzung amtlicher Quellen. VII. Jahrgang*, Berlin: Druck und Verlag der Societät der Berliner Bürger-Zeitung, 1875.

__ , *Berliner Adreß-Buch für das Jahr 1880. Unter Benutzung amtlicher Quellen. XII. Jahrgang*, Berlin: W & S Loewenthal, 1880.

Magistrat zu Berlin, "Verwaltungs-Bericht des Magistrats zu Berlin pro 1880," in "Beilagen zu Nr. 33" Communal-Blatt der Haupt- und Residenz-Stadt Berlin, 22. Jahrgang, Berlin: Julius Sittenfeld, 1881.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe, "OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings," *Applied Sciences*, 2019, *9* (22), 4853.

__ **, Uwe Springmann, and Frank Puppe**, "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" 2017, pp. 137–142.

Rose-Redwood, Reuben and Anton Tanter, "Introduction: Governmentality, house numbering and the spatial history of the modern city," *Urban History*, 2012, *39* (4), 607–613.

Schlegel, Inga, "Automated extraction of labels from large-scale historical maps," *AGILE: GIScience Series*, 2021, *2*, 12.

Shaw, Gareth and Allison Tipper, *British directories*, 2nd ed., Bloomsbury Publishing, 2010.

__ **and Tim Coles**, *A Guide to European Town Directories* A Guide to European Town Directories, Ashgate, 1997.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Lee, Jacob Carlson, and Weining Li, "LayoutParser: A unified toolkit for deep learning based document image analysis," *International Conference on Document Analysis and Recognition*, 2021.

Siodla, James, "Clean slate: Land-use changes in San Francisco after the 1906 disaster," *Explorations in Economic History*, 2017, *65*, 1–16.

__ , "Firms, fires, and firebreaks: The impact of the 1906 San Francisco disaster on business agglomeration," *Regional Science and Urban Economics*, 2021, *88*, 103659.

Spaan, Bert and Stephen Balogh, "city-directory-entry-parser," https://github.com/nypl-spacetime/city-directory-entry-parser 2021.

Spear, Dorothea N., *Bibliography of American directories through 1860*, American Antiquarian Society Worcester, Mass, 1961.

Straube, Julius, "Plan von Berlin mit Angabe der Sterblichkeitsziffer und graphischer Darstellung der Bevölkerungsdichtigkeit," map 1883.

__ , "Übersichtsplan von Berlin in 44 Blättern," map 1910.

Tesseract contributors, "Tesseract Open Source OCR Engine," https://github.com/tesseract-ocr/tesseract 2021.

van Leeuwen, Marco H.D., Ineke Maas, and Andrew Miles, *HISCO: Historical international standard classification of occupations*, Leuven: Leuven University Press, 2002.

von Gebhardt, Peter, *Die Anfänge des Berliner Adressbuches: ein bibliographischer Versuch*, Berlin: Selbstverlag, 1930.

Wick, Christoph, Christian Reul, and Frank Puppe, "Calamari - A high-performance tensorflow-based deep learning package for optical character recognition," *Digital Humanities Quarterly*, 2020, *14* (1).

Wiest, Ekkehard, *Gesellschaft und Wirtschaft in München, 1830-1920: die sozioökonomis-*

che Entwicklung der Stadt dargestellt anhand historischer Adressbücher, Vol. 3, Centaurus-Verlagsgesellschaft, 1991.

Williams, AV, *The development and growth of city directories*, Williams directory Company, 1913.

A Web Appendix

A.1 Sources

A.1.1 Berlin case study data

The Berlin city directory For the city of Berlin, city directories (*Adressbücher*) were published between 1799 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.[17] 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 neighboring districts. We also use the 1875 directory (Ludwig, 1875) to train our OCR model.

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, 2021, p.59, for further details). To map HISCO codes to the HISCAM social status scale, we obtained a crosswalk from the HISCAM website, applying the most-robust "men-only universal scale" as recommended by the authors (Lambert et al., 2013).

A.1.2 Validation data

In our validation exercise (Section 3.2), we employ high-quality administrative data reported on the level of 216 census tracts (*Stadtbezirke*).[18] We geo-reference them using a historical

[17] 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.

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

map of tract boundaries (Straube, 1883). We get the number of households per tract from the 1880 census, reported in Böckh (1883, 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, 229–236) and the magistrate's administrative report (Magistrat zu Berlin, 1881, 18–25). For more information on the income and tax data, see Section A.2.

A.1.3 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 et al., 2013; Bosker and Buringh, 2017). 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).

A.2 Estimating top-shares

A.2.1 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)[19] 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).

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.

[19] 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.

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

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

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:

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

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 generalized Pareto interpolation suggested by Blanchet et al. (forthcoming) to estimate the top-shares.

A.2.2 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.