Linking the Historical Sample of the Netherlands with the USA Censuses, 1850–1940

By Diogo Paiva, Francisco Anguita, & Kees Mandemakers

To cite this article: Paiva, D., Anguita, F., & Mandemakers, K. (2020). Linking the Historical Sample of the Netherlands with the USA Censuses, 1850–1940. *Historical Life Course Studies, 9*, 1–23. http://hdl.handle.net/10622/23526343-2020-0003?locatt=view:master

HISTORICAL LIFE COURSE STUDIES

VOLUME 9

2020

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles

This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles

This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at http://www.ehps-net.eu/journal.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

EUROPEAN SCIENCE FOUNDATION

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

EHPS NETWORK

HISTORICAL LIFE COURSE STUDIES
VOLUME 9 (2020), 1–23, published 08-09-2020

Linking the Historical Sample of the Netherlands with the USA Censuses, 1850–1940

Diogo Paiva

International Institute of Social History, Amsterdam & Iscte – University Institute of Lisbon

Francisco Anguita

International Institute of Social History, Amsterdam

Kees Mandemakers

International Institute of Social History, Amsterdam & Erasmus University Rotterdam

ABSTRACT

During the 19th and early 20th century about 220,000 Dutch born persons migrated to the USA. The Historical Sample of the Netherlands (HSN) contains about 85,500 persons born in the Netherlands between 1812 and 1922. In this article we report the way we have matched persons from the HSN with the American censuses from the period 1850 till 1940. For this purpose, a linking process was designed, comprising of three stages: harmonization, matching and validation. The different nature of the two datasets (HSN and the USA Censuses) asked for some harmonization prior to the matching. Once the data had been properly prepared, two strategies were applied in order to link the data sets. The first one, called *Similarity Approach*, matched individuals from both datasets by comparing on the basis of resemblance of first and last names. The second approach, called *Transformation Approach*, made use of dictionaries with Anglicized versions of Dutch first and last names and their most common or most likely Dutch original(s). Because of the sample character of the HSN even exact matches showed ambiguity that needs to be resolved. For this reason, a validation process comparing the household context was run to provide a more trustworthy result. In the end we identified 484 individuals present in the HSN database with reliable links to the American censuses. We also evaluated the result in the light of what we know from emigration patterns to the USA over time and period and we concluded that our efforts have produced a reasonable result. Nevertheless, we are aware that we may have missed links. We also found that at least 45% of the emigrants returned to the Netherlands at some point during their life course.

Keywords: Historical life courses, Nominal record matching, Emigration, Social history, Historical demography

e-ISSN: 2352-6343
PID article: http://hdl.handle.net/10622/23526343-2020-0003?locatt=view:master
The article can be downloaded from here.

1 INTRODUCTION

The Historical Sample of the Netherlands (HSN) database contains individual life courses from the population of the Netherlands born between 1812 and 1922. By selecting Research Persons (HSN RPs) from the birth certificates and following their life courses using civil certificates and population registers, it is possible to reconstruct complete life courses. However, for some cases, there are interruptions in the sequence of observations when individuals disappear from the available sources. Emigration is among the causes of this problem (Mandemakers, 2006). The United States is one of the destinations that attracted a considerable number of Dutch migrants in the 19th and early 20th century. During the period from 1820 to 1940, about 220,000 Dutch individuals emigrated to this destination. Searching for these emigrated HSN RPs was made possible by getting access to all households with one or more persons born in the Netherlands from the full count of the American censuses from 1850 to 1940 (Ruggles, Genadek, Goeken, Grover, & Sobek, 2015). For this purpose, a linking process was designed, comprising of three stages: harmonization, matching and validation. In the end we identified hundreds of individuals present in the HSN database with reliable links to the American censuses. This also resulted in an extension of the HSN life courses by adding new data from the American censuses to the HSN database.

The different nature of the two datasets (HSN and the USA Censuses) asked for some harmonization prior to the matching. Especially the Census datasets required harmonization, since most of the *full count* datasets were not yet prepared for public access (only samples from them). In addition, the linking involved two different languages (Dutch and English) that made it more challenging, as individuals' names went through a process of change from the Dutch original to an American English form.

Once the data had been properly prepared, two strategies were applied in order to link the data sets. The first one, called *Similarity Approach*, matched individuals from both datasets by comparing on the basis of resemblance of first and last names, within the frame of a near birth year and the same sex. This process extracted a first set of individuals present in both datasets with a reasonable resemblance. The second approach, called *Transformation Approach*, started from a set of all names contained in the census files without correspondence in the HSN database. Only 22.4% of the people in the Census dataset were found to have a first and last name similar to names in the HSN. The remaining Dutch born individuals carried a relatively rare name or a modified name: their original name had been adapted to their new environment. Usually these modified names had been, either literally translated from Dutch into English, like 'Bos' into 'Bush' or converted phonetically into a more English one or had simply been changed by clerks or officers who were not acquainted with the Dutch language. Therefore, in this second strategy we created dictionaries with Anglicized versions of Dutch first and last names and their most common or most likely Dutch original(s).

Matching small samples instead of a complete dataset has the disadvantage that even in the case of a one to one match, there is no certainty that the match is correct. Simply, because the potential matches with persons outside the sample are not known. The HSN dataset samples only 1 out of 133 or 1 out of 200 Dutch born, depending on the period. Therefore, each pair of matched individuals still has some ambiguity that needs to be resolved. For this reason, once a set of matched RPs was obtained, a validation process comparing the household context was run to provide a more trustworthy result. This process created a validation score, constructed by combining several indicators such as presence of recognizable family members and known departure to the United States of America, consistency of dates regarding emigration and the quality of the individual match (exactness of names and birth date).

Additionally, record linkage was implemented between the censuses themselves. This was accomplished by data triangulation between matched HSN persons and consecutive censuses (i.e. 1850 and 1860, 1860 and 1870, etc.). The goal was two-fold: First, to check if the quality of data linking between the HSN and consecutive censuses was comparable to those we obtain by linking consecutive censuses. And second, to be able to spot HSN RPs in other censuses which had only been matched with one of the censuses. In this way, it was possible to construct a set of linked RPs with reasonable confidence and to extract a 'Dutch migrant dataset' from the USA censuses with additional data.

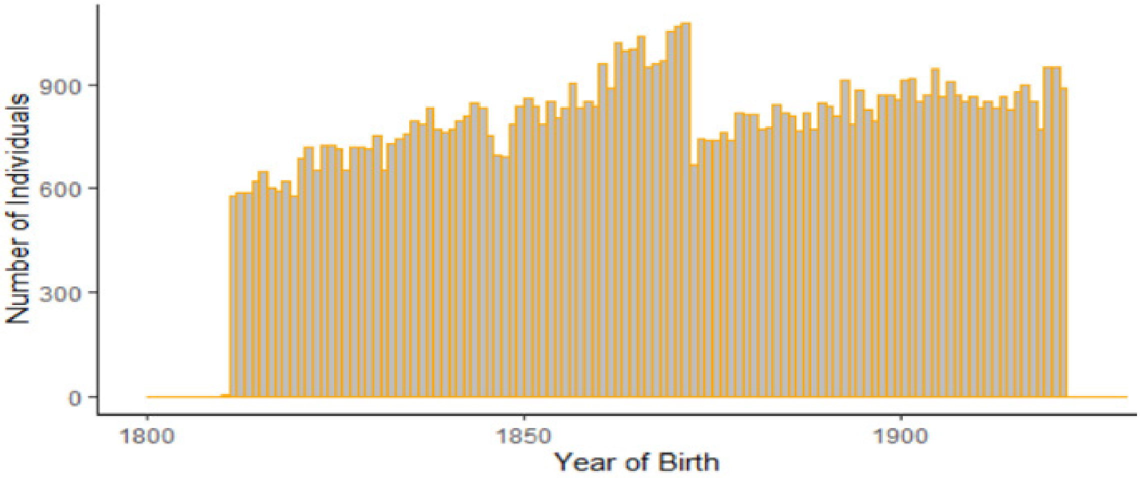
In section 2, we present both the HSN and USA Census Datasets and discuss some inherent problems of these datasets. Next, in section 3, we elaborate on the used methods of record linkage and present the results obtained from the record linkage process as such. In section 4, we test the outcomes of the matching on plausibility, by way of the validation process. In section 5, we present a 'Dutch migrant dataset' and evaluate this result with what is known from Dutch emigration to the USA. Section 6 concludes with a short discussion of the results.

2 DATA

2.1 HISTORICAL SAMPLE OF THE NETHERLANDS

The HSN contains data from 85,334 individuals sampled from the birth certificates corresponding to the period between 1812 and 1922. The sample was drawn on the basis of a sample frequency of 0.75% for the years between 1812 and 1872 and 0.5% for the period of 1873 to 1922. The goal of this sample strategy was to get more or less equivalent cohorts at the age of 16, considering the changing numbers of births and levels of infant and child mortality (Mandemakers, 2000). In Figure 1, we present the frequency of years of birth of the HSN sample. As could be expected there is a drop at the moment the sample frequency lowers from 0.75 to 0.5%.

Figure 1 *Number of HSN Research Persons by year of birth, 1812–1922*



Source: *HSN Release Civil Certificates 2017.01*

The dataset used for this research consisted of four HSN subsets created from the HSN database. In anonymized form the data were made available in three different releases (Historical Sample of the Netherlands, 2010, 2016, 2017). The datasets informed about different features of the research persons:

- *HSN Basic* provides data from the birth certificates: first and last name — with prefixes, if applicable — sex, birth date, birthplace and the RPs parents' names (n RPs=85,334, but because of changes in last names the actual number of records equals 89,956).
- *HSN Lost* presents information on those that got lost from observation, including the reason, the last period of observation, the last known location, and in case of emigration the destination. As far as known 3,847 HSN RPs emigrated of which 570 mentioned North America, America or explicitly the USA as destination.
- *HSN Marriages* gives data from the RPs' marriage certificates such as date of marriage, groom's and bride's names, their ages, occupations and birth places, as well as the groom's and the bride's parents' names and, in case they were alive at the moment of the wedding, ages and occupations (n=32,827 marriage records).
- *HSN Survival* contributes with the RPs death date (n=64,323) or in the case this is lacking, the date of last observation.

In all of these files, the RP is identified by a unique identifier. The dataset *HSN Basic* provided the information used for matching (i.e., name, birthdate, sex) except the date of death which was delivered from the dataset *HSN Survival*. The other datasets were used in parts of the validation process. As stated, about 4,500 RPs have more than one record in the HSN Basic file. This is caused by RPs being registered under two different last names. In most cases this happened when the registration of the father was posterior to the birth. With no legal father at the moment of birth the child received the last name of the mother. A large part of these children was legitimized by the wedding of the mother with the supposed father. That is why most of these names changed a couple of months after birth. We expect that only one name can be a potential match, so for all calculations we use the number of sampled RPs (n=85,334).

2.2 AMERICAN CENSUSES

2.2.1 NUMBER AND AGE STRUCTURE OF DUTCH BORN

The access to the data of the American censuses 1850–1940 was provided by the Minnesota Population Centre which is the home of the Integrated Public Use Microdata Series (IPUMS) of the USA (Ruggles et al, 2015). The full counts of these censuses are only publicly available for the census years 1850, 1880 and 1940 (see https://usa.ipums.org/usa/complete_count.shtml). However, for our research we received a compilation of full count datasets of all individuals that stated to have been born in the Netherlands from the years of 1850 until 1940 (except 1890, which census forms have not survived). The census files include information about names, age, sex, immigration year, residence, household composition and its members' roles, as well as other remarks on skills, disabilities, etc. Table 1 shows the number of available variables in each census and the number of 'Dutch born' persons that were included in the sample that we received from IPUMS.

Table 1 *Number of variables and Dutch born individuals per census year*

Census	Variables	Dutch Born
1850	22	11,546
1860	37	30,448
1870	35	47,330
1880	26	54,317
1900	88	96,455
1910	124	118,327
1920	321	133,150
1930	342	132,864
1940	395	111,843

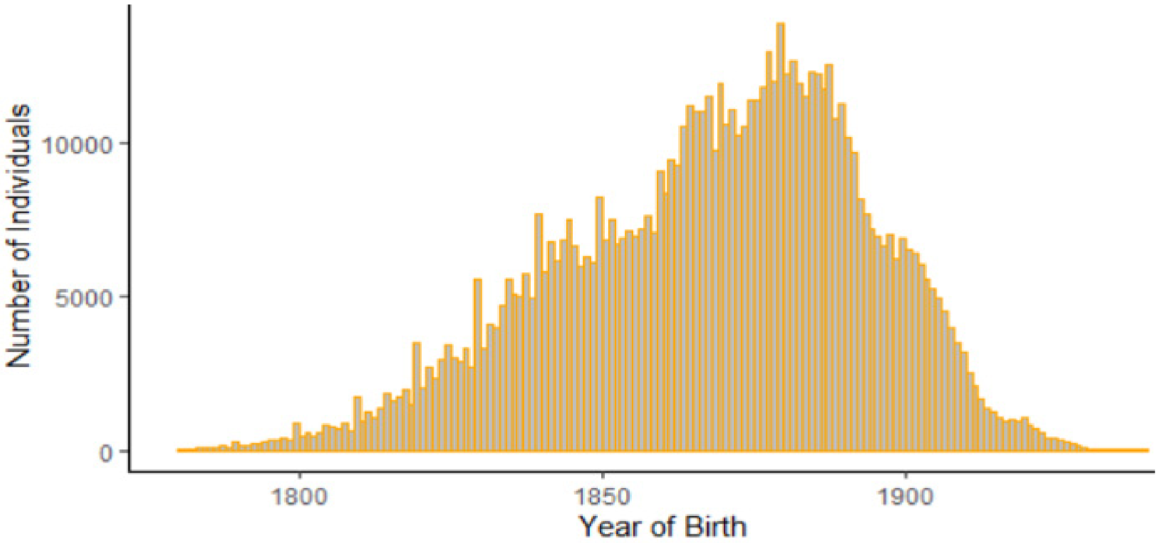
Source: IPUMS — Dutch American Census

Since the full counts of the non-public census years are still in the process of harmonization there were three basic issues in the census dataset that we had to solve: the absence of a unique identifier for each individual in every census, the fragmented information on households in some of them, and the diversity and irregularity of variables. Only the censuses of 1850 and 1880 had already been standardized and harmonized. The other ones (1860, 1870, 1900, 1910, 1920, 1930, 1940) had, at least for the full counts, not yet been harmonized and prepared for public use (when we received the data the full count of 1940 had not been published yet). All in all, the number of variables belonging to each census year increased enormously from 22 in 1850 till 395 in 1940 as shown in Table 1. Especially the harmonizations of 1910, 1920 and 1930 census were quite laborious.

Table 1 also shows the number of Dutch born individuals within each census. Combining the censuses over the years the dataset contains 736,280 records of individuals, of which almost 60% were concentrated in four states: Michigan counted 26.2% of all registrations of Dutch born), New York 12.3%, Illinois 10.5% and New Jersey 9.7%. We use the name 'Dutch American Census' for this Dutch subset of IPUMS USA.

Figure 2 shows the distribution of the years of birth for the total of all appearances of Dutch born persons in the censuses. The year of birth is imputed from the age at the moment of the census. We see a steady rise in the yearly number of Dutch born persons, until the birth year 1880 after which the number goes down. The lack of census records from 1890 helps to explain the rapid decline, as their presence would have smoothed the curve. However, the drop corroborates the progressive loss of interest of the Dutch to migrate to the USA, starting in the 1890s (Swierenga, 1985). According to this author, Dutch emigration partially shifted from being American centred (around 90% of overseas migration had the USA as destination, consistently throughout most decades in the 19th century) when interest in the Dutch East Indies rose. According to Obdeijn and Schrover (2008) immigration restrictions have already played a role since 1890, which in turn made Canada more popular as destination for Dutch migrants. Emigration to nowadays Indonesia rose after 1900 and during the interbellum 80 to 90% of the emigrants chose that destination (Bosma & Mandemakers, 2008). Before 1900 most immigrants to the Dutch East Indies were soldiers and civil servants, after 1900 and especially after 1920 the number of civilian migrants rose enormously, amongst others, due to the exploitation of the oil and rubber resources of Borneo and Sumatra (Bosma, 2010). Furthermore, the immense drop after 1920 corroborates with the very strict immigration laws implemented in the USA in the 1920s (Haines, 2000).

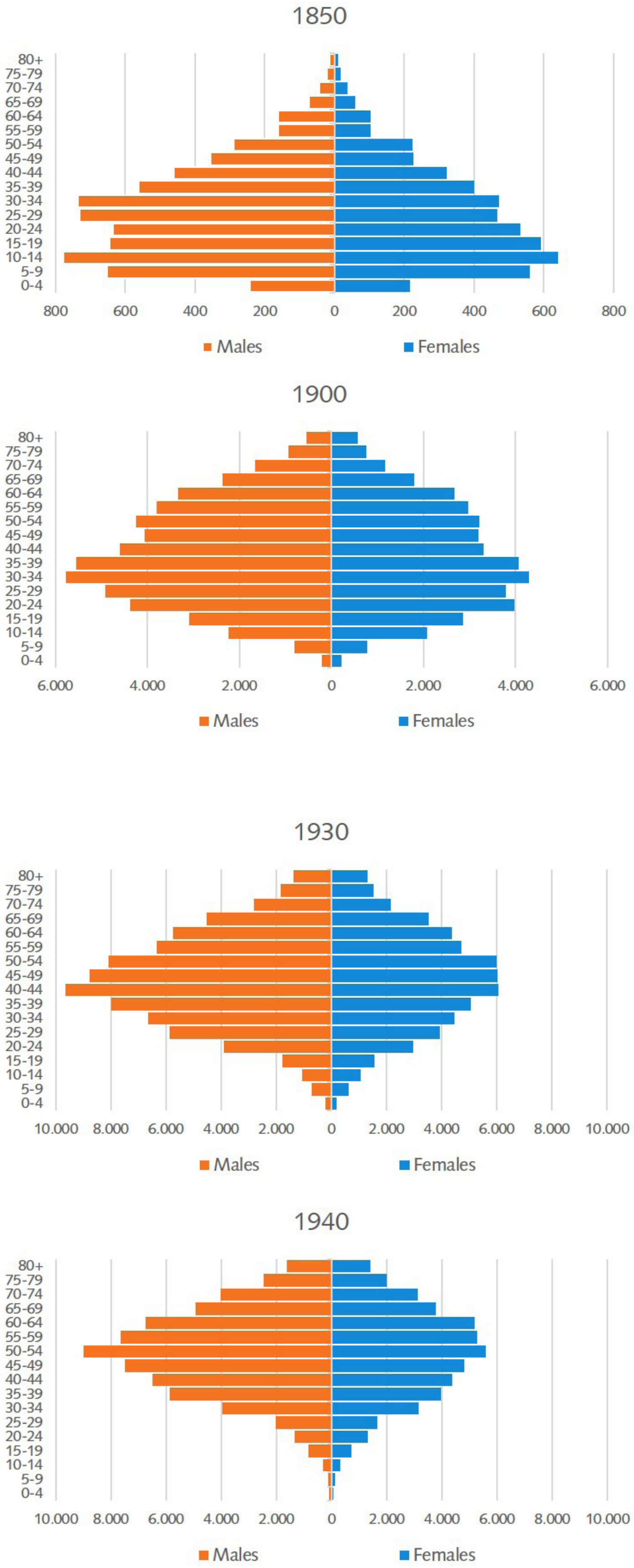
Figure 2 *Distribution of Dutch born by birth year in the USA censuses, 1850–1940*



Source: IPUMS USA Censuses — Dutch born.

In figure 3 we present the age pyramids of the Dutch born for 1850, 1900, 1930 and 1940. All four pyramids show an a-typical population structure for a 19th and early 20th century society. Most exceptional is the relatively weak presence of children. Especially for the 19th century one would expect 40 to 45% of the population is to be younger than 20 years (Engelen, 2009; Haines, 2000). Instead for 1850 it is no more than a 30% of total population and in 1930, 1940 no more than 10%. This ageing of the immigrant population is in line with the already mentioned slowing down of Dutch emigration after 1880. Secondly, we see an excess of males in the population, although 1850 seems to approach a more normal distribution. This is in line with what we know from Dutch migration to the USA (Stokvis, 1985). From 1890 onwards Dutch migration to the USA became more and more a business of single males and 'family resettlement' became less important. Swierenga reports that after 1900 over a third of all migrants to the USA were singles of which two thirds were males (Swierenga, 1993). Similarly, statistics provided by the Office of Homeland Security state a drop of Dutch immigrants obtaining legal permanent resident status, declining from 46,000 in the 1910s to 8,000 in the 1930s (Department of Homeland Security, 2009). These developments are clearly shown in figure 3 with a shift in age distributions for the Dutch born in the years 1930 and 1940. Already during the period 1890–1905 and 1914–1918 the number of Dutch USA migrants slowed down to very low numbers, which is in combination with the growing economic attractiveness of the Dutch East Indies, one of the reasons for the high share of the elderly after 1920 (Bosma & Mandemakers, 2008).

Figure 3 *Age Pyramids of Dutch Emigrant Population in the 1850, 1900, 1930 and 1940 censuses*

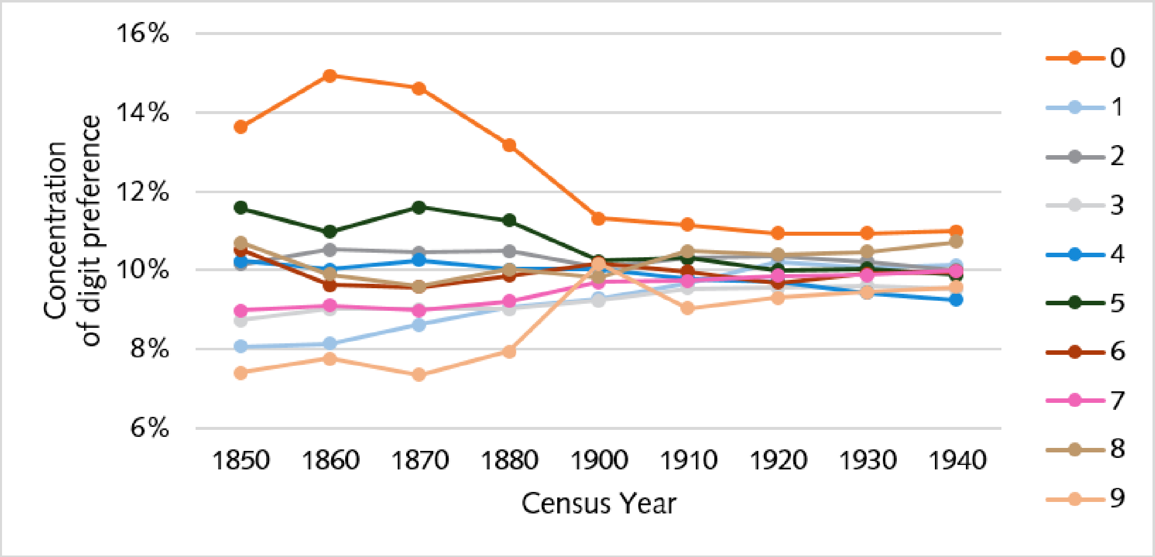


Source: IPUMS — Dutch American Census

2.2.2 AGE HEAPING

Historical data are not always as accurate as one would wish. One issue is 'age heaping' which is the phenomenon that persons are inclined to estimate their age in rounded figures, like 30 or 55. It is a well-known problem when analysing census data (Steckel, 1992; Szoltysek, Poniat, & Gruber, 2019). As we already saw from the peaking pattern in Figure 2, birth years ending on '0' or '5' seem to be over-represented, especially in the censuses before 1900. Figure 4 presents the frequency distribution (based on ages) of the last number of the birth years for the censuses 1850–1900. Given equal age distributions within a decade, we may suppose that each last digit of a birth year has a chance of 10% to appear in a census. Under-represented are particularly the years ending on '9' and '4'; over-represented are the years ending on '0' and '5'.

Figure 4 *Evolution of Age Heaping within the group of Dutch born individuals, per Census Year*



Source: IPUMS — Dutch American Census

Table 2 *Whipple's Index per Census Year*

	Censuses								
	1850	1860	1870	1880	1900	1910	1920	1930	1940
Whipple's Index	142.3	142.5	139.9	128.1	109.7	109.1	104.9	104.7	104.0
Data Quality	Rough	Rough	Rough	Rough	Fairly Accurate	Fairly Accurate	Highly Accurate	Highly Accurate	Highly Accurate

Source: IPUMS — Dutch American Census; the indicators of data quality are derived from United Nations (1990, p. 19).

One way to measure the level of age heaping is the Whipple's index. This index, developed by the American demographer Whipple, is a ratio of the number of persons aged between 23 and 62 reporting an age ending on a '0' or '5' and the total number of persons in this age category, multiplied by 500, resulting in an index ranging between 100 and 500 (Siegel & Swanson, 2004, p. 136–138). Using the United Nations categorization of age data quality (United Nations, 1990, p. 19) from table 2 we also see very clearly that the earlier censuses suffered from an age heaping problem. The improvement on the stating of age is very likely a result of modifications in the questionnaire for the enumerator. Up until 1900, enumerators only asked the age of the persons and, when in doubt, were requested to estimate the age of the individuals as exactly as possible. Aware of the skewness brought to the census by age misreporting, the enumerator instructions in 1890 warn of a tendency to report age in round numbers. In the next census 1900, the questionnaire was reformulated, to include two questions: age and birth date. This was done to better calculate age and to have a better proof of the age. From

1910 onwards, however, questioning birth date is dropped and instead the enumerator is instructed to confirm the age when people reported ages ending in '0' or '5' and to make an effort to be accurate when determining the age of the surveyed persons.[1] Regarding the execution of the matching process, as year of birth is one of the primary variables, we will use an age range of plus/minus two to prevent false negatives as a consequence of age heaping.

2.2.3 HARMONIZATION OF THE CENSUSES

Before we could start the matching, the census data had to be harmonized in several ways. Part of the files that we received from IPUMS, dated from 2016, were still not ready for public access, so they had to undergo our own cleaning process by revising and recoding variables and keeping track of data entry errors. We also had to include identifiers for all appearances of individuals and households in each census.

The recoding of variable values and names was not an easy job given the diversification available in the nine census files, but also given their different stages of standardization (visible for example in the different format of variables informing about wards, age, household serial number). Also, the number of available variables for each census increased as the years are closer to the present (see table 1). The harmonization also included the checking and removing of similar variables.

We created a unique identifier for the appearance of each person in each census. This 'CPID' — which stands for census person ID — was defined as the concatenation of the census year (i.e. '1850' for the first census, '1860' for the next, etc.) plus a sequence of digits from 1 to the total number appearances of individuals included in each census.

We also created identifying numbers for the households and the families. This was part of the reconstruction of the family units of the census. We needed these units to check the realized individual matches within the family context. The censuses of 1860 and 1870 proved to be the most problematic, as pertinent information about household or family membership was missing. The censuses from 1900 onwards provided already identifying numbers for households, for instance, the dwelling serials which included the number of persons within each dwelling.

The rules to create households provided by the publicly available censuses of 1850 and 1880 formed the basis to restructure the remaining datasets (Ruggles, Hacker, & Sobek, 1995). Main problem of the census of 1860 was that from the different roles in a household only the head of household was recorded. Therefore, relations between household members had to be inferred. This was done by applying rules that probably entail a certain margin of error. First problem was to disentangle households that contained two or more individuals with the same name and both defined as the head of the household, e.g. cases where the son had the same first name as his father. Here, the first of the potential heads is selected as head since it is assumed that the censor would first register the head, then the wife, then sons and daughters and afterwards the remaining members. The rest of the relations were attributed by using the following rules: the wife is considered to be the female next to the head and both ages should not differ more than 10 years. The children are those individuals with age differences to the head larger than 15 years. We are thus, assuming that all younger individuals from the household are sons and daughters of the head (disregarding nephews and servants, for instance), and that the wife of the head is the mother of all of them (ignoring stepmothers and stepfathers). Moreover, we assume that women with an age close to the head are the wife instead of a sister, cousin, servant or any other possible relation.

The following census (1870) was even more problematic as it did not present any relationships or defined roles (not even the head) and the geographical location of households was incompletely filled. This resulted in some oversized households, seemingly containing several distinct families. To tackle this problem, the large households were divided in smaller parts by way of grouping individuals by common last names. Of course, this can cause some errors, dividing real households in parts as in the case for servants or cousins and/or joining different households of families sharing the same family name. For constructing relationships within the household, the head was defined as the oldest male member (with the risk that the oldest member of the household could be just the father and not be the real head anymore). The rest of the household relations were defined with the same rules as for the 1860 census.

[1] For the questionnaires and enumerator instructions, see https://usa.ipums.org/usa/voliii/tEnumForm.shtml.

Censuses from 1900 onwards supplied accurate household roles, although the census of 1900 has some issues with grouping different households into one very large household. This was dealt with in the same way as 1870. Since all censuses expressed the roles within a household as a relationship to the head of the household, all have the problem of not knowing if the children present are of both the head and the wife or only of the male while the wife is actually the step-mother. Nevertheless, for the purpose of the validation process we considered all children as being offspring of both the head and his wife.

The foregoing process resulted in an unknown number of mistakes by wrongly establishing ties between individuals. However, since the purpose was only to compare — after matching — family members known in the HSN with the census family or household members, the impact of these miscalculated relationships will be limited.

2.3 EXPECTED RESULT

There is a structural difference between the census dataset and the HSN. The HSN is a longitudinal database that contains information from different sources that is linked around RPs and collected and identified through their complete life course. In the census dataset individuals are not linked at all, so only identified within the census itself. Since individuals appearing in two or more censuses are not identified as the same persons, these persons will show up several times as unique persons. It is also possible that they would not show up at all in the case they had been born or immigrated and died or emigrated during the ten years between two censuses.

Before starting the matching, we can estimate the number of linked records that we may expect to link. As pointed out in section 2.1, the HSN is a sample that represents 0.75% of the Dutch population for the birth period 1812–1872 and 0.5% for the period 1873–1922. Consequently, we would expect similar proportions among the Dutch emigrants to the USA. Estimates for total Dutch migration are mainly from Swierenga and Stokvis. Swierenga (1985, p. 33) counted 180,000 individuals from 1835 to 1920 to the USA, while Stokvis (1985, p. 59) presents around 220,000 for 1820–1910. Compared with other European countries Dutch emigration figures were comparatively low, if one considers that in total 14 million European citizens migrated to the USA during the period 1840–1910 (Bodnar, 1987).

For the estimation of the percentage of emigration among HSN RPs, we can only make a rough approximation. We can probably take 200,000 as a reasonable number for emigration in the period 1810 till 1920. If we add about 20,000 persons for the period after 1920 we count about 220,000 emigrants in total. If we consider an average sample rate of 0.63%, we can expect about 1,400 HSN RPs to be linked. However, the HSN is a random sample over time and regions (Mandemakers, 2000). And emigration to the USA was not randomly distributed. It peaked in certain time periods and certain regions (Stokvis, 1985). We also know from our HSN sources that 570 persons have registered to migrate to the America's (mainly USA, see section 2.1). This implies that the result will be lower than an average of 1,400 linked RPs. After the matching we will discuss the results in light of what is known from the emigration streams themselves.

3 MATCHING METHODS

There is a large body of literature on nominal matching and the necessary preparations before matching is possible like the standardization of names or locations (for overviews see Bloothooft, Christen, Mandemakers, & Schraagen (Eds.), 2015; Ruggles, Fitch, & Roberts, 2018; Schraagen, 2014; Schürer, 2007). Most of the approaches use forms of similarity matching in which a match is accepted if the names are considered as equal after no more than 1 or 2 character changes, examples of methods, named by the founders, are Levenshtein and Jaro-Winkler. To bring down the enormous amount of strings to be compared, methods of 'blocking' on first characters, age and sex are introduced as well. Nevertheless, Ruggles, Fitch and Roberts (2018, p. 28), overlooking the field of matching with the American censuses, concluded that 'We are witnessing the Wild West of record linkage: Almost every new study introduces some new variant in methodology'. Another approach is name standardization by building dictionaries in which names that are essentially the same but will never survive potential matches are listed as equal. For example, different equivalents of William such as Guillaume, Bill, Willem, Wim, Wilhelm are brought back to some standard to make linking possible. This method is widely used in linking French Canadian names (database BALSAC with the FONEM program, see Bouchard, Brard, & Lavoie, 1981; Vézina, St-Hilaire, Bournival, & Bellavance, 2018). Besides complete automatic linking, approaches in which more complicated candidates for matching are evaluated in a semi-automatic manual way exist as well, besides BALSAC, e.g. also the Demographic Database of Umea (Larsson & Engberg, 2016). New developments are the introduction of the family context. Not only one person is matched but pairs of persons which gives more room for uncertainty in the matching of person A given the more certainty of the match of person B (Wisselgren, Edvinsson, Berggren, & Larsson, 2014).

After the harmonization and preparation of the two large datasets (HSN and Census) the process of matching started. In our approach two matching methods were carried out for the construction of a table with linked persons: one based on similarity and one based on a dictionary. We named them the *Similarity* and the *Transformation* approach. But before we started this exercise, we made an overview of the first and last names of Dutch persons in the American Censuses that appeared in the HSN dataset as well. In this way we could check how big the relevant USA name sample was and understand how far our methods could reach. In the next chapter, after the matching, we will use a contextual method to validate the linkage results.

3.1 NAME COMPARISON

As a result of the comparison of the names in the Dutch American census with the HSN dataset, the records in the Census file are divided into four categories: those with common first and last names; those with only a common last name; those with only a common first name; and those of which both names are not present in HSN. In the case of the comparison of first names only the first part of a first name was taken into account (e.g. for Mary Jane only Mary). Table 3 presents the results of these comparisons.

Table 3 *Number of last and first names in the Census Dataset that are included in the HSN Dataset*

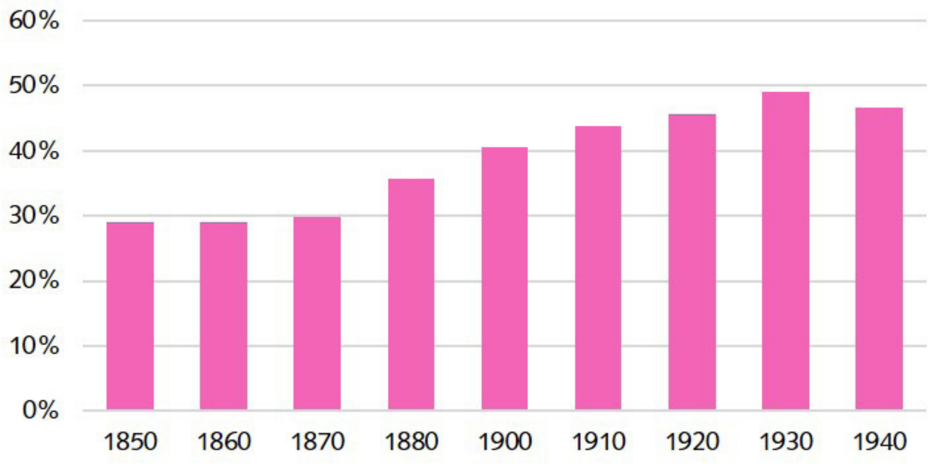
Names in Census dataset	First name in HSN dataset	Last name in HSN dataset	Number	Percentage
Full HSN	X	X	162,041	22.0
Only first name	X	–	341,552	46.4
Only last name	–	X	71,287	9.7
Non HSN	–	–	161,400	21.9
Total			736,280	100

The census group of Dutch born individuals whose first and last names are identical with names appearing in the HSN is 22%. All other ones will have already English names or adaptions, typos or translations of Dutch names. And, it should be noted that a part of these names will be genuine Dutch, but does not show up in the HSN sample because of their low frequencies. In table 3, we see that changes in the first names were less fashionable than changes in the last names. And for 21.9% of persons, both types of names do not have an equivalent in the HSN. Of course, by applying a distance string metric (e.g. the Levenshtein distance) the possibility of matching will become larger, since small differences will be equalized in this process.

Figure 5 shows the results of the appearances of the last name per census year. We see that the proportion of appearances of Dutch born with HSN last names increases over time, rising from about 25% to over 40% in the censuses from 1900 onwards.

Lastly, in table 4, we can see the amount and percentages of non-HSN last names per birth cohort. The data is clustered into five cohorts of the HSN database which cover more or less five equal lengths of birth periods. The percentage of non-HSN last names shows to be higher for those born in the earlier years, but it is never lower than 54%.

Figure 5 *Proportion of Dutch born individuals in American censuses with an HSN last name per census year*



Source: IPUMS — Dutch American Census

Table 4 Dutch born population, according to birth cohort and type of last name, appearing in the USA censuses 1850-1940

	Before 1833	**1833–1854**	**1855–1876**	**1877–1898**	**1899–1922**
Percentage with HSN last name	32.4	36.3	42.7	45.2	46.0
Percentage with non HSN last name	67.6	63.7	57.3	54.8	54.0
Total (100%)	55,253	136,964	214,355	233,315	78,382

3.2 SIMILARITY APPROACH

Due to the magnitude of the two data sets to be linked (between the HSN with 89,956 records and the Dutch American Census with 736,280 records there are about 66 billion potential matches), the matching process required a limit on the numbers to be matched. So, we used only those pairs of persons from the HSN and the Dutch American census that fitted the following criteria:

- Individuals must have the same sex
- Differences in birth year may not exceed two years
- Individuals must be alive at the moment of the census, i.e. the birth date must occur in the year of the census or before and — if known — a date of death must have occurred after the year of the census
- The length of each pair of last names may not differ more than two characters
- The length of each pair of first names may not differ more than two characters

Although string distances are not applied yet, it made no sense to keep as potential matches individuals whose names were too large or too short to fit a maximum Levenshtein distance of 2. For the similarity approach, this preparatory stage meant the creation of a set of 505,147,061 potential matches.

After the blocking we checked all potential pairs, if they complied with a Levenshtein distance of 2 or lower. In other words, all the pairs in this set of potential matches set are narrowed accepting only those whose names (last and given names) were separated maximally two Levenshtein distances. Matching with Levenshtein procedures means that every change to make string1 equal string2 counts as one, independent of the nature of the change (replacing, inserting or deleting a character) and the place of the change (for Levenshtein and alternative similarity measures, see Schraagen (2014)). After applying Levenshtein we have a lot more limited set of matches of 112,478 couples appearing both in the HSN and USA censuses.

Secondly, we checked the length of the names. For names with a length of maximum five characters tolerance was reduced to only one change (Levenshtein distance = 1), while for longer names two changes were accepted. This is done to tighten the outcome of matched pairs, keeping in mind that the similarities between names should not be equally assessed regardless of their size. After this step the matched sets remained 21,169 pairs. Given the commonness of names and the nature of the samples, this set will include both false and true matches. By way of the validation procedure (section 4) we shall try to exclude the false matches. Together with the results from the next transformative approach the results of the similarity procedure are presented in table 5.

Table 5 *Total amount of matches according to matching stage and type of matching*

Type of match	Pre-Matches	Matches	Matched	To be validated
Similarity	505,147,061	112,478	21,158	21,158
Transformed	384,573,655	119,802	18,834	25,694
Last transformed	86,114,440	49,958	7,917	12,065
First transformed	170,702,025	27,637	5,618	5,762
Both transformed	127,757,190	42,207	5,299	7,867
Total	889,720,716	232,280	39,992	46,852

3.3 TRANSFORMATION APPROACH

As shown in section 3.1, the potential number of HSN individuals we could ideally match by way of our *Similarity Approach* was only a part of the full potential of matches. We noticed that there were quite a lot of names in the Census files that were not included in the HSN dataset. Partly because they were quite unique names, partly because Dutch born had taken an English version name, for example by translating or anglicizing their names. By transforming anglicized names to their most likely original Dutch name, we try to expand the number of matches. We call this the *Transformation* approach.

This approach involved the construction of a dictionary of Dutch-English first and last names, following some known Anglicization rules like:

- Prefixes becoming attached to the last name (e.g. Van der Pol to Vanderpol)
- Endings in -ink, -els or -sen become -ing, -les, -son, respectively (e.g. Mennink to Menning)
- Direct translation of names (e.g. Vos to Fox, Smid to Smith)
- Beginning in V- may turn to Ph- or F- (e.g. Vries to Fries)
- -kk- turns into -ck-, -eo- to -oo-, -ij- to -y- (e.g. Bakker to Backer or Dijker to Dyker)

Besides, for the first names we could use a file compiling 6,999 conversions that were collected by Hoffmann (1996) and Kelly (2000). For instance, an American first name as Valentine can be translated as 'Fell', 'Felt', 'Felte', 'Feltes', or 'Feltine'.

For the last names we built a list of 12,390 last names conversions. In addition to the conversions based on the rules of Kelly, we also included, for the most frequent last names in the census, approximated name conversions, to increase the chances of matching. For instance, a Dutch last name like 'Kortschot' could evolve into 'Koskoty', 'Cruscut', 'Cortschot', 'Crosscut' or 'Crosscutt'.

A big part of the cases has more than one hypothetical version. The matching algorithm uses all the possible cases and tests their proximity to the names in HSN. We started with the same blocking as the similarity approach which implied that individuals should have the same sex, a maximum tolerance of two years in the birth range an being alive at the moment of the census. From this point we constructed three groups: a) with a transformed first name and the original last name; b) with a transformed last name and the original first name and c) with a transformed first and a transformed last name. Since for this approach we also apply Levenshtein in a second step, we applied the two last conditions of the similarity approach as well: both last and first name should have a maximal length difference of two characters. The results are included in table 5. We see that the number of potential matches is lower than the similarity approach because the dictionary covers only part of the names. And after matching with Levenshtein with maximally two characters differences the number of potential matches is much lower and after including the restriction for short names to Levenshtein 1, it totals between 5,000 and 8,000 depending on the type of transformation.

The transformation approach delivered 18,834 matches pairs that needed to be checked to find true positive matches. Since the matching by way of transformation produced ambiguous results in its own way, we actually had 24,694 matches to be validated. This happens because of translation of names (e.g. if a family of 3 persons with 'Van Dyk' as last name in the census, the validation will test for both 'Dijk' and 'Dijke' Dutch forms, representing 6 records, two for each member, the validation file will then contain two records for the matched person, when the family name is 'Dijk' and when it is 'Dijke'). If an RP is also married more than once it will also multiply rows due to name translation. Together with the 21,158 pairs from the similarity approach we ended with 46,852 pairs to be validated.

4 VALIDATION PROCESS

The large number of matches obtained with both approaches indicates an enormous amount of ambiguity. In total we had 46852 matched pairs that needed to be validated (table 6). Because the HSN, is a sample of only 0.5 to 0.75% of total population, these pairs include a lot of ambiguity, especially if we realize that beforehand we estimated the total amount of matches at maximal n=1,400 (section 2.3). The problem is comparable with the matching done by IPUMS when they linked the samples of the censuses of 1850, 1860, 1870 and 1900 with the 100% sample of 1880. Their solution was to work only with very unique names to prevent ambiguous linking (Goeken, Huynh, Lynch, & Vick, 2011). We found a solution by a further development of the linking process. We introduced what we called validity tests to be performed to all matches, in order to ascertain their reliability as links identifying the same individuals. For several reasons, especially because some results remained ambiguous, validation may fail to recognize a certain link between a census-person and HSN RP.

The results of the validation process are presented in table 6 and in table 7. Table 6 contains the results of HSN RPs that matched within the context of a household in the American censuses and table 8 the results of HSN RPs that matched with a single householder. In the first line of each table we have included the number of all matches between an HSN RP and a Dutch born person in one of the USA censuses.

The matches within a household context were validated by comparing contextual information present in the HSN on parents, spouses and parents-in-law with the information on household members provided by the census. Once the datasets had been properly structured and missing information about household relations was inferred (see section 2.2.3), the actual validation took place. We checked if parents and spouses, known from the HSN database, were present in the households recorded in the censuses. We used the same matching procedures as we did with the HSN RPs themselves (maximal Levenshtein distance of 4 for the combination of last and first names and a range in birth years of maximally 2 years). Additionally, this validation takes care of the problem that for married women — unlike in Dutch records — maiden names are not commonly used. So, if the person to be verified is a female, both maiden name and the last name of the husband are used in the matching. In the census of 1850, we see for example that 685 persons matched with HSN RPs but that only 58 family members matched in combination with an HSN RP. Since more than one family member could be matched, the actual number of unique matched HSN RPs equals a little bit lower: 53. Over all the census years we see that HSN RPs matched in total 837 times with one or more persons in a household context. From table 7 we also learn that in most cases it was the father and the mother of an HSN RP that were matching through the household context. In case of the mother the American form of the last name was of overriding importance, which of course could also simply be the last name of the husband. In case of the spouses we have more or less the same situation if we may assume that in case of marriages husband and wife both emigrated to the USA. Finally, for each pair under evaluation a validation report was produced, i.e. a list of test results for each pair with two scores: succeeded (= 1) or failed (= 0). All scores were summed up and the total ranged between 0 and 3. Positive decision on matches made in an HSN person score at least one point (= one family member). In case of ambiguity the higher score was preferred to a lower score and in case of equal scores no decision about a match was made. These results are included in the bottom line of table 6.

For the matches with single householders it was of course impossible to match with other household members. To get reasonable results we limited the matches to exact pairs in combination with information about the period of emigration. The first requirement was that a person only matched in an almost exact way, which means that two of three matching elements: first name, last name and birth year should match exactly. The second one required coherence in the timeline of emigration, by comparing a known date of departure from the Netherlands to North-America or the moment of loss of observation in the Dutch population registers and the stated year of arrival in the United States (only available in the censuses of 1900, 1910, 1920 and 1930), or a match with a census in line with the expected arrival in the USA. The results are displayed in table 7. We see that out of a potential of 4,677 matches only 147 actually met both criteria.

So, the results in table 7 are the most plausible pairs of singles matched by our matching process. All in all, we established 837 pairs of HSN RPs matching with a person in an American household and 147 RPs matching with a single householder. We see that most links are made with the census of 1880 in the 19th century and with the one of 1920 in the 20th century. We also see that the matches of singles are mostly found in the period from 1900 onwards. This is in line with what we know from the composition of the emigration flows to the USA and with the notion that over the years the census became more exact which implies that the requirement of exactness for these kinds of matchings does not result in too many false negatives. But the number of matches is not the same as matched unique HSN RPs.

The results of table 6 and 7 show the accepted matches between HSN RPs and American censuses. Although we ended up with relatively strong matches there is still a lot of ambiguity in the system, because a) the same person (RP) can still match with several persons in one census and b) an individual in the census may be matching with several RPs. In the last step of the matching process linked records are classified into four categories: *Accepted, Rejected, Ambiguous and Redundant* (see table 8). *Accepted* are the matches that have reached a validation scoring of 1 or above and without ambiguity, i.e. the HSN RP does not match with anyone else in a given census, or in case of multiple possibilities, there is only one match with the highest score. *Rejected* are the matches that failed to obtain any score or that had validation scores of zero, but present no ambiguity. *Ambiguous* are those sets of matches which showed ambiguity because the validation scores ended with the same validation score, including pairs with scores of 0. If it is not possible to solve the ambiguity by comparing ranking scores, then the matches are discarded. Despite knowing which of the possible matches is highly possible to be the true positive, we preferred to keep the matches above suspicion, regardless of the validation score. The last category, *Redundant*, are the links which are double in the sense that the linking system could produce more than one match between an HSN RP and a person in a specific American census because of the different methods of matching. These links also include those RPs who scored a lower ranking than other RPs, matching the same person in a census. We included these figures in table 8 to keep them in line with the tables 6 and 7.

Ultimately, the selection process produced 601 accepted matches between RPs and Censuses. Out of these, 86 persons from the HSN are identified in one or more censuses and 398 in only one census. From those linking with more than one census we found 1 RP linking with 5 different censuses, 5 RPs with 4 censuses, 18 RPs with 3 censuses and 62 RPS with 2 censuses.

Till now, we have discussed the matching between the HSN RPs and the USA censuses. As an extra step we also matched the Dutch in the American censuses with each other. The combination of the positive results of these two matching processes could be A) all matches of an HSN RP with persons in different censuses are confirmed by internal matches within the censuses, or B) an HSN RP matches with more censuses but not all potential internal census matchings have been found or C) an HSN RP matches with one or more censuses but not with all potential censuses as shown by internal matching. The process of checking this kind of unsolved matchings we called *triangulation* which stands for the schematic view of record linkage between two consecutive censuses and their corresponding subset of the HSN (see figure 6).

Table 6 *Validation of the matches of HSN RPs with persons in family households in USA censuses, 1850-1940*

	1850	1860	1870	1880	1900	1910	1920	1930	1940	Total
Total number of matches within family households	685	1,090	2,080	4,100	5,135	5,998	9,007	8,178	5,902	42,175
Validation according to presence of matched family member:										
Father	18	10	20	43	27	35	80	53	7	293
Mother (Dutch form)	2	0	0	5	2	0	8	2	0	19
Mother (American form)	35	7	10	39	26	38	52	55	13	275
Spouse (Dutch form)	0	2	3	8	17	29	31	44	35	169
Spouse (American form)	3	0	11	18	19	26	26	34	25	162
Total number of matched family members	58	19	44	113	91	128	197	188	80	918
Total number of matched unique family members	53	16	39	100	81	119	181	170	78	837

Explanation: Because some individuals have more than one family member the final result is not the total of the separate columns.

Table 7 *Validation of the matches of HSN RPs with singles in USA censuses, 1850-1940*

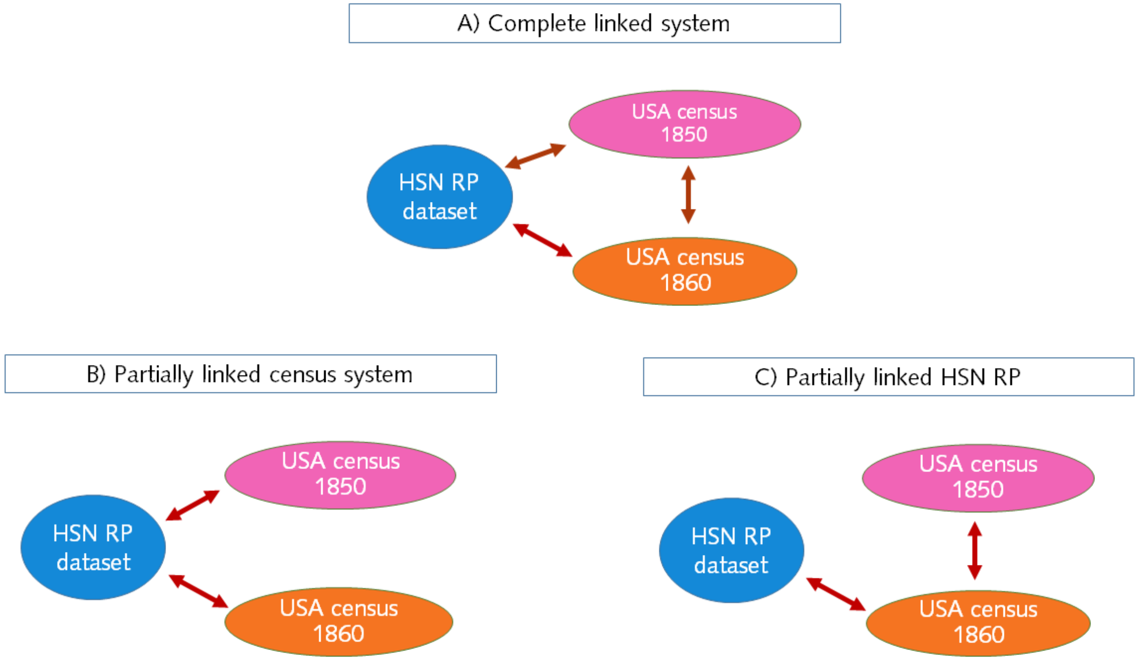
	1850	1860	1870	1880	1900	1910	1920	1930	1940	Total
Total number of matches with single householders	58	203	218	216	939	1,059	682	717	585	4,677
Validation according to exact matching										
Exact first name/Converted first name	327	550	1,055	1,725	2,246	2,628	3,550	3,336	2,372	17,789
Exact last name/Converted last name	159	206	377	999	1,172	1,335	1,896	1,627	1,194	8,965
Exact birth year	159	263	458	920	1,284	1,476	1,929	1,846	1,374	9,709
Validation according to coherence with contextual data:										
HSN RP kown to have emigrated to America	0	11	30	72	271	350	428	431	311	1,904
Concordance of last observation in Dutch administration and dates in censuses	26	97	242	523	841	989	1,135	1,143	841	5,837
Concordance of last observation in Dutch administration and immigration year					395	473	574	557		1,999
Number of validated singles										
HSN RP known to have emigrated to America + 2 out of 3 exact matchings	0	0	0	0	10	12	7	10	5	44
Concordance of Dutch sources with census data + 2 out of 3 exact matchings	0	4	1	2	23	19	15	27	12	103
Total of validated singles	0	4	1	2	33	31	22	37	17	147

Table 8 *Final coutcome of matching process according type of matching*

Type of match	Similarity		Last Transformed		First Transformed		Both Transformed		Total		Final result	
	Pairs	RPs	Pairs	RPs	Pairs	RPs	Pairs	RPs	Pairs	RPs	Pairs	RPs
Accepted	404	347	134	111	57	53	77	68	672	579	601	484
Rejected	13,200	6,578	4,447	2,282	2,696	1,650	1,918	1,156	22,261	11,666		
Ambiguous (rejected)	3,367	1,875	1,812	1,020	2,127	1,244	2,417	1,294	9,723	5,433		
Redundant	4,187		5,672		882		3,455		14,196			
Total	21,158	8,800	12,065	3,413	5,762	2,947	7,867	2,518	46,852	17,678		

Explanation: Since individual results from similarity and transformation approach overlap, the final result is not the total of the separate columns; the sum of all matched pairs is 46,852 which is the total of all matches included in table 7 and 8.

Figure 6 *Schematic overview of the possible outcomes of the triangulation process*



At first, series of subsets of matches from each pair of consecutive censuses were created. Of course a subset includes only individuals that are at least ten years older in each next census. For example, in case of the censuses of 1850 and 1860, we worked with the subset of the 1860 census individuals who were born in 1850 or before, and so on. We made subsets of all pairs of censuses up to 1940, and because of the lacking of the census of 1890, we used the census of 1900 to link with 1880. Within these subsets we continued the matching process by blocking on the basis of the following conservative rules: a) the same last and first names; b) the same sex; c) the same state within the US; d) the same county within the state. Since we matched with the same procedures as we did with the HSN-census matching, it is not surprising that for the matching situations under a) and b) we did not find any extra links that could also be validated in the same way as we did with the HSN linking. That leaves the third situation c) a partially linked HSN RP. We only found 8 HSN RPs which showed up in 11 other censuses without being linked directly. Most of the missing links were links with the censuses of 1910 and 1920 (8 out of 11 missing links). Since this is in fact a situation in which the Levenshein algorithm is further stretched, we consider this result as a conformation of our matching strategy.

We know that about 750,000 Dutch born are included in the censuses and that about 220,000 Dutch went to the USA over the period 1810–1940 (section 2.3). This implies that on average each Dutch born should appear three times in a census and two times in a triangle that connects two censuses. However, we only found 86 RPs matching with more than one census, giving an average of 1,24 census per person. We think that missing matches must be a result of several causes: a) a combination of context changing (singles becoming household members and the other way round), b) name changing (especially females) and c) the extreme difficulty to get unambiguous results when matching single persons (Goeken et al., 2011). Abramitzky, Boustan and Eriksson (2016) who matched Norwegians with the American censuses also realized relatively low matching rates (varying between 10.7 and 23.4%).

5 MISSING CASES?

We calculated a potential of 1,400 HSN RPs to be matched with the American Censuses. In the end we only matched 484. From the HSN itself we learned that from 571 persons there was an indicator that they went to the America's. This gap of about 900 persons are these 'missing' cases, or are these links not likely to be made beforehand? To get an answer we will take a closer look into the chance that persons disappeared between two censuses and the selection bias inherent to the Dutch emigration patterns.

5.1 DISAPPEARING BETWEEN CENSUS: MORTALITY AND RETURN MIGRATION

Censuses only provide a moment of observation, in the case of the USA a series of observations with intervals of ten years. It is reasonable to consider that some of the emigrating individuals present in the HSN died or returned to the Netherlands before a census moment could capture their presence.

For the purpose of making a basic estimation of how many emigrants died in the USA in an inter-census period, we need to know the crude number of deaths per 1000 persons (CDR) in the relevant period, preferably differentiated for the age structure. Table 9 presents the age structure of the HSN RPs at the first moment we matched them in the census. For the first two periods we see that 20 to 25% of the persons were younger than 20 years. This results in a relatively old age structure compared with the host country. The main reason is that infant and child mortality rated between 200 and 400 per 1000 in the Netherlands during the 19th century (van Poppel, Jonker, & Mandemakers, 2005; Walhout, 2019). This means that, in case of family migration, which was the main form of migration during the 19th century, the age structure of the children must have been relatively high. Furthermore, the improving quality of the censuses will give better second chances for persons to match at a higher age and there is a time lap between entering the USA and being included in a census. Another indicator for the relatively high age of HSN RPs is the mean age when they were first found in the census. This mean age lowered from 42 for the birth period 1812–1849 to 36 for 1850–1889 and 16 for 1890–1922. One reason for this trend to a lower mean age is that we only match till the census of 1940, another one is that we have averages that are dependent on the birth period. However, if we consider that the median age of the population raised from 19 years in 1850 till 27 in 1940 (Haines, 2000, p. 306), we can conclude that Dutch migrants were relatively old when they entered the USA.

Table 9 Relative number of HSN RPs migrating to USA according to birth period and age, 1812–1922

Birth Period	0–9	10–19	20–29	30–39	40–49	50–59	60 and older	Total	
								N	%
1812–1849	8.5	12.7	7.0	18.3	16.9	18.3	18.3	71	100
1850–1889	8.0	15.5	13.1	19.5	18.7	12.7	12.4	251	100
1890–1922	21.0	42.0	25.3	8.0	3.7	0.0	0.0	162	100
Total	12.4	24.0	16.3	15.5	13.4	9.3	9.1	484	100

So, all in all we have to be careful in estimating the average mortality in between censuses. According to Haines the CDR lowered from 23.66 during the period 1870–1880 till around 11.0 for the period 1910–1940 (Haines 2000, p. 315). Given the different age structures we cannot simply use the CDRs constructed by Haines, an average of 13% of all persons dying seems reasonable. However, on average the death risk period of emigrants was only five years, supposing an equal spread of emigration between the censuses. That puts the average mortality at 6,5%, which counts for about 90 HSN RPs (out of a total of 1,400).

There was also a risk of death at sea, particularly for voyages in the 19th century; they were long and in poor conditions. Based on Swierenga (1985, p. 25) we calculated a mortality rate of 9.1 per 1,000 for the period of 1820–1880. However, as Swierenga mentions, after 1880 better accommodations and faster ships drastically reduced deaths during the voyage. By taking this into consideration, we estimate that this risk counts for no more than 5 to 10 deaths. That means that from the 1,400 persons supposed to emigrate to the USA approximately 100 will never have reached the census because they died before the moment of census taking.

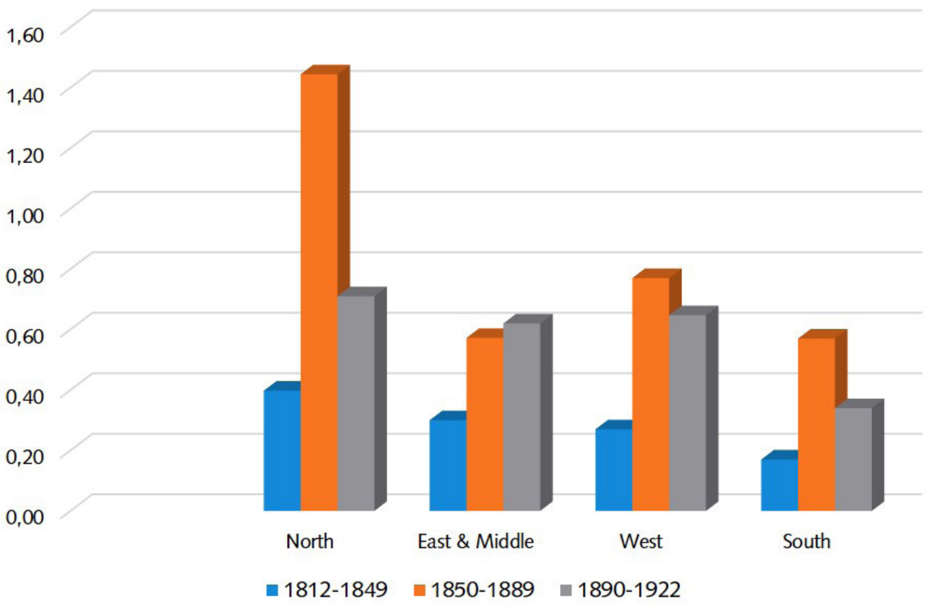
Another 'leak' is the possibility of return migration. Return migration is a well-known phenomenon from USA migration history, especially after 1870 when the journey became faster and the fares more affordable. From aggregate statistics from the early 20th century it is clear that this could be as high as 35% (Gould, 1980; Wyman, 1993). Abramitsky et al. (2016) suggest that this percentage could be even higher when using micro data looking into the whole life course. From the 484 identified persons we know that at least 221, that is 45.7%, returned to the Netherlands. We know this because the HSN dataset includes a death certificate of these persons or a personal card which also include the date of death. This relatively high percentage is due to the long time period between the match in a USA census and the moment of death. About 10% of these persons died in the Netherlands within a period of 10 year after their match with an American census. In general, the intervals between a first match and the moment of death are quite evenly spread between ranges of 0 till 90 years. But of course, a return moment on the date of death is not the same as the real moment of return. For the period between 1900 and 1910 Abramitzky et al. (2016) found that more than half of the Norwegian migrants spent no more than five years in the USA. Stokvis found percentages of 15% after 1870 and 5% before (Stokvis, 1985). If we assume for the whole period the conservative estimation of 10% then we explain another 140 RPs from the lacking 1,400 HSN RPs. This still leaves 660 'missing' HSN RPs to be explained.

5.2 EMIGRATION BIAS: PERIOD AND REGION

The HSN is an at random sample with a sample frequency of 0.5 to 0.75%. This implies that each person born between 1812 and 1922 has an equal chance to be included in the HSN (Mandemakers, 2000). And if USA emigration is equally spread over the years and regions, the HSN will include a fair share of emigrants. But this is not the case, as emigration to the USA was concentrated in some time periods and certain regions. Table 9 already showed the absolute number of matched HSN RPs for three birth periods. The total of 484 matched RPs stands for 0.57% of all HSN RPs. The first period, with 0.28%, is below this average. The main reason is that large-scale emigration to the USA did not start before 1840 and, secondly, mortality rates of the Dutch emigrants were quite high (Stokvis, 1985). The second period 1850–1889 shows a relatively high figure with 0.78%, while 0.59% for the period after 1889 is more or less on the average.

Figure 7 presents the relative shares of HSN RPs found in the USA depending on region and birth period. The country has been divided in four regions according to provincial borders (North: Groningen, Friesland and Drenthe; East & Middle: Overijssel, Gelderland and Utrecht; West: North- and South-Holland; South: Zeeland, Noord-Brabant and Limburg). For all three birth periods the North has the highest relative share in migrants. But as mentioned before, the period 1850–1889 shows the highest percentages of migrating persons and the North reaches even 1.4% of all HSN RPs which is two times higher than expected. We know that migration after 1889 was more evenly spread over the country (section 2.2), so for explaining lacking HSN RPs we will concentrate on the period before 1890, because during this period migration was heavenly concentrated in time and region.

Figure 7 *Percentage of HSN RPs migrating to USA according to birth period and region, 1812–1922*



Swierenga (1985) distinguishes three periods of relatively large migration flows before 1890: 1847–1857, 1865–1873 and 1880–1889 (continuing till 1893). Figure 8 clearly shows that migration for the period till 1880 (and in fact also during the 1880s with a severe agrarian crisis) was a rural matter. Till 1880 75% of the emigration came from only 134 mainly small rural municipalities, concentrated in no more than ten regions. Swierenga (1985) divided these municipalities into four groups according to the number of emigrants in relation with the total population: 34–92 per 1000, 93–155 per 1000, 156 –401 per 1000 and 401–1252 per 1000. Some municipalities were heavily depopulated during these emigration waves, especially the ones on the clay ground of Northern Groningen and the western part of Zeeuws-Vlaanderen. Other parts of Zeeland (except Walcheren) and the isles of South-Holland also had a lot of emigrants. Other spots were the Achterhoek in the east and municipalities along the big rivers.

In table 10 we used the spatial division as shown in figure 8 to map the HSN RPs according to the question whether they matched with the American census or did not. And indeed, we see the migration areas scoring systematically above the average of 0.55%, ranging from 1,09 to 1,91% for the nine municipalities from which half or more of the population went to the USA. In order to give a correct interpretation of these figures we need to think in terms of births not of inhabitants. The 85,344 HSN RPs stand for roughly 14 million births, of which about 9 million in the period 1812–1889. The 5,989 sampled births in these emigration municipalities stand for about 900,000 births. Assuming that about 25% did not survive the moment the family decided to emigrate because of infant and child mortality (Walhout, 2019), then we have 675,000 persons at risk. The average migration figure for the 'migration area' is 1.24% which equals 8,500 emigrated births. But in total about 75,000 persons left these municipalities during 1820–1890. Of course, not all of them were born in the relevant period and/or in these municipalities. But assuming that the number of emigrated persons also indicates the number of emigrated births, we have a number that is nine times higher than we could expect on the basis of equal chances for all births of the Netherlands. From table 10 we learn that from these migration areas 74 births were found in the USA, so 8*74=592 HSN RPs are the missing ones that can be explained from the lack of balance in emigration.

Figure 8 *Emigration Rate per Municipality, 1835–1880, per 1,000 Average Population, 1830–1878*

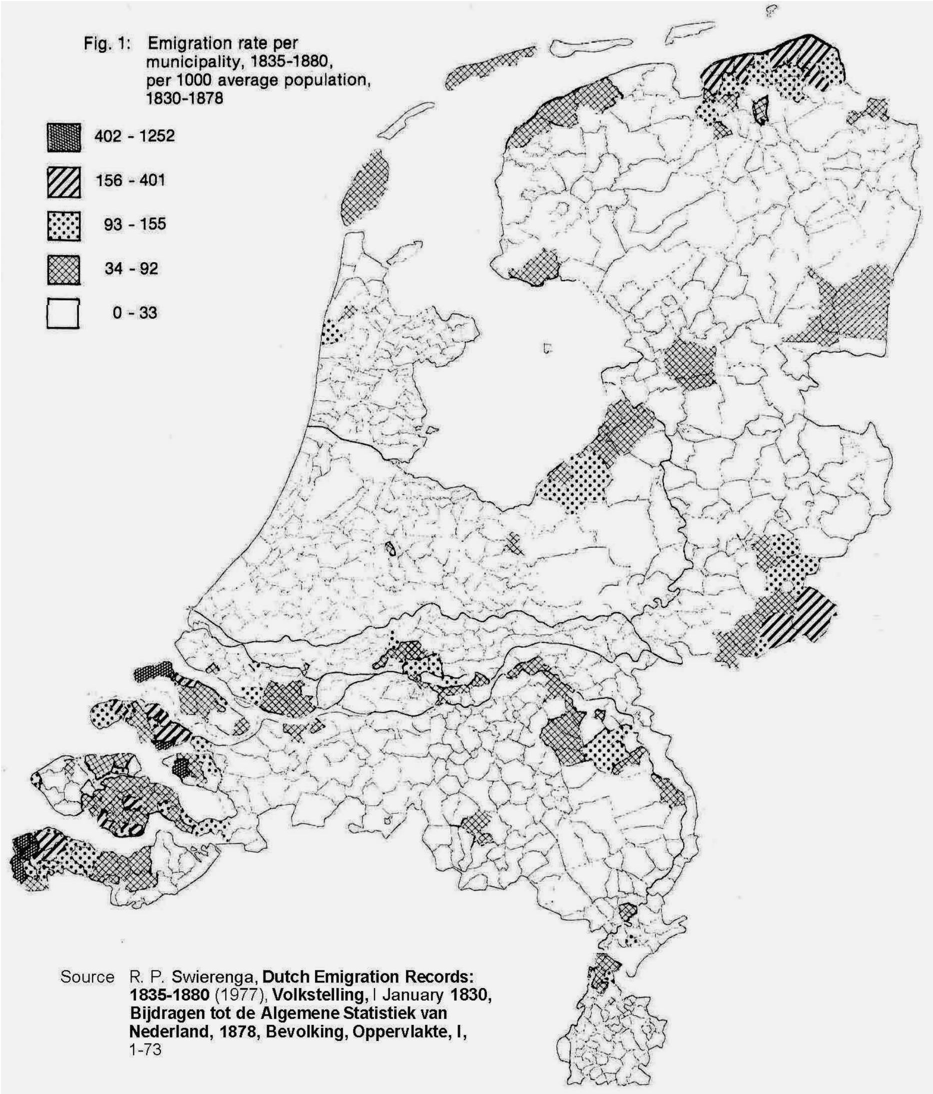


Fig. 1: Emigration rate per municipality, 1835-1880, per 1000 average population, 1830-1878
402 - 1252
156 - 401
93 - 155
34 - 92
0 - 33
Source R. P. Swierenga, Dutch Emigration Records: 1835-1880 (1977), Volkstelling, I January 1830, Bijdragen tot de Algemene Statistiek van Nederland, 1878, Bevolking, Oppervlakte, I, 1-73

Source: Swierenga (1982, p. 522).

Table 10 HSN RPs born in municipalities with large scale USA emigration, 1812–1889

N per 1000 pop.	Matched	Not matched	%	Total
0–33	248	51,969	0.47	52,217
34–1252	74	5,915	1.24	5,989
34–92	40	3,618	1.09	3,658
93–155	19	1,240	1.51	1,259
156–401	10	800	1.23	810
402–1252	5	257	1.91	262
Total	322	57,884	0.55	58,206

Explanation: Division in municipalities according to numbers of emigrants per 1000 average population size, based on Swierenga 1985, p. 34 (period 1830–1880).

6 SUMMARY AND EVALUATION

During the 19th and early 20th century about 220,000 Dutch born persons migrated to the USA. The sample of the Historical Sample of the Netherlands (HSN) contains about 85,500 persons (RPs) born in the Netherlands between 1812 and 1922. In this article we report the way we have matched persons from the HSN with the American censuses from the period 1850 till 1940. Based on the sample frequency of the HSN and known emigration figures to the USA, we expected 1,400 HSN RPs persons to be matched, given equal emigration chances over time, region, social background, etc.

The matching process was divided into three main sections: first, the data preparation; second, the criteria involved in the matching process and in obtaining the subset of matched records; and third, the validation of the linked persons obtained. At the moment the data from the Dutch born persons in the American censuses were made available by IPUMS, only the ones from the censuses of 1850 and 1880 were available in a cleaned and well-structured format. So, before matching we needed to check the other censuses and we added information to distinguish households and for the censuses of 1860 and 1870 we also imputed internal relationships within the households. The matching was based on name comparison, sex and birth period. Besides matching on name similarity with a Levenshtein maximum of two characters difference, we used the anglicised forms of Dutch names. After matching we used validation procedures to resolve ambiguity.

The application of the dictionaries with anglicized versions of Dutch last names and first names, helped to improve the results with about 14%. After both matching procedures there were 46,852 matches to evaluate. This high number is a consequence of the quite inaccurate character of the census data. So, we needed to use extra information from the census itself and Dutch registrations. In case of a household context we used other household members to validate and in case of singles we only accepted exact matches and coherence with Dutch registers indicating if and when a person emigrated to the USA (or America in general). This validation reduced the number of matches to 601 including 484 unique HSN RPs.

Final step was to evaluate the result in the light of what we know from emigration patterns to the USA over time and region. At first sight — on the basis of random chances of emigration — we expected a result of 1,400 HSN RPs. This is a gap of about 900 HSN RPs. We could explain 600 of them by differences in emigration patterns given time and space and 240 because of mortality risks and return migration before census taking. Adding these figures, we have explained almost the whole gap. Of course, we realize that this is too beautiful to be true, but is an indication that our efforts have produced a reasonable result. Nevertheless, we are aware that we will have missed links. We also found that at least 45% returned to the Netherlands at some point during their life course.

So, all in all it has been a quite big operation to get a result of almost 600 matches. We think it was worth doing for two reasons. The first one is that the HSN is a database with 85,000 life courses which consists of relatively many variables in which also a lot of money has already been invested. So, investing in an improvement of the database is not only of interest for research of emigration from the Netherlands to the USA but gives also a better perspective to other migration paths like the East Indies, Canada or neighbouring countries as Belgium, Germany and the UK. Secondly, it also elaborates methods that can be used to link another Dutch dataset, LINKS, which is based on an index of the civil certificates of the Netherlands (van den Berg et al., 2020). This includes the birth certificates which become public after 100 years. Within five years we expect to have a database of about 15 million birth certificates ranging from 1812 till 1920, linked with marriage and death certificates (Mandemakers, Bloothooft & Laan, forthcoming). We are looking forward to using our methods to link all Dutch born Americans with the context in which they were born.

ACKNOWLEDGEMENTS

We are grateful to the Minnesota Population Centre for making available this dataset with Dutch immigrants from the Integrated Public Use Microdata Series (IPUMS) and we thank Matt Nelson for doing the actual selection.

The article is written within the context of the project 'Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers' (LONGPOP) which is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676060. The results published in this article reflect only the author(s)'s view and the Research Executive Agency is not responsible for any use that may be made of the information they contain.

REFERENCES

Abramitzky, R., Platt Boustan, L., & Eriksson, K. (2016). *To the New World and back again. Return migrants in the age of mass migration*. NBER Working Paper No. 22659. Cambridge, MA: National Bureau Of Economic Research. Retrieved from https://www.nber.org/papers/w22659

Bloothooft, G., Christen, P., Mandemakers, K., & Schraagen, M. (Eds.). (2015). *Population reconstruction*. Cham Heidelberg New York Dordrecht London: Springer. doi: 10.1007/978-3-319-19884-2

Bodnar, J. (1987). *The transplanted. A history of immigrants in urban America*. Bloomington: Indiana University Press.

Bosma, U. (2010). *Indiëgangers. Verhalen van Nederlanders die naar Indië trokken.* Amsterdam: Uitgeverij Bert Bakker.

Bosma, U., & Mandemakers, K. (2008). Indiëgangers: Sociale herkomst en migratiemotieven (1830–1950). Een onderzoek op basis van de Historische Steekproef Nederlandse bevolking (HSN). *Bijdragen en Mededelingen betreffende de Geschiedenis der Nederlanden, 123*(2), 162–184. doi: 10.18352/bmgn-lchr.6779

Bouchard, G., Brard, P. , & Lavoie, Y. (1981). FONEM: Un code de transcription phonétique pour la reconstitution automatique des familles saguenayennes*. Population (French Edition), 36*(6),1085–1103. doi: 10.2307/1532326

Department of Homeland Security (2009). *Yearbook of immigration statistics: 2008*. Washington, DC: U.S. Department of Homeland Security, Office of Immigration Statistics. Retrieved from https://www.dhs.gov/immigration-statistics/yearbook/2008

Engelen, T. (2009). *Van 2 naar 16 miljoen mensen. Demografie van Nederland, 1800–nu.* Amsterdam: Boom.

Goeken, R., Huynh, L., Lynch, T. A., & Vick, R. (2011). New methods of census record linking. *Historical Methods, 44*(1), 7–14. doi: 10.1080/01615440.2010.517152

Gould, J. D. (1980). European inter-continental emigration. The road home: Return migration from the USA. *Journal of European Economic History, 9*(1), 41–112.

Haines, M. R. (2000). White population, 1790–1920. In M. R. Haines & R. H. Steckel (Eds.), *A population history of North America* (pp. 143–190). Cambridge: Cambridge University Press

Historical Sample of the Netherlands (HSN). (2010). *Life courses (Release 2010.01 (n=37.137))* [Data set].

Historical Sample of the Netherlands (HSN). (2016). *Survival dates (Release 2016.01 (n=85,334))* [Data set].

Historical Sample of the Netherlands (HSN). (2017). *Civil certificates (Release 2017.01 (n=85,334))* [Data set].

Hoffmann, J. K. (1996). *Dutch and Friesian first names anglicized. Names adopted in America by the Dutch immigrants*. Unpublished manuscript.

Kelly, A. C. M. (2000). *Names, names, & more names. Locating your Dutch ancestors in colonial America*. Orem: Ancestry.

Larsson, M., & Engberg, E. (2016). *How much do link metrics matter?* Unpublished paper presented at the 41st Annual Meeting of the Social Science History Association, Chicago 2016.

Mandemakers, K. (2000). The Netherlands. Historical Sample of the Netherlands. In P. K. Hall, R. McCaa, & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 149–177). Minneapolis: Minnesota Population Center.

Mandemakers, K. (2006). Building life course datasets from population registers by the Historical Sample of the Netherlands (HSN). *History and Computing, 14*(1–2), 87–108. Available from 10.3366/hac.2002.14.1-2.87

Mandemakers, K., Bloothooft, G., & Laan, F. (forthcoming). LINKS. The LINKing system for historical family reconstruction in the Netherlands.

Obdeijn, H., & Schrover, M. (2008). *Komen en gaan. Immigratie en emigratie in Nederland vanaf 1550.* Amsterdam: Uitgeverij Bert Bakker. Retrieved from https://openaccess.leidenuniv.nl/handle/1887/17762

Ruggles, S., Fitch, C., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology, 44*, 19–37. doi: 10.1146/annurev-soc-073117-041447

Ruggles, S., Genadek, K., Goeken, R., Grover, J., & Sobek, M. (2015). *Integrated public use microdata series: Version 6.0* [database]. Minneapolis: University of Minnesota.

Ruggles, S., Hacker, J. D., & Sobek, M. (1995). Comparability of the public use microdata samples: Enumeration procedures. *Historical Methods*, *28*(1), 33–39. doi: 10.1080/01615440.1995.9955311

Schraagen, M. (2014). *Aspects of record linkage* (PhD thesis Leiden University). Retrieved from http://hdl.handle.net/1887/29716

Schürer, K. (2007). Creating a nationally representative individual and household sample for Great Britain, 1851 to 1901: The Victorian Panel Study (VPS). *Historical Social Research, 32*(2), 211–331. doi: 10.12759/hsr.32.2007.2.211-331

Siegel, J. S., & Swanson, D. A. (Eds.). (2004). *The methods and materials of demography* (2nd ed.). United Kingdom: Emerald Publishing.

Steckel, R. H. (1992). Stature and Living Standards in the United States. NBER Chapters. In *American Economic Growth and Standards of Living before the Civil War* (pp. 265–310). National Bureau of Economic Research, Inc.

Stokvis, P. R. D. (1985). Dutch international migration, 1815–1910. In R. P. Swierenga (Ed.), *The Dutch in America. Immigration, settlement, and cultural change* (pp. 43–63). New Brunswick: Rutgers University Press.

Swierenga, R. P. (1982). Exodus Netherlands, Promised Land America. Dutch Immigration and Settlement in the United States. *Bijdragen en Mededelingen betreffende de Geschiedenis der Nederlanden, 97*(3), 517–537. Retrieved from https://www.dbnl.org/tekst/_bij005198201_01/_bij005198201_01_0027.php

Swierenga, R. P. (1985). Dutch immigration patterns in the nineteenth and twentieth centuries. In R. P. Swierenga (Ed.), *The Dutch in America. Immigration, settlement, and cultural change* (pp. 15–42). New Brunswick: Rutgers University Press.

Swierenga, R. P. (1993). The delayed transition from folk to labor migration: The Netherlands, 1880–1920. *The International Migration Review, 27*(2), 406–424. Available from https://doi.org/10.2307/2547131

Szołtysek, M., Poniat, R., & Gruber, S. (2018). Age heaping patterns in mosaic data. *Historical Methods, 51*(1), 13–38. doi: 10.1080/01615440.2017.1393359

United Nations. (1990). *1988 Demographic yearbook*. New York: United Nations. Retrieved from https://unstats.un.org/unsd/demographic-social/products/dyb/#statistics

van den Berg, N., van Dijk, I. K., Mourits R. J., Slagboom, P. E., Janssens, A. A. P. O., & Mandemakers, K. (2020). Families in comparison: An individual-level comparison of life-course and family reconstructions between population and vital event registers. *Population Studies*. doi: 10.1080/00324728.2020.1718186

van Poppel, F., Jonker, M., & Mandemakers, K. (2005). Differential infant and child mortality in three Dutch regions, 1812–1909. *Economic History Review, 58*(2), 272–309. doi: 10.1111/j.1468-0289.2005.00305.x

Vézina, H., St-Hilaire, M., Bournival, J.-S., & Bellavance, C. (2018). The linkage of microcensus data and vital records: An assessment of results on Quebec historical population data (1852–1911). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(4), 230–245. doi: 10.1080/01615440.2018.1507771

Walhout, E. (2019). *An Infants' Graveyard? Region, religion and infant mortality in North Brabant, 1840-1940* (PhD Thesis Tilburg University). Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/29027270/Walhout_An_Infants_25_01_2019.pdf

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods, 47*(3), 138–151. doi: 10.1080/01615440.2014.913967

Wyman, M. (1993). *Round-trip to America, The immigrants return to Europe, 1880–1930*. Ithaca, NY: Cornell University Press.