Record linkage in the Cape of Good Hope Panel

AUKE RIJPMA
JEANNE CILLIERS
JOHAN FOURIE

Stellenbosch Economic Working Papers: WP06/2018

www.ekon.sun.ac.za/wpapers/2018/wp062018

May 2018

KEYWORDS: census, machine learning, micro-data, record linkage, panel data, South Africa
JEL: N01, C81

Laboratory for the Economics of Africa's Past (LEAP)
http://leapstellenbosch.org.za/

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

Record linkage in the Cape of Good Hope Panel*

Auke Rijpma†, Jeanne Cilliers‡, and Johan Fourie§

In this paper we describe the record linkage procedure to create a panel from Cape Colony census returns, or *opgaafrolle*, for 1787–1828, a dataset of 42 354 household-level observations. Based on a subset of manually linked records, we first evaluate statistical models and deterministic algorithms to best identify and match households over time. By using household-level characteristics in the linking process and near-annual data, we are able to create high-quality links for 84 percent of the dataset. We compare basic analyses on the linked panel dataset to the original cross-sectional data, evaluate the feasibility of the strategy when linking to supplementary sources, and discuss the scalability of our approach to the full Cape panel.

Keywords. Census, machine learning, micro-data, record linkage, panel data, South Africa

*Paper prepared for the Workshop on Linking Historical Records, University of Guelph, Canada, 11-13 May 2017. Research made possible by the Swedish Handelsbankens and the Marianne and Marcus Research Foundation: Marianne och Marcus forskningsstiftelser (Dnr: MMW 2015.0027), Jan Wallanders och Tom Hedelius stiftelse, Tore Browaldhs stiftelse (Dnr: P2015-0159 and P2015-0409:1), CLARIAH-CORE, financed by NWO, and the South African National Research Foundation (96248).

†Department of History, Utrecht University, Drift 10 Room 3.08 3512 BS Utrecht, The Netherlands

‡Department of Economic History, Lund University, Room 2046, Alfa 1, Scheelevägen 15B, 22363 Lund, Sweden

§Department of Economics, Stellenbosch University, Room 618, Schumann building, Bosman Street, Stellenbosch 7602, South Africa

Introduction

Improvements in computing power and novel analytical techniques allow for the reconstruction of historical populations in a way that has brought historical demography into the realm of big data (Ruggles 2012, 2014). Mass digitization of historical sources, particularly those containing individual-level information is now commonplace, and not only in the developed world (Dong et al. 2015; Fourie 2016). However, to enable in-depth life-course analyses, individuals need to be identified and linked across multiple, often disparate historical records (Bloothooft et al. 2015). Introducing a degree of automation into this process increases efficiency, but raises questions of accuracy and potential bias (Feigenbaum 2016).

In this paper, we describe the record linkage strategy used to link households in the *opgaafrolle* tax records from the Cape Colony. The *opgaafrolle* were annual tax censuses collected between 1663 and 1834 of all free households of the Colony; first by the Dutch East India Company (VOC) administration and, after 1795, by the British colonial administration. Household-level information includes the name and surname of household head and spouse, the number of children present in the household, the number of slaves (and, in some cases, indigenous Khoesan employed), and several agricultural inputs and outputs, including cattle, sheep, horses, grain sown, grain reaped, vines and wine produced. Our ultimate goal is to create an annual panel of the agricultural production of households for over a century.

To create this panel, we evaluate a number of statistical models and deterministic algorithms to best identify households over time. After establishing the best approach based on a subset of manually linked records, we describe how we use this model to create a panel from the *opgaafrolle* for the Graaff-Reinet territory from 1787–1828, a 35-year, 42 354-observation dataset. We pay specific attention to the role of the panel nature of our data and the usage of household-level characteristics in attaining high linkage rates. We also compare basic correlations in the linked panel dataset to the original cross-sectional data. Next, we assess the re-usability and flexibility of our approach by adapting it to link the newly created panel dataset to a dataset of genealogical records, an independent dataset containing a different set of potential linking variables. We end with a discussion of the scalability of this approach to the full Cape panel.

Literature

Substantial efforts have already been made to link historical records in a systematic manner. Ferrie (1996) was among the first to use automated record linkage on historical data. His procedure to link individuals from the 1850 to the 1860 US census was based on the comparison of phonetically encoded

names, year of birth, and birthplaces. Improving on standard historical record linkage practices, Vick and Huynh (2011) analyse how name standardisation affects record linkage on the US and Norwegian censuses. Using manually-created first-name dictionaries to pre-process their linking data, they find improvements in linkage: while the overall linkage rate is decreased due to the removal of ambiguous links at the final stage, the higher number of candidate links created earlier in the process should mean the links are of higher quality. When linking the Swedish population registers, Wisselgren et al. (2014) use a combination of three techniques: manually standardised names, constructing surnames from patronymic naming practices, and using household information to make additional links between household members once the primary links have been established. Using this procedure on the high-quality Swedish population registers gives them linkage rates of 70%. Antonie et al. (2014) discuss their record linkage approach for the Canadian censuses of 1871 and 1881. While they use household-level characteristics to manually create their training data, automatic linkage is done based on individual-level characteristics only. Using a blocking strategy and a support vector machine (SVM) classifier, they achieve a linkage rate of 24%, with the largest share of missing links due to the discarding of ambiguous links. More recently, Massey (2017) evaluates historical record linkage practices by comparing their performance to high-quality links based on social security numbers. She finds that the use of string distance measures and probabilistic matching increases linkage rates substantially, but that this comes at the cost of a greater number of incorrect links.

Two papers deal explicitly with group versus individual level linking, which is highly relevant to our method. Goeken et al. (2011) give an overview of the record linkage strategy used to create the IPUMS project's United States Linked Samples, 1850–1930. They describe their strategy of blocking on race, gender, birthplace, and age range, the use of a SVM classifier, and a weighting procedure to correct for variable linkage rates across subpopulations. Of particular interest for the present paper is their approach to linking households and individuals. Because they do not want to create a bias towards people who remain in the same households from one census to the next, their primary linking is done on individuals, not households. Once the individuals are confirmed, other individuals are linked across censuses based on the fact that they are in the same household. While overall linkage rates are not reported, they achieve high reliability in the links (less than 2% false positives).

Fu et al. (2014) describe a group and individual-based record linkage technique for historical censuses. They note that historical record linkage is particularly problematic due to data quality. The main issue is that existing procedures tend to generate multiple links, that is, matches are created where there are actually none. They propose using household-level characteristics, which are more likely to be unique over time, to resolve

ambiguous links. Below we arrive at a similar conclusion: using additional household information can greatly improve linkage rates. The difference, however, is that we use household characteristics as the basis for primary links rather than as a tool for disambiguation after an initial linkage has been made.

Above all, the broad scope in methods and application of historical record-linkage suggest that a thorough understanding of the context of one's study is critical for developing a successful matching strategy, which is why we turn next to South Africa's *opgaafrolle*.

The *opgaafrolle*

When Europeans first settled at the southern tip of Africa, arriving in Table Bay in April 1652, their purpose was clear: to supply ships passing the Cape of Good Hope on their way to India and back to Europe with the necessary fresh produce, fuel and water. However, the small settlement was simply not large enough to produce enough food for its own survival and the almost 6000 soldiers and sailors that would frequent the Cape every year. In 1657, then, nine Company servants were released to become *vryburghers* (free settlers) and settle along the Liesbeeck River to farm. It was the beginning of a colonial society.

To keep track of the fluctuating levels of production of the fledging colonial society, the Company instituted annual tax censuses. It is unclear when the first proclamation for the *opneemrolle*, the name given to the tax censuses before 1794, or *opgaafrolle*[1], as they have since become known, were made. The inventory of the *opgaafrolle* in the Cape Archives Repository states that the 'earliest reference to the submission of a return of people and possessions is found in the journal of 12 October 1672' (Potgieter and Visagie 1974). The first return available in the National Archives of the Netherlands, transcribed by Hans Heese and Robert Shell in the 1970s, is for 1663, however.

Nevertheless, the household census of production became an annual event at the Cape soon after the establishment of a free settler society. The early censuses recorded only key household and production figures. Demographic information included the number of men, women, sons, daughters, European labourers, and slaves owned. Production information included the number of horses, cattle, and sheep owned, the number of vines planted and wine produced, as well as the volume of grains (wheat, barley, rye and oats) sown and reaped. In addition, the number of flintlocks, pistols and swords were also captured. Later censuses, particularly those after 1800, sporadically included additional production information, like the volume of brandy produced. Other assets, like wagons and carts, also appeared. Sometimes

1. *opgaafrolle* (plur.) *opgaafrol* (sing.)

information on recapitulation totals, mortality rates, church contributions and taxes paid are also included. In certain cases the head of the household signed his or her name next to the record.

It is these records the Cape of Good Hope Panel project is in the process of transcribing. Two full-time researchers, Chris de Wit and Linda Orlando, have been transcribing the original records now housed in the Cape Town Archives Repository since 2015. No digital copies of these records, which often span more than two meters when open, are available, and photography is prohibited by the Cape Town Archives Repository. Manual transcription has therefore been the only option.

The process is undertaken sequentially. First, the list of names are transcribed by De Wit, who is an expert in reading hand-written records. Values for each of the items are added in a second step, both to allow De Wit to focus on his comparative advantage, but also as a check to minimize errors. One census return spans several pages, and can take several days to complete.

The original census returns pose many transcription challenges. There are often two returns for each year, a 'concept return' and a 'final return'. There are numerous examples where the two returns do not correspond perfectly; in one of them, or both, sub-districts are amiss. We follow the general rule of transcribing the 'final return', where possible. There are also several returns missing. Figure 1 shows all the census returns that are available in the Cape Town Archives Repository between 1787 and 1842, as indicated in the inventory. The inventory also provides a short description of each return. It frequently states that some returns, or parts thereof, are missing or unreadable. Although ordered chronologically, parts of some censuses appear in other districts or years. Other information, like lists of Khoesan inhabitants, cattle lists and even a poem written about census collection, are also included.

It is also useful to note what these census returns do not include. While the names of the husband and, usually, the wife are included, childrens' names are not. Disappointingly, except for the name of the district and sometimes the name of the subdistrict, there is no information about the location of the household. Farm names are not included, for example. Occasionally the names of the household heads are listed alphabetically but are most often ordered randomly. It may be that the order is an indication of proximity: names were recorded in the sequence that the surveyor travelled through the countryside, visiting homesteads. Most importantly for economic analysis, no information about the value of land, the size of farms or homesteads or other forms of wealth (except slaves, productive assets, and wagons and carts) are included.

Between January 2015 and April 2017, 100 census returns were transcribed in full, and the names for an additional 29 have been transcribed (figure 1). The focus was initially on those records which also included num-

bers for Khoesan employed on settler farms. Such information was limited to the Graaff-Reinet and Tulbagh districts, the two frontier districts at the end of the eighteenth century on the eastern and northern borders of the Colony respectively. The third district to be transcribed was the Cape district, the returns for which differ substantially from those of the outlying regions. The Cape district returns are captured at the individual level rather than at the household level. We choose to focus here on the Graaff-Reinet district, the most complete and longest series transcribed at present.

[Figure 1 about here.]

This is not the first attempt to transcribe and use the *opgaafrolle* for research purposes. In the late 1970s, South African historians Hans Heese and Robert Shell spent several months transcribing census returns in the National Archives of the Netherlands, then known as the Rijksarchief. Their focus was on the first century of settlement. These transcribed records were never published, although some of the census returns were used in a handful of studies on early Cape inequality (Guelke and Shell 1983; Ross 1983). Johan Fourie and Dieter von Fintel uncovered these records in the Stellenbosch University Archives, and with the help of Hans Heese, digitised them into a format fit for analysis. Using more modern approaches to measuring inequality, Fourie and Von Fintel (2010, 2011) found high levels of Cape inequality, even within the settler population. Von Fintel, Du Plessis, and Jansen (2013) rely on the census returns to investigate the effect of male deaths on investment decisions. Fourie and Von Fintel (2014) used the census returns to show the skill advantage of the French Huguenot settlers, who arrived at the end of the seventeenth century and settled among the Dutch settlers. Fourie (2014), comparing the census returns to the probate inventories, calculated the extent of underreporting, while Fourie, Jansen, and Siebrits (2013) show how this mattered for the Cape's public finances. Fourie and Green (2015) use the *opgaafrolle* to provide a more accurate estimate of the number of Khoesan in the colony, showing that a more accurate reflection of Khoesan participation in the economy changes the findings of earlier work on Cape productivity, GDP and inequality.

All these studies use the census returns as repeated cross-sections. The Cape of Good Hope Panel project aims to transcribe more than five times the number of censuses and link them, for the first time, across time. Below we describe the record linkage strategy we have developed to combine the *opgaafrolle* into what will become the world's longest household-level panel dataset.

Record linkage

Our record linkage process consists of the following steps. After basic data cleaning, candidates for comparison were created (blocking). Actual comparisons and linkage decisions were then made using a classifier which was trained on a manually linked subset of the data. The results of these comparisons were then used to create links between the censuses. Each of these steps is explained in more detail below.[2]

The first step was to clean the data. All names were converted to the same case, encoding issues were fixed, non-letter characters were removed, and all characters were converted to ASCII. This latter step transliterated a few accented characters.[3] Initials were created from the first names. Table 1 provides example records from the cleaned data.

[Table 1 about here.]

Blocking – selecting candidates for more in-depth comparison – was the next step. This was necessary to prevent having to do computationally intensive full comparisons for all combinations of records (Christen 2012). We used the Jaro-Winkler string distance (with the penalty for mismatches in the first four characters set to 0.1) between the men's surnames as our blocking variable (Loo 2014). We selected as candidates all pairs of households whose normalised male surname string distance is less than 0.15.

This blocking strategy was found to be somewhat inefficient as the string distance still had to be calculated between all records. While calculating one string distance is still preferable over computing all string distances used in the in-depth comparison, it is still a computationally intensive step. A blocking strategy based on indexing would therefore be preferable, using for example ages, or soundex-encoded surnames. The *opgaafrolle*, however, do not report ages, and the diversity of origin of the settler population – including Dutch, French, English, and German origins – means that language-specific phonetic encodings are likely to be unreliable (Christen 2012). We do use phonetic string distances in the more detailed comparisons below because phonetically similar spelling variations can be important (for instance exchanging the letters C and K). At that stage of the matching procedure, the use of soundex for the South African population are not problematic because uninformative variables do not end up contributing greatly to the predictions. However, selecting the candidates based on a potentially uninformative variable carries a high risk.

Since we only compare the observations from one year in one district (usually between 1 000 to 1 500 observations) to all the others in the district

2. The R scripts (R Core Team 2015) for the procedure can be found at `https://github.com/rijpma/opgaafrolle/`.

3. This was done because we did not expect accents to be consistently applied between census years.

(the remainder of the 42 354 observations), the computational inefficiency of our procedure does not yet lead to problems. Future expansions of the procedure will, however, probably require an improved blocking strategy. Scalability issues are discussed in more detail below.

Next, we use machine learning techniques to predict whether two candidate records were the same household. To do this, we began by manually creating a training dataset of 454 links based on 608 records from Graaff-Reinet in 1828 and 674 records from 1826. When manually matching individuals the following steps were followed: Names were arranged alphabetically in the two censuses. A large number of straightforward matches could be made on the basis of male first-names and surnames alone, as unique male first-names and surnames often existed across several years. The fact that 80 % of males in our manually matched sample had more than one first name, also aided the matching process. Note that minor spelling variations between names were permitted – e.g. the surname 'Ackerman' could be either 'Akkerman' or 'Ackermann' – but given the same unique first names, a true match was assigned. Knowledge about the idiosyncrasies of the Dutch/Afrikaans naming traditions proved particularly useful, for example, that 'Johan' and 'Jan' are both common diminutives of the first name 'Johannes'.

In some cases, however, owing to the tradition of naming oldest sons after their fathers or paternal grandfathers, certain first names, their ordering, and surnames repeated within a given census year. This necessitated the consideration of the wife's name and surname. Although we did not record how many names could be matched by simply using this procedure, our impression is that at least 70 % of all pairs were successfully matched using only these four variables (names and surnames of husband and wife). Where there was no wife present and two similar names appeared in the same year, manual matching became much more difficult, and subjective. Here we erred on the side of caution; if two names could not be distinguished, we did not attempt a match. Sometimes, though, additional information, like 'junior' or 'widow', would enable disambiguation between identical names. Sometimes a name was unique within a specific district, which allowed us to match the individual (assuming no migration). Occasionally we also considered the number of children in the household or assets owned, although here, too, we erred on the side of caution. Quite frequently, while the husband's name and surname might have been similar across years, the wife's name would be somewhat different: for example, 'Maria Magdelena' might become 'Maria Elizabeth'. We then assumed that these were errors in the transcription process. There were also several cases where it was clear that the wife had died, and that the husband had remarried.

Almost 75 % of names in the 1828 census could be manually linked to 1826. A retention rate of 75 % is surprisingly high in this context given that this was a turbulent period on the Cape frontier, characterised by frequent skirmishes between the settlers, the indigenous Khoesan and amaXhosa,

and the settlers' semi-nomadic lifestyle with poor access to basic services. New migrants constantly entered and exited the district in search of land, or better economic opportunities. As far as we could establish from casual observation, the unmatched consisted of largely two groups: (i) Where a unique name and surname combination only existed in one year (this was the most likely reason for non-matches), and (ii) where two similar names existed in one year matched by similar names in another year but with no additional information to separate the two.

Using the blocking strategy described above, candidates were created for linkage from these 608/675 records to end up at a final training data set with 7 585 candidate links containing 454 true links. To train the models and assess their performance we split this dataset in half. Having a separate test dataset is important to check for overfitting (the tendency to fit the model to the training data's idiosyncrasies rather than general features that will also hold for the rest of the data), a common issue with machine learning algorithms.

Table 2 presents the variables or features of the data that were used in the models. Some of these are commonly used in historical record linkage (Feigenbaum 2016; Goeken et al. 2011), while others are more specific to the *opgaafrolle*. We include string distances between the first names and surnames of the husbands and wives from one year to the next. The old and young indicator variables are also included. The *nrdist* variable is included to capture information about the order of the households in the returns.[4] The wife presence-variables are important to include because the absence of the wife makes it far harder to identify a link.[5] The string distances between the name of the husband and wife is meant to capture changes in recording the wife's maiden name or her husband's surname. Finally, the frequency of each surname was also included as a predictor variable, as common surnames are likely to be less predictive of linkages compared to rarer surnames.[6]

[Table 2 about here.]

We experimented with a number of models (including support vector machines and neural networks), but focus here on two: logistic regression and

4. We experimented with other ways of capturing order information, especially using the string distance of shifted observations, but these did not improve the model's performance.

5. We also tried to use separate models for observations with and without the wives' names to make sure that the importance of the wives' names did not hinder our ability to make links where the wife was absent. However, the minor improvement in performance did not compensate for the additional complexity of having to create matches from two different models.

6. This frequency was calculated on the basis of uniformized surnames, as we want to capture the frequency of the name regardless of small spelling variations. Note that this uniformization procedure was not used in the direct string comparisons of names described earlier (as suggested by Vick and Huynh 2011), because in that case we want to capture the exact difference between two names rather than just the broader class of name they belong to.

the random forest classifier (Breiman 2001; Liaw and Wiener 2002). Logistic regression is included as a high-performing, yet easy-to-interpret classifier (Feigenbaum 2016). The random forest classifier is discussed because it is the best performing model.[7]

Figure 2 presents the regression coefficients for the logistic regression and the importance of variables in the random forest model. The models agree on a number of features. The string distance between the male first names, the initials, and the female last names are important in both models. The low importance of the string distance between male surnames in both models can be explained by the fact that they have already been used to select candidates (blocking). Thus, the male surnames within each block will be similar and contain little further for prediction of matches within the block.

[Figure 2 about here.]

Most important for our purposes is the predictive performance of the models. Since both models give estimates of the probability that a link is true, a threshold at which we declare a link must first be determined. This is done by using the error rates based on the number of false positives (where the model predicts a link that was not present in the training and test data) and false negatives (where the model incorrectly fails to predict a link that was present in the training and test data) as a share of the total number of observations (James et al. 2013).

Figure 3 shows that the error rates are minimised well below the conventional threshold of 0.5. However, because we are more concerned about creating false links than we are about missing true links, we use a more conservative threshold of 0.5. By including the predicted probability (or in the case of the random forest model, "votes") of a true link, future users of the *opgaafrolle* database can increase the required confidence level and exclude less certain links. Note, however, that the number of false negatives increases sharply when we raise the threshold in the random forest model above circa 0.7, meaning one would miss a large numer of true links.

[Figure 3 about here.]

The predictive performance of the two models is further investigated using the confusion matrices in tables 3 and 4. These matrices show the true positives (where the model correctly predicts a link that was present in the training and test data), true negatives (where the model correctly predicts

7. The random forest classifier is an extension of the decision trees classifier which repeatedly segments the data to predict outcomes. Random forest lowers the variance of decision trees by averaging over a large number of trees, each based on a subset of the predictor variables to decrease the correlation between the trees (James et al. 2013).

that a link was not present in the training and test data), false positives, and false negatives for the training and the test data. Because there are many candidates that are not actual links relative to the number of actual links, true negative rates are typically very high in record linkage and not informative about the quality of the matches (Christen 2012). We therefore focus on the sensitivity (the true positives as a share of true positives and the false negatives, also known as the true positive rate) and the precision (the true positives as a share of true positives and false positives).

Both models have a high sensitivity. In case of the logit model it is 87 percent on the training data and 89 percent on the test data (table 3). The random forest model (table 4) performs well on the training data (99.6 percent sensitivity), though its performance is closer to the logit model on the test data (89 percent), a sign of overfitting. Overall, our models perform well compared to other historical record linkage efforts, a point which we explain below.

[Table 3 about here.]

[Table 4 about here.]

Of the two main classifiers tested here, we prefer the random forest model for our linkage procedure. While both classifiers have similar sensitivity on the test data, the main reason for preferring the random forest model is that it has fewer false positives. The precision is 94 % compared to 90 % in the logit model.[8] Including false links is arguably worse than missing true links, since missing observations, while an issue, is a well-researched issue; consider, for instance, the literature on missing data, selection bias, and survey weighting (Little and Rubin 1987; Heckman 1979; Solon, Haider, and Wooldridge 2015; Antonie et al. 2014). On the other hand, there are no satisfying methods to deal with incorrectly-matched observations (a similar point in Antonie et al. 2014). Another reason for preferring the random forest classifier is that the decision trees on which it is based are well suited to find any non-linearities and interactions that might exist in the data (Hastie, Tibshirani, and Friedman 2009, 587). This should be relevant for record linkage. For example, the string distance between the men's first names should be more important if there is a large string distance between the wives' surnames (due to remarriage or a change from maiden name to husbands name). We should be less tolerant of large string distances in these cases.

While we prefer the random forest classifier, it should be noted that logistic regression performs surprisingly well here (see also Feigenbaum 2016), and could be preferable if interpretability of the classifier is important.

8. With precisions of 91 % and 92 %, the support vector machine and neural net classifiers also performed worse than the random forest model.

By comparison, a manually weighted combination of string distances performed far worse than all models we tested. To create a set of weights that reflected what we deemed to be important in the linking procedure, we drew from our experience of manually matching the training data.[9] We identified only 122 out of 229 true matches (53 %) in the training data correctly. The false positive and false negative rate for this procedure were 7 % and 47 % respectively. Clearly, even relatively straightforward classifiers are superior in the case of the *opgaafrolle* data.[10]

It is useful to investigate the false positives created by the random forest classifier in greater detail to understand in what cases our procedure fails. The preferred model creates 25 false positives in the test data (see appendix, section B.1). Of these, 16 lack information on the wife in one or both records. Based on only the husband's name, all but one of these could be true links omitted in the creation of the training data, but were omitted because the link was ambiguous. Of the remaining nine observations for which information about the wife is available, seven are likely correct, but were not recognised as true links in the construction of the training data. On the basis of differences in the wives' surnames, the remaining two seem to be incorrect links. In short, many of the false positives could be actual links in the training data, while a small number of false positives that do remain are truly wrong.

The random forest model is not dependent on including all of the variables in table 2. In appendix C, tables C.1 and C2 show that the AUC, a measure of the rate of true positives to the rate of false positives, remains close to the preferred model's AUC of 0.94.[11]

Since we are building a panel we would expect to find one household in multiple years. This means the application of the model outlined above as well as the construction of the panel out of the suggested links require us to deal with a number of further issues.

Our approach has been to apply our linkage from each year to all earlier years (e.g., 1828 to 1826, 1825, …; 1826 to 1825, 1824, …, and so forth). We use this backwards procedure to exploit the fact that information is typically better for the more recent records. Once we move to the next base year for comparing, it is unnecessary to include the previous base year in the comparison. The string distance relations we use are symmetric, so if two years haven been compared in one direction they do not need to be compared again. This allows us to economize on the number of comparisons

9. Weights: mlastdist 0.27; mfirstdist 0.16; minidist 0.05; winidist 0.03; wlastdist 0.14; wfirstdist 0.07; mlastsdx 0.11; mfirstsdx 0.05; wlastsdx 0.05; wfirstsdx 0.03; mtchs 0.03. Other weighting schemes were tried, but performance was generally similar or worse.

10. The sensitivity on the test data of the support vector machine (Venables and Ripley 2002) and a neural net classifier (Meyer et al. 2017) was 86 % and 85 % respectively.

11. We were not able to test the exclusion of every combination of variables because more than 8 million combinations are possible.

we need to make. Working from one base year has the additional advantage of keeping the string distance matrix that serves as the basis for creating candidates small.

Working in a panel setting also creates the possibility of creating incomplete series of links (where for example observations 1 and 3, and 2 and 3 are matched, but observations 1 and 2 are not). In dealing with this issue we have been permissive and have used the linkage information from various base-years to connect disjointed series. The main reason for doing this is that finding that a household is linked between certain census-years means this household should probably be present in other years as well, but has been missed as a result of a transcription error.

Model evaluation

We now turn to the evaluation of the dataset. While success on the test data is the only certain assessment of the linkage procedure, manual inspection of the series is also useful. This has so far not revealed obviously incorrect linkages. On average, the share of households in any given year being linked to another is over 80 % (figure 4). Usually the figure is somewhat higher than that, but linking individuals to and from 1801 and 1803 proved more difficult because the *opgaafrolle* in those years rarely if ever contain information on wives. In these years linkage drops to 67 % of the households and many of our series end in one of these two years.

[Figure 4 about here.]

The longest series created are 32 observations long (figure 5). There are 3 of these series in the data, making for 96 observations in total. A shorter minimum length of linkages yields more observations: a minimum series length of 23 yields over 100 series covering almost 2 500 observations. A lower threshold of a minimum series length of 9 yields over 2 000 series, covering almost 20 000 observations. With the Graaff-Reinet *opgaafrolle* for 1787–1828 containing 42 354 observations, this means that almost half of the observations are contained in these moderately long series. If series of at least length three are sufficient for an intended analysis, over two thirds of the dataset (more than 10 000 series and over 30 000 observations) would be covered.

[Figure 5 about here.]

Another way to assess the quality of the created panel is to inspect the correlation of variables within linked households over time (figure 6). To check the correlations over time, we compare the number of settler children (children who are not Khoesan or slaves) and cattle on the farms with their

respective one year lags. As expected, it can clearly be seen that the panel displays a strong correlation between the variables and their lags (Pearson correlation coefficient are 0.9 for the number of settler children and 0.83 for the number of cattle).

[Figure 6 about here.]

Differences in the data between the linked and unlinked observations can reveal if there are any biases in the resulting data, as well as provide a check on the characteristics that drive the record linkage procedure (that is: are the biases as we expect them to be?). Figure 7 shows the distribution of the number of cattle owned per household broken down by the length of the created series. It can be seen that the distribution for short links is similar to the one for unlinked households. However, longer linked series show fewer households with no cattle (the spike at the left) and in the case of the longest linked series (bottom right panel), more households with a high number of cattle. One possible reason for this pattern is that households with a high number of cattle were less likely to migrate compared to households without such valuable assets. They are therefore more likely to be captured by our linkage procedure for the Graaff-Reinet territory. Wealthier families in the *opgaafrolle* are more likely to have characteristics that make them easier to link. Economic and social historians may be familiar with the idea that rare surnames are easier to link and can also be informative of economic outcomes (Guell, Mora, and Telmer 2014; Clark et al. 2015). Besides this general phenomenon, it is particularly important in our case that household heads with large farms were also more likely to be married (households with a wife present on average had more than twice as many cattle in Graaff Reinet), thus increasing the linkage rate through the information contained in the wives' names.

[Figure 7 about here.]

To deal with these biases we follow Goeken et al. (2011) and provide weights based on the inverse of the linking rate of variables showing a strong linking gradient. We have considered two weighting variables: one based on the presence of a wife (indirectly capturing marital status), and one on the name frequency. The weight based on the presence of a wife is probably the most important as it strongly predicts linkage (linkage rates are 88 % versus 77 %) and is probably correlated with many outcomes of interest (Solon, Haider, and Wooldridge 2015). Surname frequency shows a strong gradient at frequencies lower than 10, with rarer surnames less likely to be linked (see figure 8 below). Above 10, no linkage gradient can be observed. Altogether, this means that rare surnames are not improving our linkage rates, but rather only capture the fact that some names were so infrequent

that they were unlikely to be linked. Moreover, name frequency is probably not as highly correlated with the outcome variables as the presence of a wife. While we provide weights based on four categories of name frequencies (1, 2–5, 6–10, greater than 10) in the final dataset, we suggest that the presence of a wife is the most important linking variable. Finally, because the panel contains the unlinked individuals as well as all the variables used to construct the links, users of the data can create further weighting variables to meet the specific requirements of their analyses.

While our linked dataset is not without biases, we have been able to create a number of quality links. The sensitivity and precision of the model on the training and test data are high and at more than 80 %, the overall linkage rate on the complete data is also high (cf. Massey 2017; Feigenbaum 2016). This is especially striking given the fact that the *opgaafrolle* lack a number of the conventionally important variables for record linkage such as age or place of birth. The transcription of names is clearly not the reason. Small spelling variants between the names from one year to the next are frequent. We also do not think the naming practices in South Africa are to credit for the high linkage rates. Figure 8 shows the distribution of surnames over our entire panel as well as the linkage rate by frequency. It can be seen that linkage rates are actually lower for infrequent names.

[Figure 8 about here.]

The reason we are able to achieve high linkage rates with a low number of false positives despite the existence of common surnames is that we also use other household information to make the links (see also Fu et al. 2014). While a surname and a first name of the husband are frequently unable to provide an unambiguous link, adding the first name and surname of the wife often allow us to make that distinction. Figure 4 confirms this: 1801 and 1803 are years in which the wives' names are only recorded infrequently in the *opgaafrolle* and our linkage rate drops below 70 %, in line with individual-level historical record linkage strategies. Additionally, the fact that our data are close to annual means that events such as migration, death, or changes in the composition of the household are less frequent than in datasets with a greater time gap. This will also increase linkage rates.

Linking external data: South African Families

Thus far, our record linkage efforts have used relatively consistent source material: the various years of the *opgaafrolle*. Most information in one year is usually also present in the next. However, many record linkage tasks will have to deal with more heterogeneous data. The data in the *opgaafrolle* also lack certain information that would be useful for analyses. Most importantly,

it is lacking demographic data as no ages, dates of marriage, details on childbirths, or family links are provided. The latter would allow us to know the composition of the households in more detail and also provide insight into intergenerational mechanisms of economic success.

For these reasons, we also attempt to link the *opgaafrolle* to an outside dataset. One such source is a large genealogical database of the South African settler population, *South African Families* (SAF, see Cilliers 2016). SAF is a complete register of European settlers and their descendants at the Cape spanning over two hundred and fifty years from settlement in 1652 to the beginning of the 20th century. Unique in its size and scope across time and space, it offers a longitudinal account of individual life histories of white settlers.

Linking this dataset to the *opgaafrolle* requires some modifications to our strategy used to link within the *opgaafrolle*. These modifications concern the creation of a linkage window to prevent having to match against a large number of implausible candidates and dealing with the fact that we cannot recreate all the *opgaafolle* variables in the SAF-data. Other than these modifications, we have used the same procedure as outlined above.

Information contained in the genealogies includes names of all family members (also maiden names of wives), dates and locations for birth, baptism, marriage, and death, as well as occasionally occupations. However, not all entries contain complete information for every event. While close to two thirds of the entries contain a birth or a baptism date, only one quarter contains a death date, and less than one fifth contains a marriage date. Nevertheless, these dates, where available, allow for an additional selection step, in which individuals in the genealogies who could not possibly be a match to the *opgaafrolle* given their birth and death dates can be excluded.[12] This is important for the prevention of false negatives as well as computational feasibility given the size of SAF. We use the date-of-death of the husband (after the year of the *opgaafrol*, but no later than 100 years after) and date-of-birth (16 years before the year of the *opgaafrol* tax, but no earlier than 100 years before) to construct our linking window. This step reduces the size of potential matches from over 670 000 (the full SAF database), to 153 156. Date-of-death and date-of-birth are, however, often missing (119 548 and 110 232 observations respectively). Further selection based on the date of birth of children in a household reduces the number of individuals in SAF further. We include men whose children were born between 48 years before the year of the *opgaafrol* (meaning a male fathering a child at age 12 would be 60 at the time of the *opgaafrol*) or 88 years after

12. Since the *opgaafrollen* for Graaff-Reinet only cover the period 1787-1828 and SAF spans 1652-2012 there are many SAF-persons that could not be a match to an opgaafrol-person given that some will already be dead before the opgaafrollen begin, while others will not have been born until long after the *opgaafrollen* end. We therefore only consider SAF-persons who were conceivably alive during the *opgaafrollen* period.

the year of the *opgaafrol* (a male household head aged 12 at the time of the *opgaafrol* could conceivably still father a child at that time).

Because some of the variables are not available or would not contain a great deal of information in SAF, the full random forest classifier we use to predict linkages within the *opgaafrolle* cannot be used here. One set of variables we exclude are cross-spouse surname string distances. In the within-*opgaafrolle* linking these variables were meant to capture year-to-year variation in wife's surname – especially whether the wife had dropped her maiden name. But since wives are reported only once with their maiden name in the genealogies, such a distance measure would have no meaning in the linking between SAF and the *opgaafrolle*. We have also excluded the old and young dummies since it is unclear when a person in the *opgaafrolle* qualified as either, making it difficult to construct a similar variable for the genealogies. The order of the observations is also excluded. This was meant to capture the order in which households were recorded in the *opgaafrolle*, which has no equivalent in the genealogies. Also excluded were the wine producer and district dummy variables.

This model was trained and evaluated on the *opgaafrolle* training data (that is: no new training and test data was created for the SAF-to-*opgaafrolle* linkage). While this model has to work with less information, it still performs well on the *opgaafrolle* training data. 96 % of the true matches are correctly classified in the training data and 87 % of the true matches are correctly classified in the test data. This is only slightly lower sensitivity than the full model (99 % and 88 %). The number of false negatives on the test data (14) is also similar to the full model. The fact that the name distances are the most important predictor variables in the original model explains why the model's predictive power deteriorates only slightly while omitting a number of variables.

We select the best match from the genealogies for each person in each year. Because the *opgaafrolle* have already been linked over time, this allows for the possibility that the same *opgaafrol*-person over time (according to our linkage procedure) gets linked to multiple genealogy-persons. Reassuringly, this does not happen often (1 729 out of 21 496 linked individuals in the *opgaafrolle*). We chose to drop these links to the genealogies altogether. However, it would probably be possible to disambiguate a few of these links. Figure D.1 in appendix D shows that the differences of the mean linking scores of the doubly linked persons can be substantial. While the largest differences can probably be resolved, developing a systematic rule would require a new training data set which is beyond the scope of this paper.

The share of observations in the *opgaafrolle* that our procedure managed to link is shown in figure 9. Again, it shows that in the years where the names of the wives are usually absent (1801 and 1803), linkage is difficult. Elsewhere, the share of observations matched is fairly high in the latest years (over 60 percent), but decreases to 40 percent as we go further back in time.

One reason for this is that the number of persons in the genealogies increases exponentially over time (Cilliers 2016), so the further back in time we try to link, the smaller the number of candidates will be. Overall, it is feasible to scale the procedure to larger datasets.

[Figure 9 about here.]

Overall, the efforts to link the *opgaafrolle* to the genealogy data have been successful in two ways. First, it adds valuable demographic data to the economic information contained in the *opgaafrolle* panel. We are hopeful that additional supplementary data can be linked to the final panel as well. Second, it shows that our linkage strategy is more widely applicable than to the *opgaafrolle* alone. We think that any household-level data that contains the surnames and first names of the husband and the wife and uses the naming conventions of South Africa in the eighteenth and nineteenth century (a mix of Dutch, English, German, and French) could potentially be linked using our strategy and classifier. Creating additional training data for these datasets could improve performance further.

Model scalability

While we are generally satisfied with the record linkage procedure, a few issues with scalability remain. Memory usage is especially an issue because we have to use string distances for our blocking strategy. By comparing one base year of observations with the rest of the dataset rather than all observations at once, we have already limited the computational burden this imposes. Rather than a 13 GB object containing all the string distances that we would get if we compare all the 42 354 male surnames at once, we are now left with an object smaller than 1 GB. This means that the creation of each distance matrix takes less than a minute on commodity hardware.[13]

Once we begin to expand the dataset, however, computation may become more difficult. Currently, the Graaff-Reinet *opgaafrolle* contain 42 354 observations, spread over 35 years and 1–14 sub-districts (depending on the year). The eventual goal of the Cape of Good Hope Panel project is to cover 150 years for each of the territories that make up the Cape Colony. If we take an upper limit to the comparison depth of 60 years and restrict ourselves to within-district linkage, this should still keep the size of the candidates string distance, currently our most computationally intensive step in the procedure, below 1.5 GB.

However, computing difficulties might arise as the project expands. Neighbouring territories are one such complication. Borders were not stable, so

13. Of course, such a matrix needs to be created as many times as there are base years, so there is no advantage in processing time to doing the procedure one year at a time. However, the advantage of keeping the base years separate for memory limits is substantial.

we can expect to find households appearing in different territories in later years even if people did not migrate. This could increase the number of records that needs to be considered, though a detailed graph of neighbouring (sub-)districts would help. Should we want to follow migrants in the panel, record linkage will become more computationally expensive as well. This means we would have to search for ways to make our record linkage more scalable. Options to do this include: using a better indexing method (blocking) for candidate selection, parallelisation of string distance matrix calculation, parallelisation of string comparisons, and parallelisation of estimation of random forest model and its predictions.

Conclusion

This paper has explained the record linkage strategy used to create the Cape of Good Hope Panel. The basis for this panel are the *opgaafrolle*, annual census returns for settlers of the Cape Colony, an area at the southern tip of Africa settled by Europeans in the seventeenth century. The tax censuses contain valuable information on agricultural production and demographic characteristics of the settlers. To get the most out of these censuses, it is necessary to create a panel by linking the households in the repeated cross-sectional censuses. This paper constructs a matching algorithm to link settlers residing in one region of the Cape Colony, Graaff-Reinet, for the years 1787-1828.

The first step was to manually create 454 links between 608 and 674 households across two years. From this starting point we created linkage candidates based on the male surname string distances. This training dataset was then used to estimate a model to classify new, unlinked observations. We preferred a random forest classifier over alternative classifiers, most notably logistic regression, because it resulted in fewer false positives. The model takes on board as much information in the *opgaafrolle* as possible, but the string distances of the husband and wife names are the strongest predictors of a link. The training and test data show that our model has a sensitivity of 90 % and a precision of 94 % meaning that it correctly classifies 90 percent of the manually identified links and has an acceptable rate of false positives. This high linkage rate is due to our use of household-level characteristics and near-annual data.

The number of links created in the resulting dataset is more than sufficient for most analyses. Three-quarters of the dataset consists of series of at least four observations per individual and nearly half is of length nine or more. Moreover, the created links show the expected correlations in demographic or economic variables. It is however important to be aware of the biases that are created in the linkage process. Notably, linkage rates of households with a married couple were 10 percentage points higher, skew-

ing the subset of the longest series (length eight or more) compared to the overall dataset. Weights are provided in the data to correct for this issue in analyses.

We have also explored the possibilities of linking the Cape of Good Hope Panel to a genealogical database (SAF). The SAF adds valuable demographic information to the *opgaafrolle*. Matching is done using the same training data and random forest classifier as for the within-*opgaafrolle* linkage task, but using fewer variables to match the variables in the genealogical data. Nonetheless, 87 percent of the links are correctly identified in the test data and this allows us to link a person from the genealogies to half the households in the *opgaafrolle*. This means our approach is more generally applicable and should enable us to create a more detailed and complete dataset of Cape Colony settlers.

We expect our method to be able to scale beyond Graaff-Reinet, the district we analysed here. If we want to follow households across districts, however, it will be necessary to compare far more records at once. Memory usage would especially become a concern and for this reason, a better blocking procedure than male surname string distance will be necessary.

References

Antonie, L., K. Inwood, D. J. Lizotte, and J. AndrewRoss. 2014. "Tracking people over time in 19th century Canada for longitudinal analysis." *Machine Learning* 95, no. 1 (): 129–146.

Bloothooft, G., P. Christen, K. Mandemakers, and M. Schraagen. 2015. *Population Reconstruction.* Cham: Springer.

Breiman, L. 2001. "Random forests." *Machine learning* 45 (1): 5–32.

Christen, P. 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media.

Cilliers, J. A. 2016. "A demographic history of settler South Africa." Thesis, Stellenbosch University. `http://ir.nrf.ac.za/handle/10907/497`.

Clark, G., N. Cummins, Y. Hao, and D. D. Vidal. 2015. "Surnames: A new source for the history of social mobility." *Explorations in Economic History* 55 (1): 3–24.

Dong, H., C. Campbell, S. Kurosu, W. Yang, and J. Z. Lee. 2015. "New sources for comparative social science: Historical population panel data from East Asia." *Demography* 52 (3): 1061–1088.

Feigenbaum, J. J. 2016. "Automated census record linking: A machine learning approach." http://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking.

Ferrie, J. P. 1996. "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29, no. 4 (): 141–156. https://doi.org/10.1080/01615440.1996.10112735.

Fourie, J. 2014. "The remarkable wealth of the Dutch Cape Colony: measurements from eighteenth-century probate inventories." *Economic History Review* 66 (2): 419–448.

———. 2016. "The data revolution in African economic history." *Journal of Interdisciplinary History* 47 (2): 193–212.

Fourie, J., and E. Green. 2015. "The missing people: accounting for the productivity of indigenous populations in Cape Colonial History." *Journal of African History* 56 (2): 195–215.

Fourie, J., A. Jansen, and K. Siebrits. 2013. "Public finances and private company rule: The Dutch Cape Colony (1652-1795)." *New Contree* 68:1–22.

Fourie, J., and D. Von Fintel. 2010. "The dynamics of inequality in a newly settled, pre-industrial society: the case of the Cape Colony." *Cliometrica* 4 (3): 229–267.

———. 2011. "A history with evidence: Income inequality in the Dutch Cape colony." *Economic History of Developing Regions* 26 (1): 16–48.

———. 2014. "Settler skills and colonial development: the Huguenot wine-makers in eighteenth-century Dutch South Africa." *Economic History Review* 67 (4): 932–963.

Fu, Z., H. Boot, P. Christen, and J. Zhou. 2014. "Automatic Record Linkage of Individuals and Households in Historical Census Data." *International Journal of Humanities and Arts Computing* 8, no. 2 (): 204–225.

Goeken, R., L. Huynh, T. A. Lynch, and R. Vick. 2011. "New Methods of Census Record Linking." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44, no. 1 (): 7–14. https://doi.org/10.1080/01615440.2010.517152.

Guelke, L., and R. Shell. 1983. "An early colonial landed gentry: land and wealth in the Cape Colony 1682–1731." *Journal of Historical Geography* 9 (3): 265–286.

Guell, M., J. V. R. Mora, and C. I. Telmer. 2014. "The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating." *The Review of Economic Studies* 82 (2): 693–735.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Second edition, corrected 7th printing. Springer series in statistics. New York: Springer.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–161.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning.* Vol. 6. Springer.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. `http://CRAN.R-project.org/doc/Rnews/`.

Little, R. J., and D. B. Rubin. 1987. *Statistical analysis with missing data.* New York: Wiley New York.

Loo, M. P. J. van der. 2014. "The stringdist package for approximate string matching." *The R Journal* 6 (1): 111–122. `http://CRAN.R-project.org/package=stringdist`.

Massey, C. G. 2017. "Playing with matches: An assessment of accuracy in linked historical data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* (): 1–15. `http://dx.doi.org/10.1080/01615440.2017.1288598`.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2017. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* `https://CRAN.R-project.org/package=e1071`.

Potgieter, M., and J. Visagie. 1974. *Inventaris van Opgaafrolle.* Cape Town, South Africa: Cape Town Archives Repository.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Ross, R. 1983. "The rise of the Cape gentry." *Journal of Southern African Studies* 9 (2): 193–217.

Ruggles, S. 2012. "The Future of Historical Family Demography." *Annual Review of Sociology* 38 (1): 423–441. `http://dx.doi.org/10.1146/annurev-soc-071811-145533`.

Ruggles, S. 2014. "Big microdata for population research." *Demography* 51 (1): 287–297.

Solon, G., S. J. Haider, and J. M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50, no. 2 (): 301–316.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth. New York: Springer. `http://www.stats.ox.ac.uk/pub/MASS4`.

Vick, R., and L. Huynh. 2011. "The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44, no. 1 (): 15–24. `http://dx.doi.org/10.1080/01615440.2010.514849`.

Von Fintel, D., S. Du Plessis, and A. Jansen. 2013. "The wealth of Cape Colony widows: inheritance laws and investment responses following male death in the 17th and 18th centuries." *Economic History of Developing Regions* 28 (1): 87–108.

Wisselgren, M. J., S. Edvinsson, M. Berggren, and M. Larsson. 2014. "Testing Methods of Record Linkage on Swedish Censuses." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 47, no. 3 (): 138–151. `https://doi.org/10.1080/01615440.2014.913967`.

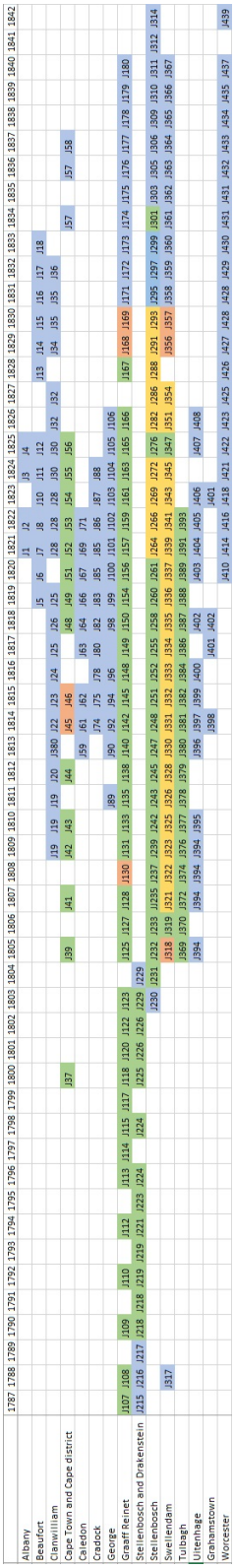

Figure 1: The series of *opgaafrolle*, 1787-1842, by district. Fully transcribed censuses in green, partially transcribed ones in yellow, illegible ones in red and the un-transcribed census returns in blue.

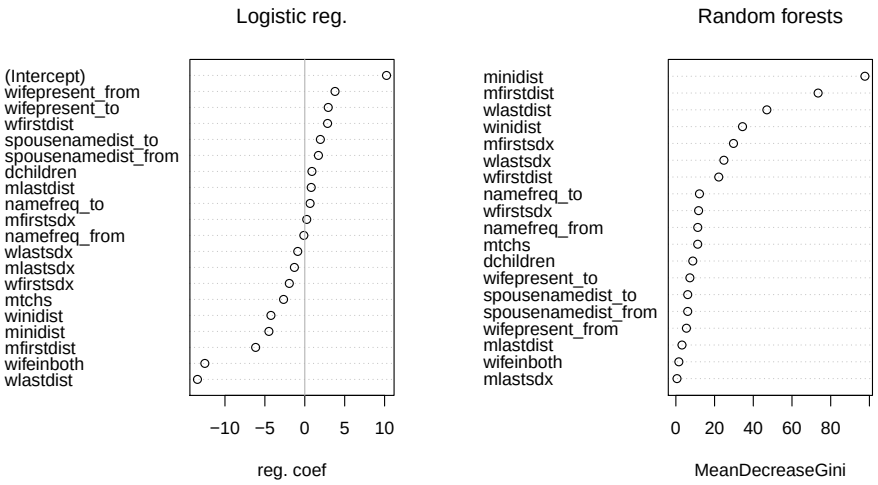

Figure 2: Regression coefficients for logistic regression (left panel) and variable importance plot for random forest model (right panel).

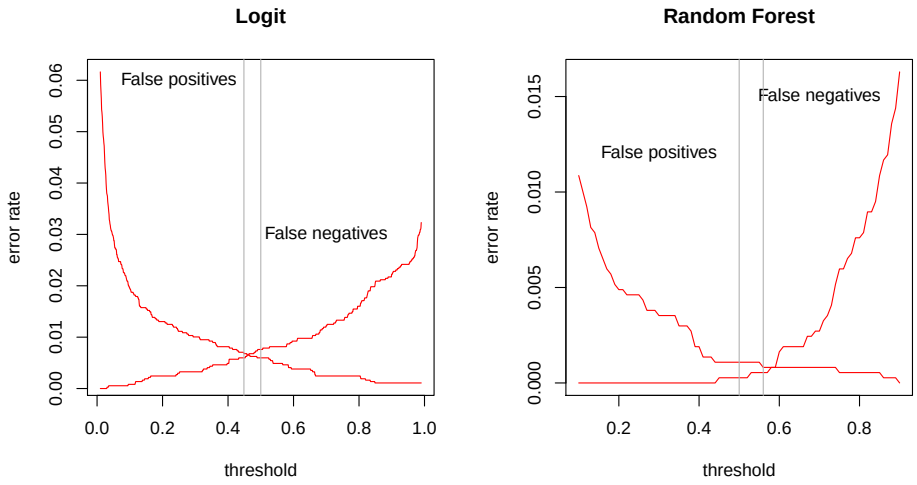

Figure 3: Errors as share of total candidates in the training data as a function of the threshold for logistic regression (left panel) and for random forest (right panel) model. Vertical reference lines at error rate minimising-vote share and 0.5.

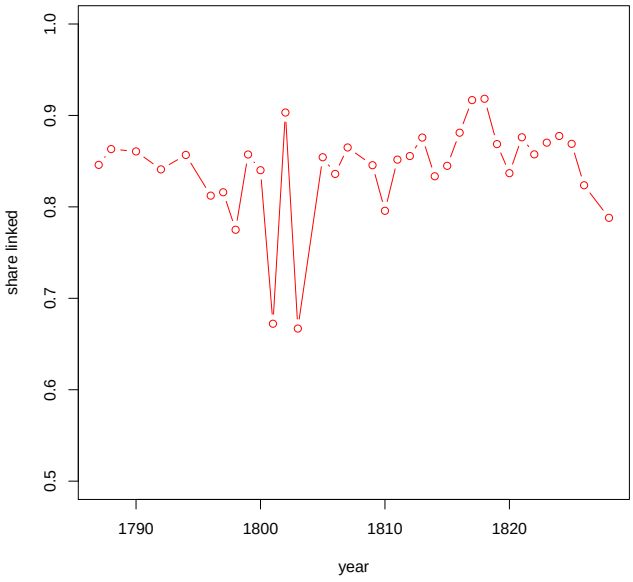

Figure 4: Share of households linked by year.

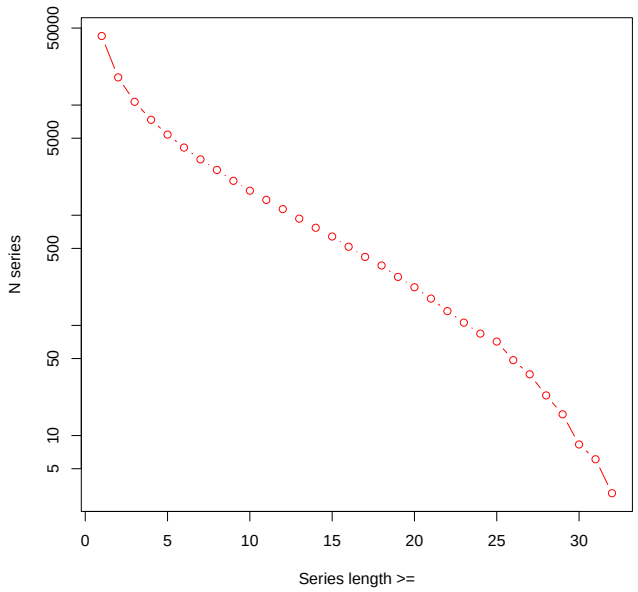

Figure 5: Cumulative links in *opgaafrolle* panel.

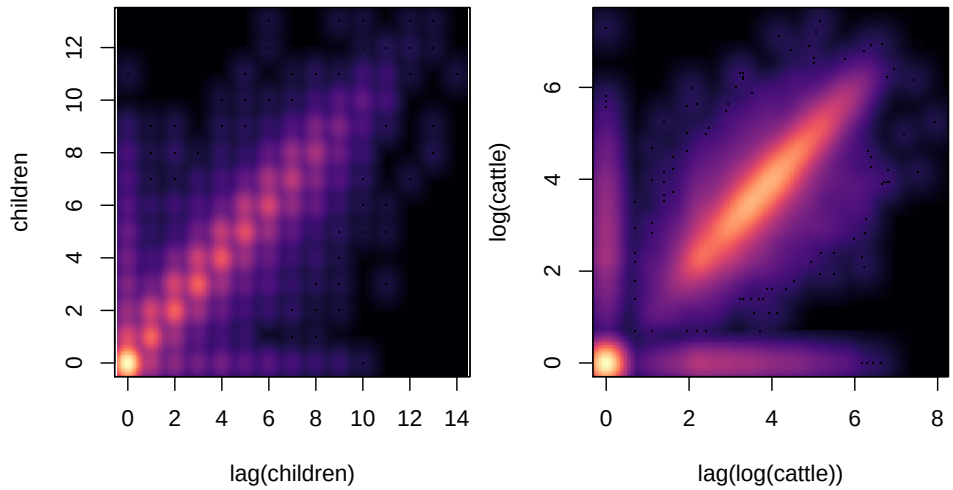

Figure 6: Smoothed scatter plot of the number of children v. lag of number children (left) and log number of cattle v. lag of log number of cattle (right) in panel created through record linkage.

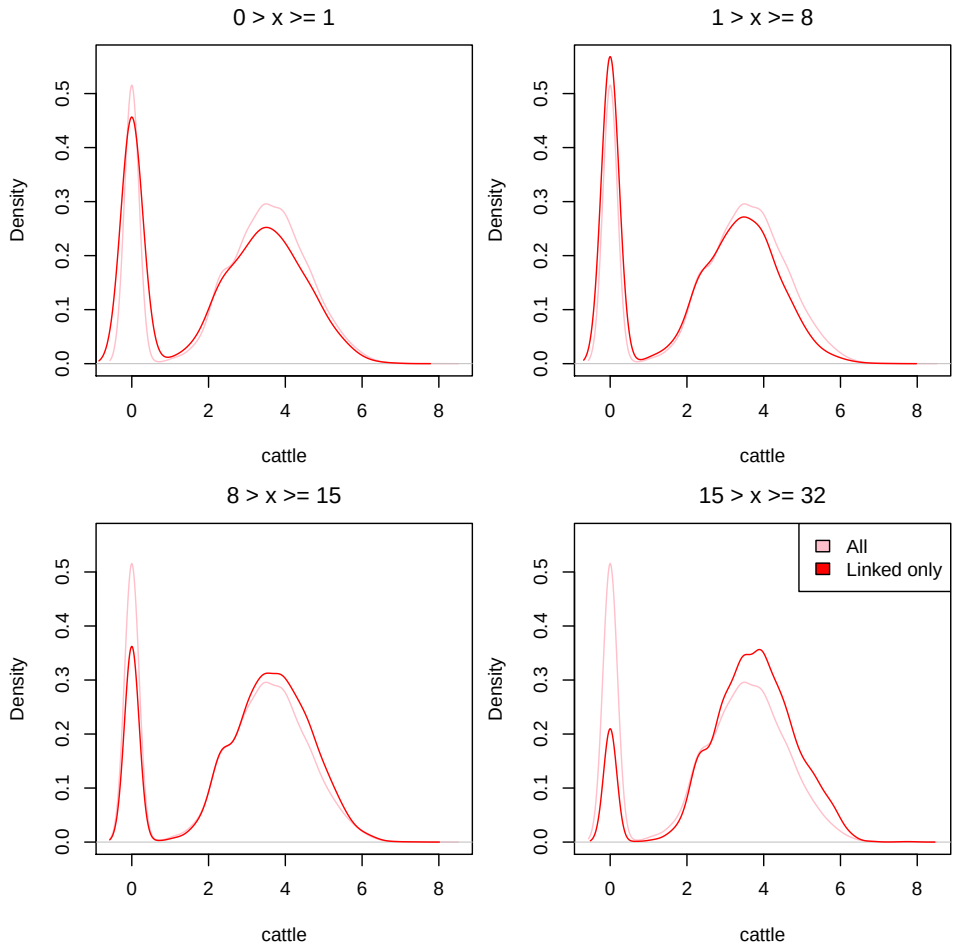

Figure 7: Distribution of log number of cattle by length of link.

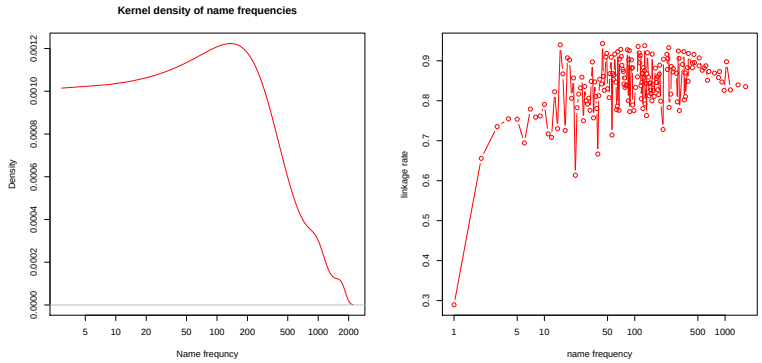

Figure 8: Kernel density of name frequency in the Graaff Reinet district and linkage rate by name frequency. Name frequency is plotted on a logarithmic axis.

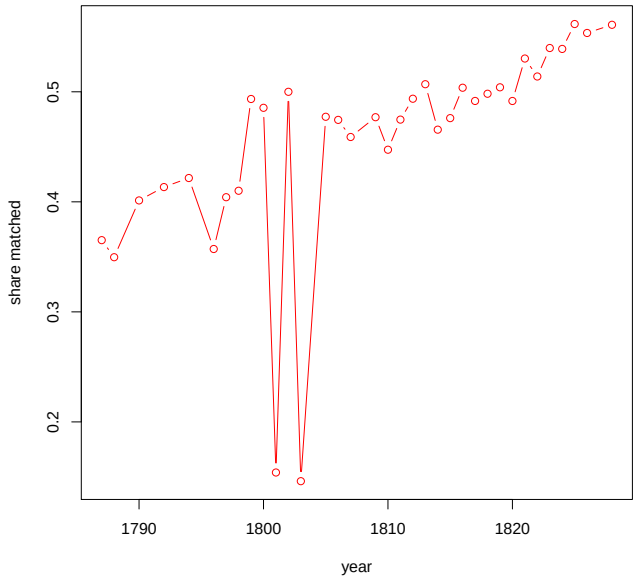

Figure 9: Share of observations in *opgaafrolle* that is linked to an observation in the genealogies by year.

year	mfirst	mlast	wfirst	wlast	old	...
1828	dirk jacobus	coetzee	engela geertruida	grobler	0	...
1828	jan willem	mienie	anna maria	els	0	...
1828	alewijn petrus johannes	van heerden	anna willemina hendrina	lubbe	0	...
...	...	...	...	...	...	...
1826	johannes willem	minnie	anna maria	els	0	...
...	...	...	...	...	...	...
1824	dirk jacobus	koetzee	engela geertruida	grobler	0	...
...	...	...	...	...	...	...

Table 1: Example records from Graaff-Reinet *opgaafrolle*

variable	explanation
namefreq_from/to	Frequency of name and similar variants in full *opgaafrolle* dataset
mlastdist	Jaro-Winkler string distance between husbands' last names.
mfirstdist	Jaro-Winkler string distance between husbands' first names.
minidist	Jaro-Winkler string distance between husbands' initials.
wlastdist	Jaro-Winkler string distance between wives' last names.
wfirstdist	Jaro-Winkler string distance between husbands' first names.
winidist	Jaro-Winkler string distance between husbands' initials.
mlastsdx	Soundex string distance between husbands' last names.
mfirstsdx	Soundex string distance between husbands' first names.
wlastsdx	Soundex string distance between wives' last names.
wfirstsdx	Soundex string distance between wives' first names.
nrdist	Difference between position in year's *opgaafrol*.
bothyoung	Both individuals are identified as young.
bothold	Both individuals are identified as old.
dchildren	Difference in number of children present in households
spousenamedist_from/to	Jaro-Winkler string distance between husband and spouse surname
wife_present_from/to	Wife present in the "from" or "to" record
wifeinboth	Wife present in both records.
bothwineprod	Both records are indicated as wine producers (they have
mtchs	Number of candidates for this record.

Table 2: Description of variables used in the record comparisons.

		Predicted			
		Train		Test	
		FALSE	TRUE	FALSE	TRUE
Actual	0	3432	22	3558	20
	1	28	201	23	189

Table 3: Confusion matrix for logit models

		Predicted			
		Train		Test	
		FALSE	TRUE	FALSE	TRUE
Actual	0	3450	4	3565	13
	1	1	228	23	189

Table 4: Confusion matrix for random forest model

A Logistic regression model

	Model 1
(Intercept)	10.24 (3.76)**
wifepresent_from	3.79 (0.73)***
spousenamedist_from	1.71 (1.20)
namefreq_from	−0.12 (5.23)
wifepresent_to	2.94 (0.83)***
spousenamedist_to	1.95 (1.48)
namefreq_to	0.67 (5.24)
mlastdist	0.81 (1.23)
mfirstdist	−6.16 (1.89)**
minidist	−4.49 (0.97)***
wlastdist	−13.45 (4.62)**
wfirstdist	2.86 (2.37)
winidist	−4.24 (1.16)***
mlastsdx	−1.30 (1.12)
mfirstsdx	0.25 (0.54)
wlastsdx	−0.88 (1.61)
wfirstsdx	−1.94 (1.02)
wifeinboth	−12.52 (3.46)***
dchildren	0.90 (1.79)
mtchs	−2.66 (1.07)*
AIC	281.10
BIC	405.33
Log Likelihood	-120.55
Deviance	241.10
Num. obs.	3683

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table A.1: Logistical regression predicting record matches

B False positives

mlast_from	mfirst_from	wlast_from	wfirst_from	mlast_to	mfirst_to	wlast_to	wfirst_to
coetzee	a. p.	potgieter	anna magdalena	coetzee	hendrik petrus	potgieter	anna magdalena
de villiers	jacobus jonathan			de villiers	jacobus stephanus	van aswegen	cornelia margaretha
jacobs	willem hendrik			jacobs	willem hendrik	barnard	maria elisabeth
coetzee	gert lodewijk	kunn	magdalena	coetzee	gerrit lodewyk		
cloete	jan hendrik	rabie	margaretha	cloete	jan hendrik	margaret	rabie
knoetze	frederik willem	oosthuizen	jacoba m.	knoetze	willem	oosthuisen	elisabeth
bezuidenhout	wijnand frederik	swanepoel	cornelia	bezuidenhout	wynand frederik		
booijens	petrus johannes			booyens	petrus johannes	breitenbach	maria magdalena
geering	willem george			geere	willem		
geel	carl jacobus			geel	carel francois		
coetzee	johannes	kruger	francina anna elisabeth	coetzee	johannes		
coetzee	stephanus jacobus			coetzee	stephanus jacobus	van der walt	susanna maria lacya
louw	pieter johannes	van loggenberg	anna elizabeth	louw	petrus johannes		
kruger	g. h. i.	coetzee	anna johanna	kruger	gerrit hendrik jacobus	coetzee	anna johanna
erasmus	johannes	venter	maria claudina	erasmus	johannes		
lessing	isaac jacobus	kruger	elizabeth josina lacya	lessing	johan jacobus		
kruger	stephanus johannes	steenkamp	sophia margaretha	kruger	petrus johannes	steenkamp	sophia margretha
brits	willem marthinus	brits	johanna adriana	brits	willem maartens		
beetje	johannes andries	joosten	magdalena catharina	beetge	johannes andries	van der linde	sara geertruida jacoba
badenhorst	hendrik johannes	nagel	catharina florina	badenhorst	hendrik johannes		
davel	cornelius hendrik			davel	cornelis hendrik	aucamp	maria catharina
liebenberg	christiaan jacobus	liebenberg	susanna carolina	liebenberg	christiaan jacobus		
burger	andries petrus	pienaar	anna sophia	burger	andries jacobus	pienaar	am. sophia johanna

Table B.1: False positives created by random forest classifier

C Sensitivity analysis

V1	AUC
minidist	0.93
wlastdist	0.93
wfirstdist	0.93
wlastsdx	0.93
winidist	0.93
wfirstsdx	0.93
bothyoung	0.93
bothwineprod	0.93
spousenamedist_from	0.94
wifepresent_from	0.94
wifepresent_to	0.94
spousenamedist_to	0.94
mfirstdist	0.94
correct	0.94
mlastdist	0.94
dchildren	0.94
namefreq_to	0.94
mlastsdx	0.94
wifeinboth	0.94
mtchs	0.94
nrdist	0.94
mfirstsdx	0.94
bothold	0.94

Table C.1: AUC after omitting one variable

V1	V2	AUC
mfirstdist	minidist	0.90
minidist	wlastsdx	0.91
wlastdist	wlastsdx	0.91
minidist	mtchs	0.92
minidist	winidist	0.92
minidist	wlastdist	0.92
minidist	wfirstsdx	0.92
correct	minidist	0.92
minidist	mlastsdx	0.92
wifepresent_from	winidist	0.92
wifepresent_to	winidist	0.93
minidist	wifeinboth	0.93
spousenamedist_to	minidist	0.93
namefreq_to	minidist	0.93
winidist	wlastsdx	0.93
wifepresent_from	wifepresent_to	0.93
wifepresent_from	minidist	0.93
minidist	wfirstdist	0.93
wfirstdist	winidist	0.93
spousenamedist_from	minidist	0.93

Table C2: AUC after omitting two variables (full model AUC: 0.94). Note: only 20 lowest AUC values reported.

D Genealogy linking

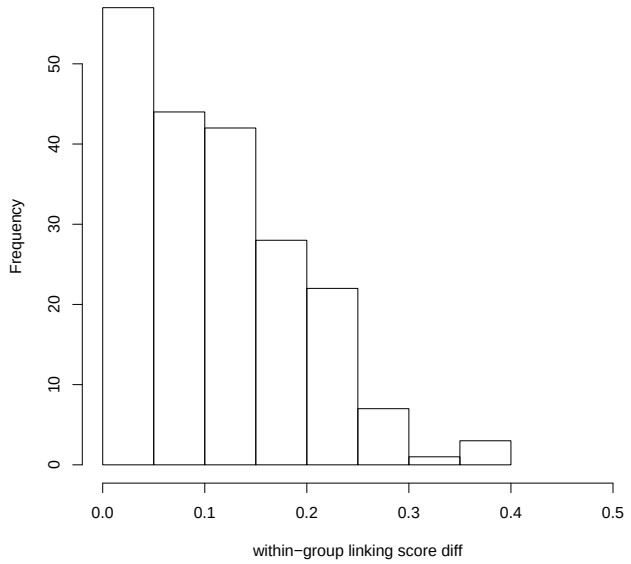

Figure D.1: Distribution of between-group linking scores. For each indexed person in the opgaafrollen that was linked to more than one person from the genealogies, the difference of the mean random forest classification score for each geneaology-person was calculated. The maximum possible difference is 0.5