An Historical Income Panel for the Netherlands, 1850–1920.

Auke Rijpma
Eva van der Heijden
Alex Nagel
Paul Puschmann
Rick Schouten

This version: April 2025

Abstract: This paper presents the Historical Income Panel of the Netherlands (HIPNL), a large scale historical database that follows the incomes and key demographic variables of hundreds of thousands households over the period 1850-1920. The data can be integrated with the overall Dutch microdata infrastructure based on civil registries and population registers, allowing the welfare of these household members to be studied from a multidimensional perspective, and in turn enriching these databases with key economic indicators. This paper describes the HIPNL database, outlining the overall data pipeline, discussing the municipal income taxes it is based on (the *Hoofdelijke Omslagen*) and how the data was sampled and collected using both manual and automated processes. We also address the record linkage needed to link the tax data to external sources, and introduce two case studies on inequality and migration to show the potential of the database

Introduction

There has been a growing interest among economic and social historians in topics like welfare, inequality, and social mobility. Scholars like Piketty (2014), Alfani (2021), and Scheidel (2017) have put the topic of economic inequality on the top of the research agenda. Social mobility research has seen a great impetus in the work of Clark and Cummins (2014), as well as Chetty et al. (2014). These topics are all the more relevant given current societal concerns about exactly these points (see Deaton 2024 and the articles collected in that special issue).

These questions also make new demands on the kind of data we work with. Concretely, they require us to measure economic outcomes such as income or wealth, at the individual level, at multiple points in time, and for multiple generations. Ideally, we should be able to view welfare from a multidimensional perspective (Van Zanden et al. 2044; Rijpma et al. 2024).

Collecting and linking this kind of data is time and resource intensive, however. Moreover, much of the literature so far has relied on imperfect measures of economic welfare, such as occupations in the case of social mobility; or proxies such as house rents. Ideally, we directly

The Netherlands are a good example of these challenges. Historical demographers have used large-scale databases such as the HSN and LINKS to reconstruct demographic behaviour in the late nineteenth and early twentieth centuries (Mandemakers 2000), a key period encompassing the demographic revolution, as well as massive changes in the economic and political structure of the country.

However, the economic dimension has not seen similar high-resolution data. There are world class aggregate statistics such as the nineteenth-century time series collected for the Dutch Historical GDP estimates (Van Zanden and Van Riel 2004; Smits et al. 2014; Van Riel 2021). From 1917 onwards there are inequality reconstructions such as those by Salverda and Atkinson (2007), based on national income tax tabulation. Prior to 1917, there are tentative reconstructions by Soltow and Van Zanden (1998), which rely on a wide variety of proxies. However, systematically collected microdata on the economic dimension of welfare for this period does not yet exist.

Internationally, a similar picture can be seen (Ruggles 2012). Demographic databases are the state of the art in terms of coverage, richness, and the ability to study long-term change. The economic dimension, however, is not part of these data sources other than in the form of occupations. Historical census microdata is a hugely popular resource among historians and social scientists, but generally lacks information on economic welfare (the US after 1940 being the exception). The main exception to this overall picture is the “Swedish Incomes, 1870–1950” project (Bengtsson and Molinder, 2024; forthcoming).

This paper presents the Historical Income Panel of the Netherlands (HIPNL), a large scale historical database that is designed to address these lacunae. The database follows the incomes and key demographic variables of hundreds of thousands households over the period 1850–1920. The data can be integrated with the overall Dutch microdata infrastructure based on civil registries and population registers (Raad et al. 2020; Mandemakers 2000; Van den Berg et al. 2020), allowing the welfare of these household

members to be studied from a multidimensional perspective, and in turn enriching these databases with key economic indicators.

This paper introduces the HIPNL database. After outlining the overall data pipeline, we next discuss the municipal income taxes it is based on (the *Hoofdelijke omslag*) and how we sampled collected this data using both manual and automated processes. We next turn to the record linkage needed to link the tax data to external sources and to subsequent , which given the sparsity and heterogeneity of our source data is a big challenge for the process. After describing the resulting dataset, we also introduce two case studies on inequality and migration on the basis of the database.

Overall process

The overall process to process the tax data is presented in figure X below. Archival tax records are our starting point, which we scan and store. These are transcribed into spreadsheets forms. Different methods for this were used, as better technology allowed us to replace manual transcription (which did the bulk of the data) by ever better automated tools. These tax spreadsheets are then cleaned and harmonised into a tax database. To link our data, we rely on the population registers made available by archives. These are typically provided as XML data, which we harvest and harmonise.

The next step is to link these two datasets so that each household in the tax registers can be connected to a name-place of birth-date of birth X of the household head in the population registers. This combination usually uniquely identifies individuals, and thus allows us to link households over time and to other data sources such as the civil registries. We combine the two datasets into a candidate set that contains plausible links. We manually identify true links in a sample from this data, and use this ground truth data to train a model that can predict links.

Based on these predictions, we create a version of our tax database with the demographic information allowing for further linkage. In turn, this database is used to create a Linked Data version of the database enabling further integration into the overall Dutch microdata infrastructure (Hoekstra et al. 2016), as well as data extracts such as historical income scores for occupations (HINCO).

The final dataset thus created spans a period of X years, and contains Y households, measure a total of N features. All data is openly available on Utrecht University's YODA platform[1] as well as the CLARIAH triplestore.[2]

[1] https://i-lab.yoda.uu.nl
[2] https://druid.datalegend.net

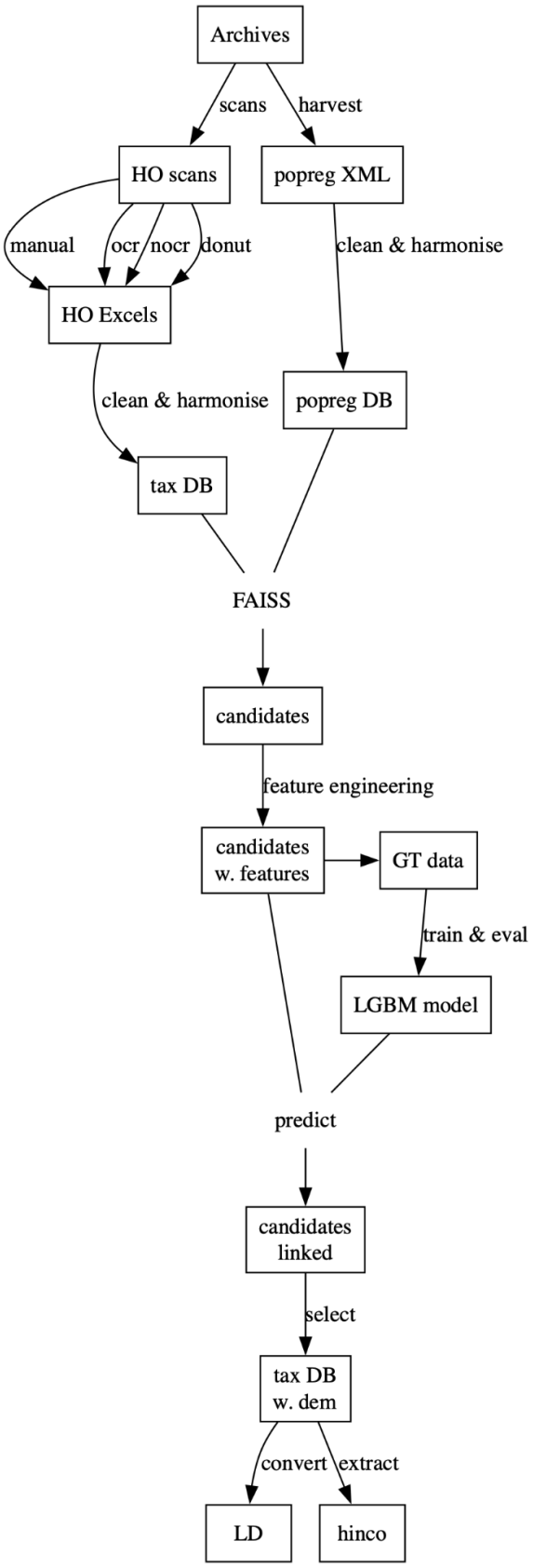

Figure x. The HIPNL data pipeline.

The hoofdelijke omslag taxes

The core sources for the Historical Income Panel for the Netherlands are municipal income tax registers. Taxation by the municipal level legislature in the first decades of the 1800s was split into local excise duties and taxation on assets – such as housing, furniture, and personnel – and land. The *Gemeentewet* (Municipalities Act) of 1851 provided the legal framework that allowed municipalities to levy income taxes locally. After the local excise duties were abolished in 1865 (Staatsblad nr. 79), local income taxation gained importance. The tax registers of municipal income taxation are often called the *Hoofdelijke Omslagen* (HO). Though HO taxes existed from the early 1800s, from 1865 onwards we can see the number of tax registers on the municipal level increase substantially (see Figure x), and the same holds for the number of observations in the tax registers (see below). As of 1920, municipal income taxation became a responsibility for the national governement, hence the stark decrease of hoofdelijke omslagen from 1920 onward (Klep et al. 1982: 7).

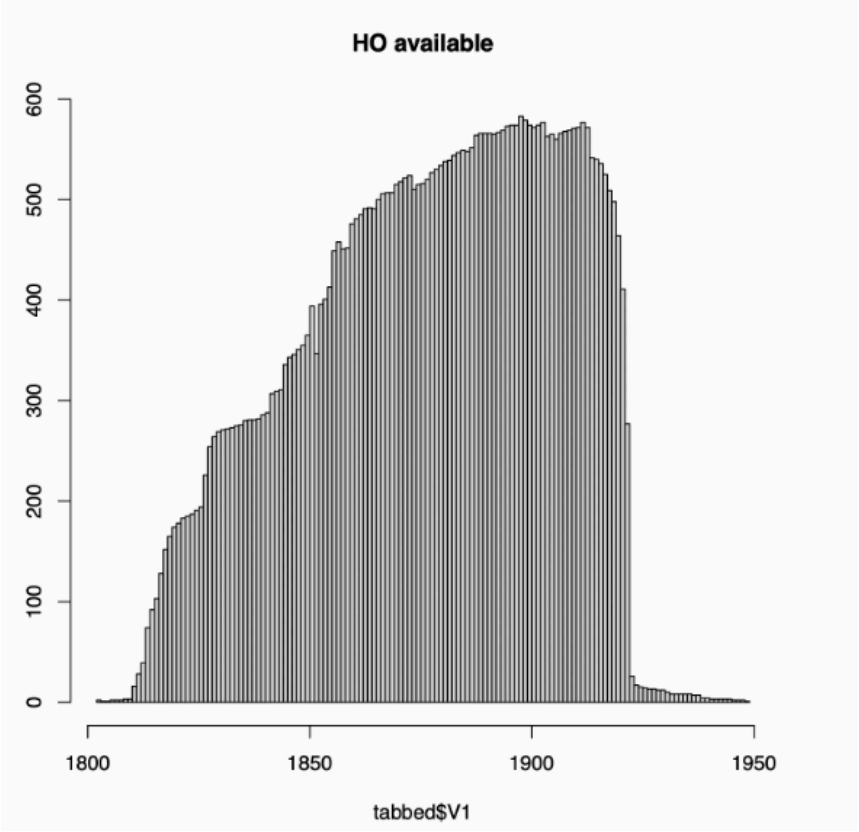

Figure x. Number of municipal archives holding HO tax registers, 1800–1950.

The exact implementation of the HO tax was left to the municipalities, including the decision whether to levy it at all. Moreover, great variation existed in the tax procedure and reporting which makes processing and harmonising the HO taxes challenging. That said, some common characteristics can be identified.

The registers recorded taxation based on personal or household annual income and are commonly self-reported and at times estimated by municipal officials. Registrations in the

hoofdelijke omslagen are on the individual's name and usually, the head of the household is listed. Often this means we can consider the HO a household tax, though for some municipalities and periods individuals making an independent income within the household are listed.

The income information reported in the registers varied between municipalities and years. The bare minimum of reported income information concerns the taxes levied as well as the name of the head of household and some basic location information, i.e. street plus house number or the neighbourhood the household resided. In other municipalities and particularly in later years, the *hoofdelijke omslagen* included income measures, demographic data relevant to tax deductions – e.g. the number of children or marital status – and occupations. Income measures can vary from gross income, taxable income (gross income net of deductions), unspecified income or income classes. The first three are – if present in the register – reported in guilders, whereas the income class reports the stratified income brackets. The example of Utrecht in 1919 (figure x below) is an example that provides detailed information on the location, demographics, occupation, and vitally the incomes. Most tax registers reported at least the names, taxes levied and incomes, such as in Doesburg in 1879.

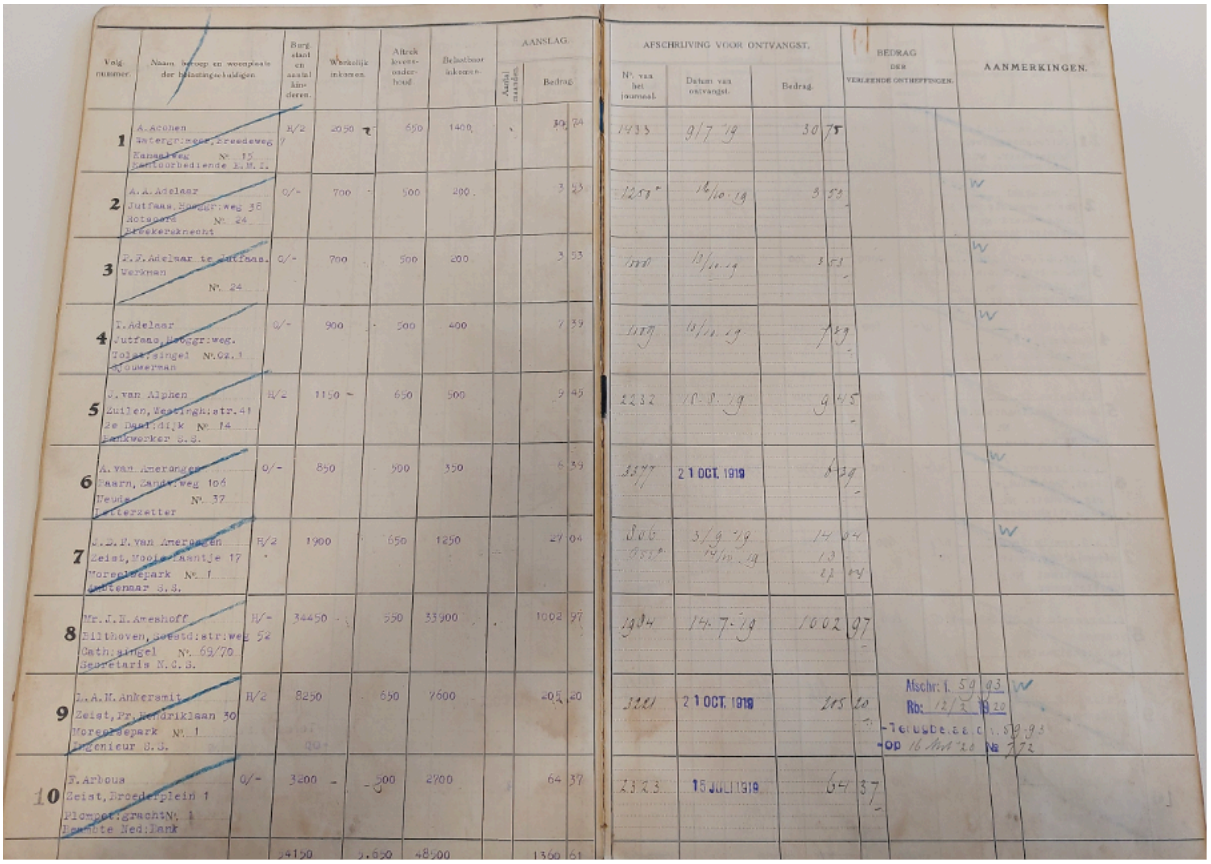

Figure x. Tax register for Utrecht in 1919.

Incomes are commonly reported by the taxpayers themselves, especially in later years. In earlier years we also see that municipal officials estimated the incomes ("*vermoed inkomen*" – suspected income). Often there was a system of deductions, where a basic subsistenc income could be subtracted from the gross income. In addition, there could be a deduction

for each child under a given age. The most straightforward implementation of deductions were reporting gross income, the deductions in fl., and the taxable income. In some cases the registers do not make clear whether deductions have been applied.

The system of deductions means that the HO tax was progressive, where low-income households paid a smaller percentage over their gross income. In addition, we find that the progressive tax rates were applied on top of the deductions. Earlier tax years, especially in the 1850s and 1860s, often only report the taxes due, and not the underlying incomes and deductions. Finding out whether these early taxes were also progressive is one of the key challenges to using the HO taxes over the entire period.

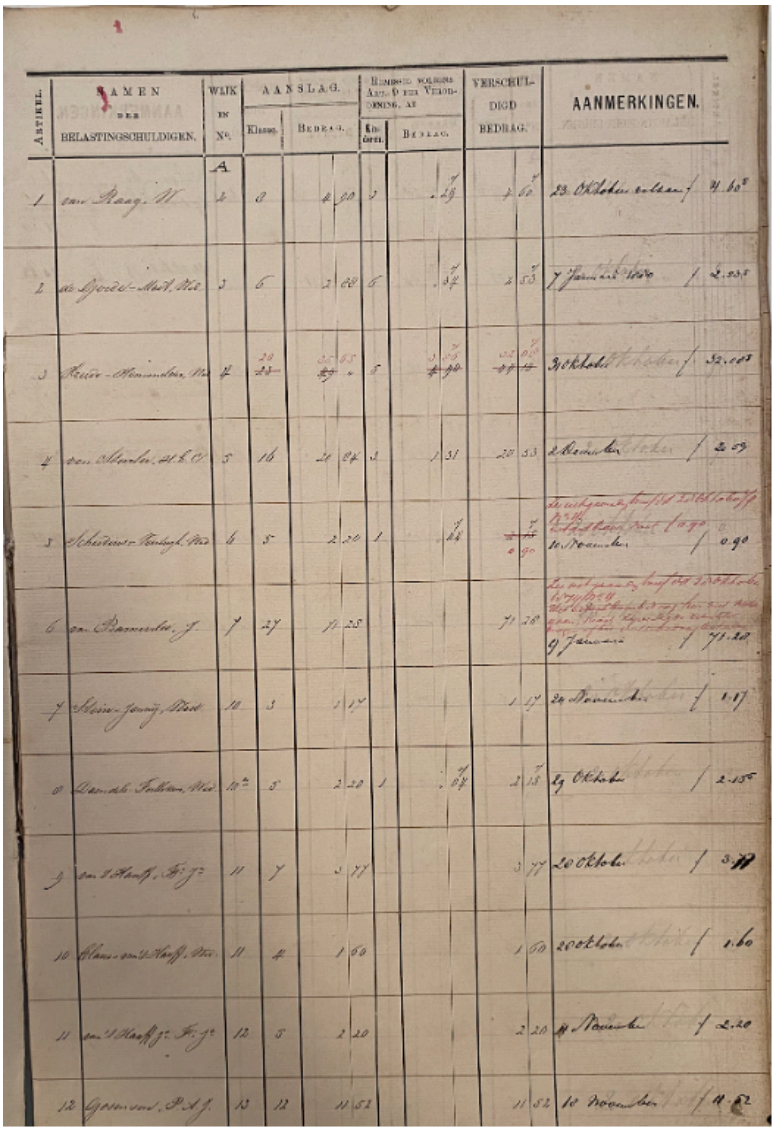

Figure x. Hoofdelijke Omslag tax register of Doesburg in 1879

The registered individuals are by no means a representative sample of the population of the respective municipality. First, the male head of household often is registered as paying taxes, most commonly this comes. Married women earning an income are barely registered in the hoofdelijke omslagen; the majority of women in the tax records are widows. The tax unit in the HO is often the household. A deviation from this rule is that older – often unmarried – children can be registered to pay taxes independently from their parents. This

particularly occurred in the case of widowed mothers with unmarried sons living at the same address. In rare cases married women are also registered. Another exception to the household rule is the case of siblings earning separate incomes and thus each paying taxes and living at a singular address. Rarely, institutional households are found in the hoofdelijke omslagen. This can be a larger household with multiple people earning an income but paying income taxes jointly, and an example of this is an abbey or convent where the head of household is taxed. Another option of an institutional household is a larger group of individually taxed people that are registered at the same address, such as the women registered paying taxes individually at Agnietenstraat 2 in Utrecht. A psychiatric home and hospital for women – particularly from wealthy families – was housed at this address.

Second, there was an income tax threshold, which resulted in excluding households living at or below subsistence from taxation. The threshold varied by municipality, and ranged between fl. 200–500. The implication of this is the exclusion of a sizable share of the population earning an income below this subsistence level. The households present in the tax registers are therefore the relatively well-earning share of the population. To correct for this, a reference population is needed. The most straightforward approach is to use the total number of households for a city from the Dutch censuses and municipal reports, estimated from the total population if the number of households was not reported (Boonstra et al. 2003).[3] The share of the population covered by the HO taxes in the HIPNL database is shown in figure x below. For most of the nineteenth century, coverage was about 60%. By the turn of the twentieth century, coverage started to increase, reaching more than 90% by 1920. This is exceptionally good coverage for this period. Many income taxes around this time covered little more than half the population (e.g. Roikonen and Heikkinen, 2018), including the Dutch national income tax of 1917 (Salverda and Atkinson 2007). For microdata analyses, the reference population estimates can be used for bounding exercises on the estimates. When high-quality population registers are available, they can be used to identify non-taxed individuals, allowing researchers to obtain income estimates for all tax units in a municipality. The linking process and the requirements on the population registers are discussed below.

[3] Doing this does lose information in municipalities where independent income earners who were not household heads were taxed. In these cases more detailed information from Dutch historical censuses could be used, including counts of individuals living independently, non-kin household members and working-age population.

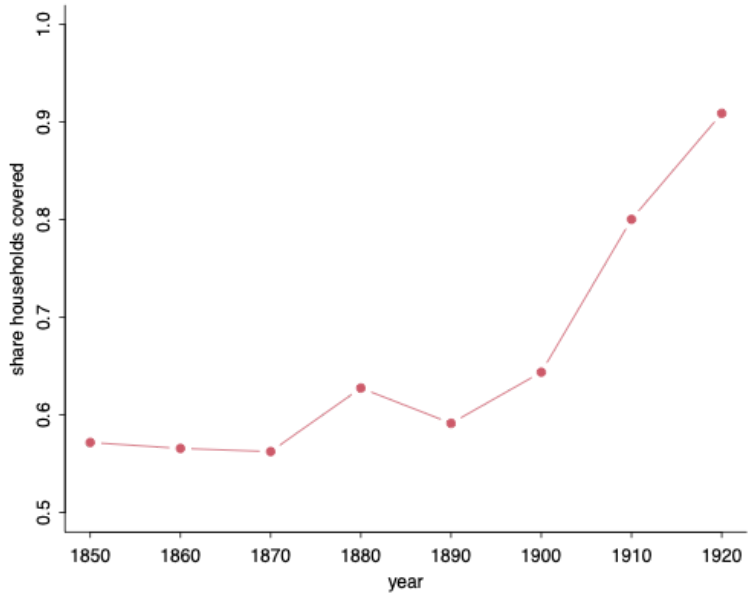

Figure x. Coverage of the HO taxes, 1850–1920.

Sample design

Our sample of Hoofdelijke Omslag taxes span over seventy years (1849–1921). Sampling started in 1851, and then continued at ten year intervals from 1859 onwards to coincide with the Dutch censuses. We aimed for a sample of municipalities that would cover 10% of the Dutch population. We sample municipalities for three reasons. First, the HO taxes are located in municipal archives, and sampling individuals would have forced us to visit every archive. Second, sampling municipalities allows us to get a grip on the reference population. Third, sampling within municipalities allows us to follow those people who did not move over time, turning the data into a panel. The original sample is stratified by geography (province) and population size, see figure x. The municipalities with substantial population sizes were hand-picked to ensure the availability of HO registers. The cities of Utrecht, Leiden, Nijmegen, Enschede, and Eindhoven were included for this reason. Amsterdam was excluded because at the time the sample was designed, its civic registers were not yet digitised, limiting the degree to which that data could be integrated in the larger Dutch microdata infrastructure. Beyond the large, handpicked cities, municipalities were selected at random within provinces and within municipality size strata (<5000, 5000-10000, 10.000–50.000), after a quick check into the digital availability of the population registers for the respective municipalities. This check meant that the northern province of Groningen and most of Overijssel had to be excluded from the sample, as there were no digitised population registers available. The availability of the population registers is important for further use of the income panel. The tax registers provide information on the name and address or location whereas the population registers include name, address/location, and year of birth, which is necessary to ensure unique identification of observations and to allow for further linking.

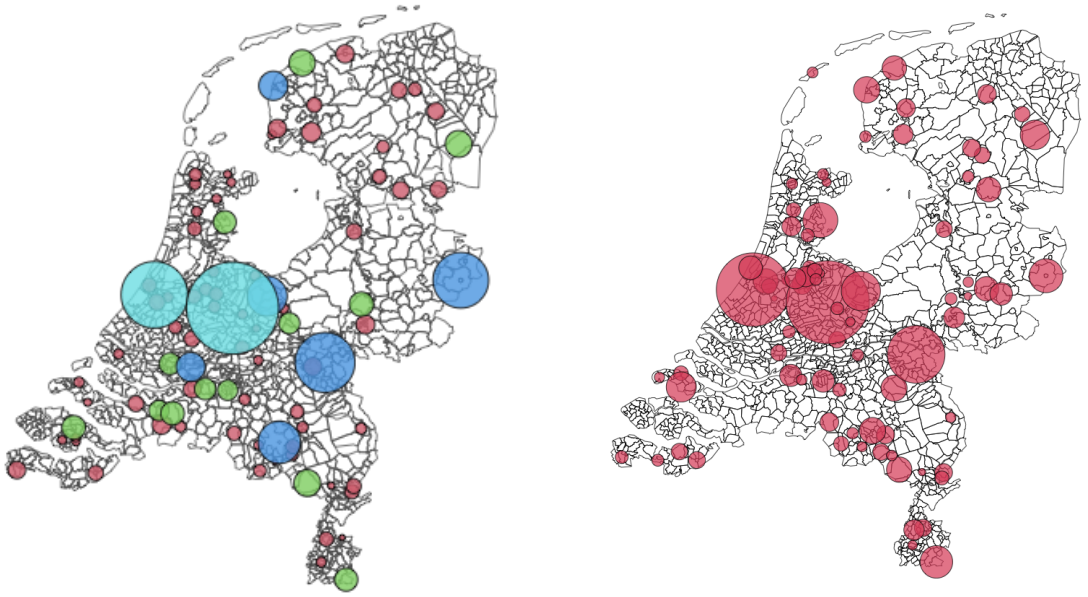

Figure x. Municipalities in the planned (left) and realised (right) HIP-NL sample. Circles are proportional to the municipality's population size in 1899.

A year into the project, we discovered that for about forty of the municipalities, there were no population registers digitally available - despite the initial check for the availability of the population registers. This happened because the reported event place in the population registers did not refer to the event of population registration, but to the birth place. Especially for smaller municipalities this could be misleading. This turned out to be a challenge in the province of Limburg, where population registers rarely were digitally indexed or published. The original sample was adjusted after searching for the availability of population registers. In case the population register was present, the municipality remained included in the dataset. In case the population register was absent, an alternative municipality – yet similar in population size in 1900 and located in the same province – was listed to be included in the adjusted sample. The municipalities that were already entered into the dataset but had missing population registers thus far remain in the dataset but cannot be linked.

Data collection

The data collection is conducted primarily by research assistants. They visited the archives to scan the Hoofdelijke omslagen. HO registers were scanned from cover to cover to include potential key information on the tax rules or tax schedules on the first or last pages of the book. The scans were made for every tenth year from 1859 onwards, i.e. 1859, 1869, 1879, 1889, 1899, 1909, and 1919.[4] In case of missing books for a specific year, the nearest year was scanned. In the case of missing hoofdelijke omslagen in a municipal archive, an alternative municipality was sought of similar population size and located in the same province.

[4] Taxes for c. 1850 turned out to be rare.

Most municipalities have multiple years of observations, but data availability does increase over time (figure x). Note that most of the rise in the number of observations (a fourfold increase between 1860 and 1920) can be explained by increasing population coverage in each municipality (nearly doubling see above).

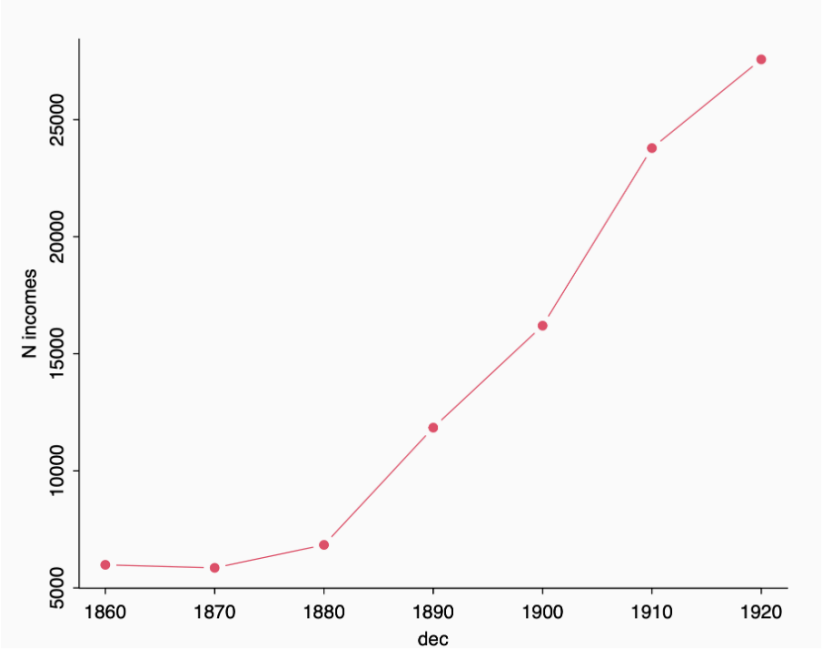

Figure x. Number of taxed households

An overview of the variables in the dataset can be found in Appendix A. The research assistants were provided with an entry form that stuck close to the typical HO tax register for speed of entry and adherence to the original. The information reported in the registers that we transcribed includes of course the income (gross, taxable, or unspecified), deductions, and taxes due. In addition surname, initials or first names, address, neighbourhood, and titles of the head of the household or tax-paying individual were included. Details about children and marital status are also provided. Furthermore, occupation is reported in some of the municipalities. We can often infer the sex of the taxed individual from the name, titles such as widow, ms, or mrs, or the given name itself. In some municipalities, the number of months the tax was due is also reported for residents who did not live there for the entire period. Finally, while information about the actual payment of the tax is usually present in the tax registers, we did not include this information in our transcription.

A variable that was added over the course of the project is ‘moved_address’. This variable indicated whether someone moved; in case the new address was reported this observation was entered under the original address and with the ‘moved_address’ variable at ‘1’, and this entry was duplicated – i.e. same name, unique identifier etc. – yet the new address is entered for the regular address variable. Due to these iterations of the form, the current dataset has staggered entry of new variables, i.e. the municipalities that were entered in the early stages of the project contain fewer variables than those entered in the later stages.

Originally, 90 municipalities were in the sample. Of these, 41 turned out to have missing population registers. For 35, replacement municipalities were found, with the remaining issues concentrated in the province of Limburg.

Data collection was prioritized in three ways. In the first year of the project, as many municipalities as possible were covered without a specific regional dispersion and with a focus on accumulating scans over data entry of tax registers. This resulted in a subsample concentrated largely in the South of the Netherlands – i.e. the provinces of Limburg, Noord-Brabant, and Zeeland. In the second year, the data collection shifted focus to entering the data from the gathered scans of the tax registers, particularly for larger cities like Utrecht and Amersfoort. Simultaneously in the second year, the challenge of missing population registers meant certain municipalities had to be swapped for alternatives. In the third year, the spatial dispersion and population size were emphasized in prioritizing places, therefore municipalities in regions that were underrepresented in the current subsample were scanned and entered with priority. This meant a focus on the provinces of Gelderland and Overrijssel as well as Friesland, as well as entering smaller municipalities surrounding larger cities (in particular those surrounding Utrecht or Leiden).

Automatic transcription

In addition to the manual input of tax registers, we have also explored methods for the automated transcription of both printed and handwritten registers. In this process, we initially utilized a traditional Optical Character Recognition (OCR) pipeline and later transitioned to a locally finetuned Visual Document Understanding (VDU) transformer model (DONUT, see Kim et al. 2022), to finally move to a few-shot cloud-based VLM approach. The latter two approaches, which we call NOCR, offer significant advantages. They do not try to transcribe every piece of text as literally as possible but rather extract information directly from a document. Consequently, it does not require labour-intensive training datasets with text boxes and literal transcriptions of the text therein, as is necessary for traditional OCR models. Moreover, post-processing of the model output is much reduced. We describe our approach in more detail in a separate paper (Rijpma et al. 2025), including benchmarks and best practice. Here we briefly describe the three approaches.

Traditional OCR

For the printed tax registers of Leiden, we initially explored a traditional OCR pipeline. We employed an OCR and text recognition model from EasyOCR,[5] which we fine-tuned specifically for the Leiden registers. This choice was made because the text recognition model performed well with our data and because the accompanying python package made it easy to finetune and deploy the OCR model.

To collect training data for fine tuning the model, we had the base model predict several pages, which we checked and corrected if necessary. This approach allowed us to quickly gather a substantial amount of training data. Additionally, we created synthetic training data by placing names, amounts, and house numbers in fonts resembling those in the Leiden registers onto backgrounds derived from these same registers.

Accurately recognizing the table structure required that the pages be perfectly straight. The text recognition creates text boxes around each piece of text, with each piece of text comprising only one line. Moreover, it only occupies one column if there is a significant white space following the text. Consequently, it is possible to determine the table structure afterward, as text boxes at the same height likely belong to the same row of the table. Text boxes of approximately the same width that are aligned at the same position on the horizontal axis are likely part of the same column.

In addition to ensuring the pages were not rotated, it was crucial to remove all warping caused by the bound nature of the booklets from the scans. We succeeded in finding a solution for this, but it significantly limited the model and pipeline's application to the Leiden registers, requiring recalibration for each new booklet.

Although this approach yielded reasonably good results, the entire pipeline became quite complex. Over time, additional challenging situations arose, such as names spanning two lines or text boxes encompassing two or three lines, while the model could only recognize

[5] https://github.com/JaidedAI/EasyOCR.

one line of text at a time. These issues were manageable but further limited the application of the method to Leiden, implying that considerable effort would be needed to adapt it for other municipalities.

DONUT

Due to the aforementioned reasons, we sought a method that would be capable of extracting information directly from a page without intermediary steps, regardless of any page distortions, and without the need for complex post-processing steps. As the registers of many of the big municipalities – Leiden, Nijmegen, and Enschede – were printed, getting a better OCR pipeline was worth the effort. Because of the potential applications for handwritten tax records, it was also crucial that a model could be directly trained to produce the desired output. For handwritten tax records, we lack datasets with text boxes, having only the transcriptions. Moreover, a significant portion of the handwritten records were photographed with a handheld camera rather than professional scanning equipment, thus exacerbating the problem of distortions.

Ultimately, we opted for an OCR-free model named DONUT (Document understanding transformer, Kim et al. 2022). A significant advantage of this model is that it incorporates a language model to generate the output. Unlike a traditional OCR model that classifies each letter, this model is capable of 'interpreting' the image. It is an example of a Vision Language Models (VLMs) that combine their image understanding and language modeling components. Unlike traditional OCR systems that might pipeline distinct visual analysis and text processing stages, VLMs often employ a more unified approach. The image understanding module processes the input image, identifying potential text regions and recognizing character shapes, spatial arrangements, and layout. Simultaneously, the language model component leverages its knowledge of written text, and real-world context. This integration allows the VLM to not only decode the visual characters but also to use linguistic context to disambiguate visually similar characters, correct recognition errors based on probable word sequences, and interpret the text within the broader visual context of the image. Moreover, these models can be trained to answer questions about an image, simply transcribe the text from a page, or even return a table from an image directly as a table.

We trained this model on corrected output of RAs and the traditional OCR. Notably, the output are the direct tables we want to extract from the documents. This approach was applied to the Leiden and Nijmegen records. RAs checked the output for mistakes.[6]

NOCR/Gemini

While the DONUT approach was very promising and an improvement over traditional OCR approaches, Finetuning a VLM is still time-intensive. What is more, while we could directly leverage the tabular training data from the RAs to fine-tune the model, a new fine-tune did

[6] In addition we developed a training dataset for handwritten records, but this quickly became obsolete.

seem needed for every municipality. Fine-tuning a model that could understand generic HO tax registers seemed possible, but very time-intensive.

We suspected that general methods may have already superseded our case-specific approach based on a 2022 model, in turn based on the 2018 BERT and 2021 Swin architectures (Kim et al. 2022; Devlin et al. 2018; Liu et al. 2021).[7] An exploration of a number of models showed that the Gemini models (Gemini Team 2023) performed very well on our tasks, requiring little more than a good prompt describing what we wanted the model to extract from the scan, and sometimes a few examples of what the output should look like (that is, we use zero- and few-shot prompting).

Through Google's generative AI API, we used this model to create the data for the printed records in Enschede, the typewritten records of Utrecht, and – notably – the handwritten records for Utrecht 1899. As the quality of the output was very high, we decided to redo the Leiden and Nijmegen printed records as well. The model output was checked by the RAs, but this process was far less time-intensive than full transcription. We think this approach holds great promise for future digitisation processes of all but the hardest handwritten structured sources. We detail it in a separate paper (Rijpma and Schouten 2025).

Record linkage

Linkage of the households represents a crucial step in the construction of the HIP-NL data. It enriches the sparse taxation data with "demographics," nearly always including date of birth, place of birth, and sex, and sometimes marital status, household composition, and address. By linking households to a common observation in the civil and population registers, we connect the households in the taxes over time, effectively turning our tax data into a panel. Finally, this linkage allows our data to be integrated into the Dutch historical microdata infrastructure, above all LINKS and the HSN.

Our primary aim is to link the tax observations to the population registers, and sometimes their successors, the family cards (though these are rare and we rely on essentially the same information as we get from the population registers). The population registers record all the residents of a municipality over a period of time, which reduces the search space for individuals and makes it possible to uniquely identify the individuals in the HO taxes. In our experience, households in the HO taxes can nearly always be found in the full population registers, as the combination of name and address in the HO taxes uniquely identifies them.

However, automatically linking these sources is much more challenging. There are two main reasons for this. First, we want to allow for variation and small mistakes in records. This is a familiar record linkage problem (Christen 2012), and it has been dealt with successfully in social, demographic, and economic history using many different approaches (Rijpma et al. 2020; Abramitzky 2021; Raad et al. 2020).

Our biggest problem, however, is that we have to deal with extreme data heterogeneity and sparsity. As discussed above, the information contained in the tax registers varies from one

[7] On the growth in efficiency of general methods, see Sutton (2019), "The Bitter Lesson" at http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

municipality to the next; full names, occupations, streets, and house numbers are not always reported. Furthermore, we rely on volunteer transcriptions of the population registers. While these registers and their transcriptions are an amazing resource that should in principle allow us to uniquely identify most individuals in the HO with the full information they contain, the information transcribed varies from one municipality to the next. Some volunteer transcriptions only report names and date/place of birth,[8] meaning we can only rely on the name and age at the time of the HO tax for matching, which is challenging. In others, we also obtain address or occupation information, making matching easier.

Processing the population registers is also challenging, as they are hosted on a number of repositories and stored as XML with the data organized differently from one archive organisation to the next. We retrieved the population registers stored in XML data from the archive repositories and then, due to some archives having poor search functions, cross-referenced our XML data against the listings scraped from archive websites. We then retrieved any missing registers, in some cases this necessitated scraping archive websites directly, in other cases we were able to retrieve XML data, from the same repositories, which we had initially missed. Once we had a complete dataset we standardized column names and register data. The exact operations that we applied varied per digitization service used by the archive, common operations included: standardizing columns between archives, splitting names into first names, infixes, surnames, etc., splitting data stored as a 'remark' (such as an individuals address, widowhood or partner information, occupation, etc.) into separate columns, and standardizing recorded dates (e.g. dates of birth, departure, and arrival dates). We also created unique identifiers corresponding to pages (where possible we used scan id's) and municipalities (using the province, municipality, as well as start and end years of the register). We also had to train a model which could distinguish whether a set of pages was ordered in order to ensure that as much of our data as possible was in the same order as the original registers.[9] The final result was a set of population registers that, aside from metadata generated during the collection and harmonization process, reflected the registers sources as closely as possible and allowed for easy and efficient linking to specific municipalities and periods for eventual linking.

With both the HO tax data and the population registers in place, the linkage process takes the following steps. First, we employ blocking using FAISS (Facebook AI Similarity Search; Douze et al. 2024) on text embeddings of a combined string of shared identifiers in the HO and population register data, made using the multilingual E5 sentence encoder (Wang et al. 2024). This string at the very least contains initials and surname, but we use more if it is available for both sources. We keep the records of the top 20 most similar strings to create our candidate blocks. This "semantic blocking" offers significant advantages over conventional string distance methods. Specifically, blocking on embeddings allows us to capture semantic similarity between records, even when there are variations in spelling (Maria and Marie), abbreviations (straat and str.), or the use of synonyms or diminutives (Willem and Wim). Unlike traditional string distance metrics that focus on exact character

[8] The term used for these sparse transcriptions is indexation. The resulting indices are meant to facilitate searching for people, not provide full research-grade data.
[9] Our model works as follows: split each page into a separate group, calculate two 'age scores' (done by averaging the differences between each person's birth date) for the page in its original and randomized order, then calculating a host of second-level statistics for the group of pages (average score, min score, difference between average original and randomized scores).

matches, embedding-based blocking groups records together based on the underlying meaning of the information, leading to more effective and robust candidate link generation, especially in the presence of noisy historical data.

Second, a portion of this blocked training data was sampled to create Ground Truth (GT) data, where our team manually labelled potential links as either correct or incorrect. We used Label Studio to organize this annotation process and incorporated further information from the scan to ensure the links were accurate, thereby preventing the training of our model on erroneous ground truth. This time-intensive step is crucial for the accuracy of our linkage, and we share our GT data with the research community to improve on our approach.

Linking pilot: Utrecht 1909

For the development of a linking model, we first conducted a pilot study to explore the linking challenge in depth. We chose Utrecht 1909 due to its high data availability and quality, which included full addresses in both the tax registers and population registers. Being the largest city in our sample, achieving accurate linkage for Utrecht would furthermore be a significant accomplishment in itself.

The original Utrecht pilot used 1000 individuals from the tax registers to make GT data. In addition, we specifically sampled some complex cases, particularly widows and women, to enrich the GT data and ensure sufficient representation of these challenging groups in ourdata. This GT data was then used to train and evaluate a LightGBM model specifically for Utrecht (Ke et al. 2017). The train/test split was 70/30. The following features were used to train the model:

feature	**Description**
name_initials_cos_sim	cosine similarity between query and candidate initials
address_cos_sim	cosine similarity between query and candidate address
surname_sim_ratio_fuzz	Normalised Indel distance between query and candidate surname
Age	Candidate age
initials_sim_avg_fuzz	Average of the ratio and wratio score on the query and candidate initials
street_cos_sim	cosine similarity between query and candidate street
address_sim_wratio_fuzz	Wratio score between query and candidate address
house_nr_street_sim_ratio_fuzz	Normalised Indel score between query and candidate street number
is_junior_tax	Boolean: is query junior in tax

surname_sim_partial_ratio_fuzz	Optimal string alignment score between query and candidate surname
same_seniority	Boolean: query and candidate are both junior or senior.
is_senior_pop_not_junior_tax	Boolean: candidate is senior and query is not junior
is_senior_tax	Boolean: query is senior
is_junior_pop	Boolean: candidate is junior
same_gender	Boolean: query and candidate are both same gender
female_tax	Boolean: query is female
female_pop	Boolean: candidate is female

Postprocessing steps?

We accept links above a threshold of 0.437. The model's performance was extremely strong, as indicated by the test set precision of 0.99 and recall of 0.99. The group correct score is 98.5%. When evaluating the performance, we focus in particular on the “group correct” statistics, which measures for each candidate block whether it was correctly predicted, meaning either a link has been predicted to a correct individual in the population registers, or it has been correctly predicted that there was no true link in the candidate block. Typically, individuals have multiple entries in the population registers because they are expected to have moved house multiple times within the 10+ year period covered by a single register (see Kok et al. 2005 on the high frequency of changing residence in this period). While we labelled the data and trained the model to identify the absolute best match, it is acceptable if the model predicts a less optimal match that still corresponds to the same individual. This is reported as [group%], and is the main performance metric we track.

The main reasons this model performs so well are (i) the large amount of information contained in the combination of the name and address info that we can link on; and (ii) the ability of the gradient boosting framework to capture the richness of the data in terms of interactions and non-linearities. Given this exceptional performance, we decided to maintain this pilot as a separate, dedicated model for Utrecht, rather than train a single model for all linkage tasks.

Full linkage

The second phase in the linkage process involved taking the lessons from the pilot phase to design a procedure and train a model capable of predicting links across all of the tax data. This is where the inherent challenges of data heterogeneity in both the tax registers and the population registers become particularly apparent. As mentioned earlier, the name and initials represent the only consistently shared feature between these two primary data sources. Other potentially useful linking information may be missing entirely, and the specific

additional linking details that are available vary considerably from one municipality to the next. Because of the variation between municipalities, we take a municipality-by-municipality approach to linking, adjust the blocking procedure and postprocessing to the specific context

municipality	year	N	type
Lochem	1920	110	Large, no fn
Harlingen	1919	104	Large, fn
Amersfoort	1909	93	Large, no fn, neighbourhood
Nijmegen	1909	87	Large, no fn, neighbourhood
Zierikzee	1879	79	Medium, fn, occupation
Edam	1879	78	Medium, fn, neighbourhood
Uden	1879	75	Medium, no fn
Made en Drimmelen	1879	74	Small, fn
Brummen	1909	69	Medium, no fn, neighbourhood
Helvoirt	1920	67	Small, no fn
Hilvarenbeek	1909	58	Small, fn, neighbourhood
Bergambacht	1909	55	Small, fn, occupation
Philippine	1899	32	Medium, no fn, occupation
Alphen	1864	31	Small, no fn, occupation

To address this variability, we selected 14 municipalities that represent a spectrum of data availability, ranging from sparse to rich, as well as different population sizes, operating under the assumption that larger municipalities might present more complex linking challenges. Within these municipalities, we identified several common combinations of available linking information, such as initials and surname only, first name and surname, name and neighbourhood, and name and occupation.

Again, we extracted a sample of 1000 households from the tax registers for these twelve municipalities for this phase. This sample was stratified to ensure that challenging cases, such as records pertaining to widows and women, were adequately represented in our GT data. It is important to note that this oversampling of difficult cases means that the performance metrics reported below likely underestimate the model's true performance on the broader dataset, where these harder-to-link cases will be less frequent.

This sample was manually labelled with true links and a model was trained using a LightGBM ranker model.
[train/test split
other relevant steps]

This model required a substantial number of features to perform well, including semantic distances between strings, regular Levenshtein string distances, dummy variables, and various machine learning inferences (such as gender prediction from names and identification of household positions within the population registers). The specific features utilized in this 12-municipality model are detailed in table X.

[table x model features goes here]

The performance of this 12-municipality sample was not as high as that of the dedicated Utrecht model. The primary reason for this difference is the lack of complete address data in the population registers of most of the twelve municipalities, which was consistently available in both the Utrecht HO tax registers and population registers. Nevertheless, considering the inherent sparsity of the original data and the significant variation in available features across the different municipalities, the model's overall performance remains very good. The performance of the model on non-widows shows a high test precision of 0.94 and an even higher test recall of 0.99, resulting in an F1 score of 0.97 and a group correct score of 94%. The results for widows, the harder group, indicate a test precision of 0.89 and a test recall of 0.79, leading to an F1 score of 0.84 and a group correct score of 75%.

The model's ability to recover a significant number of links despite these challenges can be attributed to several key factors: - reason1 - reason2

The trained model is then applied to predict links across the full candidate set. We select the highest-ranked link that exceeds a defined THRESHOLD. In instances where there are very plausible alternative links (indicated by a standard deviation of predicted probabilities within the top three link candidates being less than 0.012), we remove the link to maintain higher confidence in the remaining linkages. The resulting set of predicted links is then used to combine the tax and population register data.

In the project repository, the complete linksets are provided, allowing users the flexibility to adjust the link criteria (tightening or loosening them) to better suit their specific research objectives. It is important to acknowledge, however, that assessing the overall quality of the links across the entire dataset is a complex task due to the sheer volume of records in both the tax and population register datasets. The current threshold is set to optimize the model's performance on our representative sample, which was designed to reflect the characteristics of the full dataset in terms of available linking features. Based on these optimized links within the linksets, we also generate a version of the tax data that incorporates relevant information from the population registers for each linked household. This includes a link to a persistent identifier within the population registers, as well as key demographic variables that will facilitate further linkage to other historical records, such as the civil registries.

Data description

Applications of the data

Looking at two applications demonstrates the promise of the data assembled, building upon the work initiated by the team. The first application involves the reconstruction of income inequality in the Netherlands between 1860 and 1920, a study that primarily requires the tax data and reference population figures. The second example leverages the linked data as well by investigating the income premium/penalty of migrants to the city of Utrecht, whom we are able to identify using the place of birth information derived from the population registers.

Inequality

An obvious application of this assembled data lies in reconstructing income inequality for the crucial 1860–1920 period. This era witnessed significant large-scale economic, political, and demographic changes, yet our current understanding of inequality during this time is based on limited data, primarily the pioneering work of Soltow and Van Zanden (1998). More robust series derived from national income taxes only start in 1917 (Salverda and Atkinson, 2007).

Here, we briefly present one initial reconstruction of inequality in the Netherlands over the 1860-1920 period. This particular application does not require the linked data but still demonstrates the potential of HIP-NL to provide new insights into these historical trends. The reconstruction presented here requires a number of steps that we can only outline superficially here. Incomes across the various tax systems need to be harmonized, which we do using a model-based approach that leverages tax registers where rich data is available. Furthermore, tax-exempt households require imputation, which we achieve by drawing from a truncated lognormal distribution. The resulting inequality series is illustrated in figure X below.

[fig: inequality in NLD, 1860-1920]

One of the striking findings emerging from this new series is that the decline in inequality begins after a peak around 1890. This timing is notable as it precedes the traditional dating of the Great Compression by approximately 30 years. Interestingly, this aligns with early findings for Sweden (Bengtsson and Molinder). Together, these works suggest that underlying structural changes (e.g. Kuznets 1955; Goldin and Katz 2009), rather than factors like taxation (Alvaredo et al .2013), war-related destruction (Scheidel 2017), or deglobalisation (Milanovic 2016), were the primary drivers of this shift in inequality.

Migration

The second application focuses on an analysis of the incomes of migrants compared to natives. Existing research in this area has largely relied on occupational titles (e.g., Paping & Pawlowski 2018, Puschmann 2015, Schrover, 2002), an approach that likely obscures considerable income variation within occupational groups. Notably, there have been no studies conducted on this topic using income data directly. This analysis is relevant to the

ongoing debate surrounding the labor market inclusion of migrants, with pessimistic, optimistic, and intermediate perspectives – the latter often differentiating between short-distance and long-distance migrants, as well as stayers versus leavers. The linked HIP-NL data has the potential to bring new precision to this important discussion.

Here, we present a comparison of migrants and natives in Utrecht for the years 1889 and 1909, based on income data from the taxes, and place of birth data from the linked population registers. Figure X illustrates that migrants are over-represented in the middle of the income distribution and under-represented at the top.

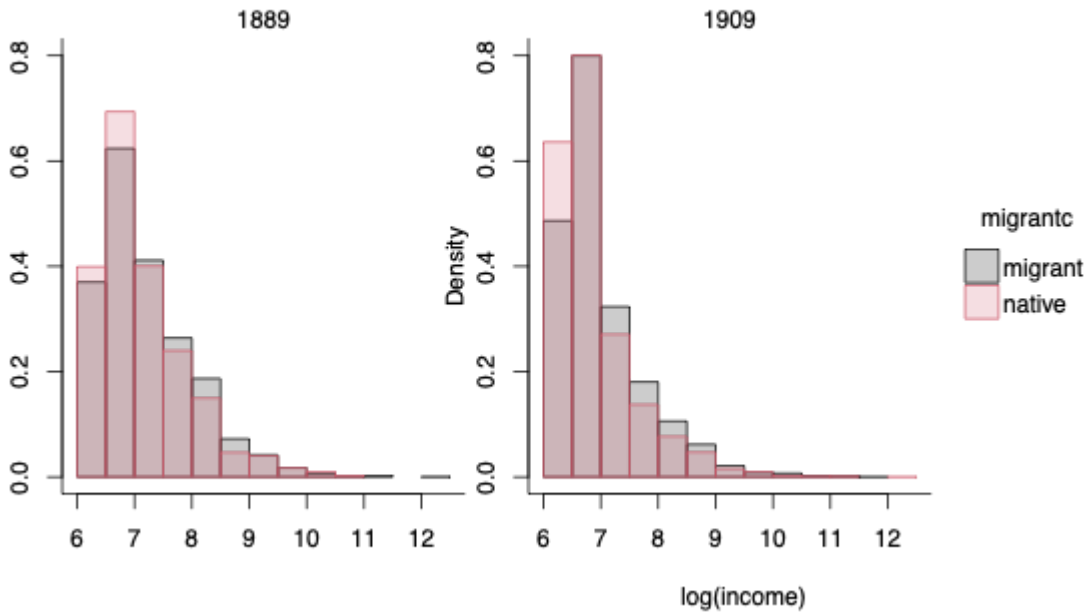

Figure x. Income distribution of migrants v. natives in Utrecht, 1889 and 1909.

We can also more formally test these observations within a regression framework. The detailed demographic information available in our linked data allows us to adjust these income estimates for age, a potentially crucial factor if migrants exhibit a different age profile compared to Utrecht natives.

Dependent Variable:	log(mid)		
Model:	(1)	(2)	(3)
Variables			
migrantTRUE	0.1018***		
	(0.0083)		
nobleTRUE	1.420***	1.433***	1.415***
	(0.0671)	(0.0710)	(0.0712)
far_migrant		0.1359***	
		(0.0091)	
close_migrant		-0.0977***	
		(0.0128)	
urban_migrant			0.1750***
			(0.0106)
rural_migrant			0.0076
			(0.0098)
Fixed-effects			
year	Yes	Yes	Yes
age5	Yes	Yes	Yes
Fit statistics			
Observations	26,036	24,113	24,107
R^2	0.12121	0.13178	0.13024
Within R^2	0.04473	0.05550	0.05387

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table x. Regression of log(income) on migrant status in Utrecht, 1889 and 1909.

Our regression analysis reveals that migrants, on average, have incomes approximately 10% higher than natives (table x). This income premium is concentrated among urban, long-distance migrants, as opposed to those originating from Utrecht's immediate hinterlands. These findings provide clear evidence supporting a more positive view of the economic position of migrants during this period. We anticipate that the full HIP-NL dataset, with its longitudinal dimension, will further enrich this picture by allowing us to follow migrants over their life courses.

Conclusions and future avenues

- Boring summary
- Opportunities to apply the data
- Link our data to other sources (again example of voorbeeld munic/year)
    - V. van Achternamen opzoeken in openarch hetutrechtsarchief
        - Link (some persons from the initial example Utrecht page?) to other datasets, such as:
            - Militia registers
            - Nationale Spoorwegen staff data

        - Notarial deeds
        - GIS options with addresses
            - Plot voorbeeld munic/year
- Link to other databases: tafel vbis
- Additional data collection through OCR
- Occupational HISCO scores

Appendix

- Evaluate metrics
- Metadata table: grossincomes, n_observations etc (DANS?)
    - List munics/years in sample: report on:
        - entered manually or OCR
        - Refpop
        - Finished munics/year and unfinished
- Variable list tax records
    - Definitions
    - Availability variables (total and by munic/year)

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. 'Automated Linking of Historical Data'. Journal of Economic Literature 59 (3): 865–918. https://doi.org/10.1257/jel.20201599.

Alfani, Guido. 2021. "Economic Inequality in Preindustrial Times: Europe and Beyond." Journal of Economic Literature 59 (1): 3–44. https://doi.org/10.1257/jel.20191449.

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. 'The Top 1 Percent in International and Historical Perspective'. Journal of Economic Perspectives 27 (3): 3–20. https://doi.org/10.1257/jep.27.3.3.

Bengtsson, Erik, and Jakob Molinder. 2024. 'Incomes and Income Inequality in Stockholm, 1870–1970: Evidence from Micro Data'. Explorations in Economic History 92 (April):101568. https://doi.org/10.1016/j.eeh.2023.101568.

Bengtsson, Erik, and Jakob Molinder. Forthcoming. 'What Happened to the Incomes of the Rich during the Great Levelling? Evidence from Swedish Individual-Level Data, 1909–1950'. The Journal of Economic History. https://doi.org/10.1017/jeh.2021.28.

Berg, Niels van den, Ingrid K. van Dijk, Rick J. Mourits, P. Eline Slagboom, Angelique A. P. O. Janssens, and Kees Mandemakers. 2020. 'Families in Comparison: An Individual-Level Comparison of Life-Course and Family Reconstructions between Population and Vital Event Registers'. Population Studies 0 (0): 1–20. https://doi.org/10.1080/00324728.2020.1718186.

Boonstra, O.W.A., E. Beekink, Th.L.M. Engelen, and Hans Knippenberg. 2003. ‘De Historische Databank Nederlandse Gemeenten (HDNG)’. In Nederland in Verandering. Maatschappelijke Ontwikkelingen in Kaart Gebracht 1800-2000, edited by Erik Beekink, O.W.A. Boonstra, T.L.M. Engelen, and Hans Knippenberg, 169–74. Amsterdam : Aksant. http://hdl.handle.net/2066/125510.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. 2017. “The Fading American Dream: Trends in Absolute Income Mobility since 1940.” Science 356 (6336): 398–406. https://doi.org/10.1126/science.aal4617.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. “Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States.” The Quarterly Journal of Economics 129 (4): 1553–1623. https://doi.org/10.1093/qje/qju022.

Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin; New York: Springer.

Clark, Gregory, and Neil Cummins. 2014. “Surnames and Social Mobility in England, 1170–2012.” Human Nature 25 (4): 517–37. https://doi.org/10.1007/s12110-014-9219-y.

Deaton, Angus. 2024. ‘What Is Wrong with Inequality?’ Oxford Open Economics 3 (Supplement_1): i2–3. https://doi.org/10.1093/ooec/odad079.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. ‘The Faiss Library’. arXiv. https://doi.org/10.48550/arXiv.2401.08281.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2023. ‘Gemini: A Family of Highly Capable Multimodal Models’. arXiv. https://doi.org/10.48550/ARXIV.2312.11805.

Goldin, Claudia Dale, and Lawrence F. Katz. 2009. The Race between Education and Technology. 1. paperback ed. Cambridge, Mass.: Belknap.

Hoekstra, Rinke, Albert Meroño-Peñuela, Kathrin Dentler, Auke Rijpma, Richard Zijdeman, and Ivo Zandhuis. 2016. ‘An Ecosystem for Linked Humanities Data’. In The Semantic Web, edited by Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange, 9989:425–40. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47602-5_54.

Kim, Geewook, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. ‘OCR-Free Document Understanding Transformer’. In Computer Vision – ECCV 2022, edited by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, 13688:498–517. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19815-1_29.

Klep, P.M.M., A. Lansink, and W. van Mulken. 1982. De kohieren van de gemeentelijke hoofdelijke omslag 1851-1922. Broncommentaren 1. Arnhem: V.A.N.-commissie Broncommentaren. https://resources.huygens.knaw.nl/broncommentaren/retro.

Kok, Jan, Kees Mandemakers, and Henk Wals. 2005. 'City Nomads: Changing Residence as a Coping Strategy, Amsterdam, 1890-1940'. Social Science History 29 (1): 15–43. https://doi.org/10.1017/S0145553200013237.

Kuznets, Simon. 1955. 'Economic Growth and Income Inequality'. American Economic Review 45 (1): 1–28.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. 'Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows'. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002. Montreal, QC, Canada: IEEE. https://doi.org/10.1109/ICCV48922.2021.00986.

Long, Jason, and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850." The American Economic Review 103 (4): 1109–37. https://doi.org/10.1257/aer.103.4.1109.

Mandemakers, Kees. 2000. "Historical Sample of the Netherlands." In Handbook of International Historical Microdata for Population Research, edited by Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, 149–77. Minneapolis, Minn.: Minnesota Population Center. https://www.international.ipums.org/international/resources/microdata_handbook/1_10_netherlands_ch11.pdf.

Milanovic, Branko. 2016. Global Inequality: A New Approach for the Age of Globalization. Harvard University Press.

Piketty, Thomas. 2014. Capital in the Twenty-First Century. Cambridge Massachusetts: The Belknap Press of Harvard University Press.

Raad, Joe, Rick Mourits, Auke Rijpma, Ruben Schalk, Richard Zijdeman, Kees Mandemakers, and Albert Meroño-Peñuela. 2020. 'Linking Dutch Civil Certificates'. In WHiSe 2020 Workshop on Humanities in the Semantic Web 2020, edited by Alessandro Adamou, Enrico Daga, and Albert Meroño-Peñuela, 47–58. CEUR Workshop Proceedings. CEUR-WS. http://ceur-ws.org/Vol-2695/paper6.pdf.

Riel, Arthur van. 2021. Trials of Convergence: Prices, Markets and Industrialization in the Netherlands, 1800-1913. BRILL. https://doi.org/10.1163/9789004460805.

Rijpma, Auke, Jeanne Cilliers, and Johan Fourie. 2020. 'Record Linkage in the Cape of Good Hope Panel'. Historical Methods: A Journal of Quantitative and Interdisciplinary History 53 (2): 112–29. https://doi.org/10.1080/01615440.2018.1517030.

Rijpma, Auke, Robin C. M. Philips, and Bas J. P. Van Bavel. 2024. ‘Multidimensional Composite Indicators of Well-being: Applications in Economic History’. Journal of Economic Surveys.

Rijpma, Auke and Rick Schouten. 2025. "NOCR! VLMs are superior to traditional OCR". Working paper.

Roikonen, Petri, and Sakari Heikkinen. 2018. ‘A Kuznets Rise and a Piketty Fall: Income Inequality in Finland, 1865–1934’. European Review of Economic History, December. https://doi.org/10.1093/ereh/hey032.

Salverda, W, and A B Atkinson. 2007. “Top Incomes in the Netherlands over the Twentieth Century.” In Top Incomes Over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries. Oxford University Press. https://doi.org/10.1093/oso/9780199286881.003.0010.

Scheidel, Walter. 2017. The Great Leveler: Violence and the History of Inequality from the Stone Age to the Twenty-First Century. The Princeton Economic History of the Western World. Princeton, New Jersey: Princeton University Press.

Smits, Jan-Pieter, Edwin Horlings, and Jan Luiten van Zanden. 2000. “DUTCH GNP AND ITS COMPONENTS, 1800-1913,” 260.

Soltow, Lee, and Jan Luiten Van Zanden. 1998. Income and Wealth Inequality in the Netherlands: 16th - 20th Century. Amsterdam: Het Spinhuis.

Van Zanden, Jan Luiten, Joerg Baten, Marco Mira D’Ercole, Auke Rijpma, Conal Smith, and Marcel Timmer, eds. 2014. How Was Life? Global Well-Being since 1820. Paris: OECD Publishing. https://doi.org/10.1787/9789264214262-en.

Van Zanden, Jan Luiten, and Arthur van Riel. 2004. The Strictures of Inheritance: The Dutch Economy in the Nineteenth Century. The Princeton Economic History of the Western World. Princeton, NJ etc.: Princeton University Press.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. ‘Multilingual E5 Text Embeddings: A Technical Report’. arXiv. https://doi.org/10.48550/ARXIV.2402.05672.