Reconstructing historical populations from genealogical data: an overview of methods used for aggregating data from GEDCOM files

Corry Gellatly
Department of History and Art History
Social and Economic History
Utrecht University, Drift 6, 3512 BS
Utrecht, The Netherlands
c.gellatly@uu.nl

Abstract

The GEDCOM file format is by far the most widely used means of exchanging genealogical data and extensive collections of these files are available online. There is a huge potential benefit for historians and other academics who are able to make use of the data contained in available GEDCOM files, as these effectively represent hundreds of thousands of hours of crowd-sourced work and a considerable source of knowledge about individual families. This paper details a number of methods that are being used to clean and aggregate such genealogical data; this includes a series of steps for screening out substantially flawed files, as well as for cleaning date and place information. A group-linking method is described for identifying duplicates / linkages within a genealogical database based on comparison of family structures. This is tested alongside conventional methods (i.e. comparison of name and birth date) and an estimation of the power of the differing methods is provided. It is proposed that use of the group-linking method provides advantages over conventional methods, because this provides a way of increasing the size and timespan of datasets that may be extracted from a genealogical database with confidence that they do not contain duplicates. The method will be further improved by incorporating probabilistic record linkage techniques, which take into account the frequencies of values in the linkage arrays.

1 Credits

This work is part of the research programme 276-53-008, which is financed by the Netherlands Organisation for Scientific Research (NWO).

2 Introduction

GEDCOM files are widespread as a means for storing and exchanging genealogical data. The reasons for this are clear:

- The format provides a systematic and standardised way of structuring information about individuals, their families and life events.
- Exchanging of family trees is a way that people may use to identify ancestral connections and expand their own family trees, so there is a demand for a common exchangeable file type.
- The files are in plain text format, which allows for easy development of software applications that can read or export to the files - hundreds of such applications have been developed.
- The format is flexible in terms of what can be added to a file, allowing users and software to easily make use of the format, even without having to conform to the correct standard specifications[1] .

There are a number of criticisms of the GEDCOM format, in particular:

- There is a list of tags that can be used to describe events, but there is no allowance for additional tags, so unusual events typically have to be added in the notes. In some instances, people (or software) will use non-standard tags, which results in confusion or

[1] These are defined by the GEDCOM Standard Release, which is currently at version 5.5 (Church of Jesus Christ of Latter-day Saints, 1996) with an XML step-change version in development.

data loss when the file is read by a different person or software application. In these cases, the flexibility of the format is arguably problematic, even though the ability to add non-standard tags may have been useful at some point.

- There are no constraints over the data that may be entered under each tag. For example, it is possible to enter a date of birth from the future, or to enter e.g. '< 1900' or 'Born between 1900 and 1920' in a date field, which may not be understood by software. It is often the case that address identifiers are entered incorrectly, so that confusion may arise about geographical status, e.g. 'Washington' is both a city and a state in the US.
- There are technical constraints with the format, e.g. multiple people cannot be linked to the same record, so a single event, source citation or note may have to be replicated within each file. In technical terms, the data format is not normalised, leading to excessive replication of content and increased risk of data corruption.

The above problems with the GEDCOM format have undoubtedly led to files of poor quality becoming available online. However, there is an abundance of online genealogical data, much of which is well researched and involves knowledge that people naturally have about their own families, and as such incorporates indirect techniques of obtaining family history data (United Nations, 1983). There is an opportunity for those with an interest in reconstructing historical populations or following life history variables through generations to utilise this data, with the proviso that it must be assessed for probability of correctness.

In this paper, I detail the genealogical data collation methods that we are developing as part of a Dutch NWO funded project - *Nature or nurture? A search for the institutional and biological determinants of life expectancy in Europe during the early modern period*. In this project, we are comparing historical changes in life expectancy between south and north-west Europe, alongside institutional arrangements for care of the elderly, development of the European Marriage Pattern and changes in reproduction and fertility.

The methods extend those initially developed to study inter-generational patterns of human sex ratio variation (Gellatly, 2009). The starting point is a database of >900 primarily north American and western European genealogies, which is being expanded in terms of number of genealogies and geographical coverage through collaboration with the Dutch Central Bureau for Genealogy, engagement with amateur genealogists and investigation of further online repositories.

3 Screening, aggregating and cleaning

There are some large GEDCOM files in existence that contain many thousands of individuals, but most contain far fewer. If we are seeking to understand historical trends and population level phenomena, then the potential research questions that can be addressed with single GEDCOM files are limited. Hence, we aggregate individual files together by importing them into a single database, where we also screen out poor quality files, carry out data cleaning and identify linkages between separate genealogies.

3.1 Screening

Once the target GEDCOM files are downloaded, they are screened for the following errors that are assumed to indicate that they contain poor quality research or are incorrectly constructed:

- If the file has a low mean number of offspring per family due to inclusion only of the author's direct lineage and exclusion of their ancestors' siblings. This type of bias in genealogical research excludes much of the familial information that we require for reconstructing extended kin networks, which is important for understanding familial aspects of life history – e.g. availability of elderly care (Post *et al.* 1997). If the mean number of offspring per family is 1, then a file clearly contains a lineage. If it is close to 1, then further inspection may reveal branches of the tree where consecutive generations of individuals have one parent and no siblings.
- If there are individuals older than ≈ 110 or younger than 0, this is a potential cause to exclude files; except where there are clear typo errors. For example, an individual with parents born in the 1930's and siblings born in 1961, 1962 and 1963 was likely to have been born in 1965 and not 965. In such clear-cut cases, the typo may be ignored and the family tree included in the initial database. Lists of potential typo errors are later generated for the whole database, which allows all potential errors to be worked through sys-

tematically and typos either corrected or problematic files deleted.

- If the time between dates of birth precludes the possibility of a stated relationship between individuals, files are excluded. If an individual becomes a parent younger than 11 years old, or if a woman is older than 55 at the birth of a child or a man older than 80, then the file may need to be excluded unless the errors are clear typos.
- If the spacing between births is less than one year for any mother in the file, then this should be looked at as a cause to exclude the file or correct typos.
- A relationship to an individual not listed in the family tree is a reason to exclude a file; for example, a family may contain the ID of an offspring or parent, though the actual record of that person does not exist within the file. The ID should not have been created.
- In cases where individuals have more than one mother or father, it may be an error; however, it may also indicate adoption because biological and adoptive parents may both be linked via the child <CHIL> tag, whilst the adoption <ADOP> tag is an event tag that tends to be used to indicate the date adoption occurred rather than the relationship itself. This is one of the more taxing problems with the GEDCOM format. The absence of an adoption tag on multiple parentage should flag up the file for exclusion.
- Incestuous parentage is also a challenging problem, because there is no tag for it, whilst incorrect linking of individuals to offspring can happen. An investigation of other events, e.g. marriages and deaths will often allow incest to be ruled out.
- If the stated number of offspring in a family does not match the actual number, the file may be excluded. It can occur, for example, if there is a duplicate entry for a child in the family record.
- If individuals are listed as offspring of one sex, but occur as parents of another sex, the file may be excluded on the basis that this is a fairly major error on the part of the researcher. However, as with most errors, the context is important and some judgement is used. Is this the only error in a large, otherwise good, genealogy? Can the error be resolved with confidence (e.g. by cross referencing with another information source)?
- Another reason for excluding GEDCOM files is that people may trace their ancestry back so far that it is not credible. If an author traces their ancestry to before the Early Middle Ages (500-1000 AD) then it may be presumed that their methods were not particularly rigorous.

3.2 Aggregating

Having screened each GEDCOM file for the errors described above, all individuals with that file are ascribed unique keys that identify within the first four digits what GEDCOM file they belong to and in subsequent digits which families they belong to, before all information is aggregated by loading it into a common database. In this way, the meta-data from each file, e.g. author, download date, etc., is encoded with every data point via the unique keys. From a technical point of view, this eliminates the need for junction tables in the database and reduces the complexity of the queries used for extraction and deletion of data. The keys from the original GEDCOM files cannot be used because there would be duplication of these when data from the files is aggregated in a single database.

Any number of datasets may be selected from the database using SQL queries to select for specific variables or parameters. In any extracted dataset, all information can be linked back to the source GEDCOM file or selected on the basis of its meta-data.

3.3 Cleaning

As mentioned, the process of correcting typos and other minor errors in the database is carried out once the files have been aggregated, because SQL queries can be used to build lists, for example, of all women becoming mothers under the age of 13, or families with less than 18 months between sibling births. These lists can then be worked through systematically to check for typographical errors in the entry of dates. If a typo error cannot be identified and e.g. birth spacing or parental age is highly improbable, then the file is deleted.

These potential typo date errors were not uncommon, with an average of 1-2 in each screened file, whilst actual typo errors constituted the majority of these. An actual typo error was determined by weighing up several factors:

First, the existence of other event dates, such as marriage. If date of birth of a mother is 1834, birth of first child is 1845 and marriage date is 1854, then there is a good chance that the child's date of birth was actually 1854.

Second, dates of birth for parents, siblings, spouse and offspring. If, for example, younger siblings are born in 1865 and 1868, whilst parents were born in 1835 and 1840, then there is a good chance that the individual was born in 1861 and not 1816.

Third, an internet search for the individual. In many cases, the individual may be recorded in a separate source, such as another genealogy, census or gravestone record.

In the screened genealogies, dates were possibly the 'cleanest' fields, because most genealogy software enforces some control over date format, although some do not check for impossible dates, e.g. 31 June or 29 February in a non-leap year. There was no effort made to clean dates, because all queries incorporating dates used the MySQL date functions, which ignore impossible or incorrectly formatted dates.

The geographical information required the most effort, in terms of data cleaning. In GEDCOM files, place names are entered in a comma separated hierarchy under the PLAC tag, in an order which roughly corresponds to:

- Town/Village
- District/Region
- State/Province
- Country

However, although the correct order is usually followed (i.e. smaller to larger place), churches, graveyards, etc. are often included and higher level information is missing (i.e. country and/or region). For example, a record may look like this:

-
- Episcopal church
- Billings
- Montana

when it should look like this:

- Episcopal church
- Billings
- Montana
- USA

or perhaps this:

- Billings (Episcopal church)
-
- Montana
- USA

The process of cleaning the place information was semi-automated, but relied heavily on human input. SQL queries were used to automatically identify country names and US states anywhere in the place hierarchy. Lists of places were then manually checked to see that country names had been correctly applied to these places and to attribute country names where one was not identified automatically. For the European data, a similarly semi-automated method was used to attribute NUTS (Nomenclature of Territorial Units for Statistics) codes to regions.

In terms of name data, first names were more problematic than surnames, as they were more likely to contain nicknames, titles and punctuation. For the purpose of duplicate identification, all punctuation and spacing was removed and all text set to lower case for first names and surnames, though no manual cleaning was carried out.

3.4 Duplication and linkage

The issue of identifying duplicates between genealogies is important, because duplicates typically have to be removed for statistical analysis purposes (to avoid the problem of having non-independent variables within a dataset); moreover, identification of duplicates is the way that linkages between genealogies are identified. The issue of duplicates is dealt with at the point where a working dataset is extracted from the database.

The problem of data matching - also known as record linkage, has been tackled using a number of mathematical, computational and statistical techniques and has occupied the attention of scientists working in diverse fields. Newcombe *et al.* (1959) first described the principles of probabilistic matching, which recognises that records may be imprecise, inaccurate or 'fuzzy' and takes into account factors that may influence the likelihood of a match, such as the relative frequency of a particular surname within a sample. The modern academic literature describes how this principle is applied using various techniques in various types of data to match individuals based on their names, dates of birth and other information (e.g. George *et al.*, 1998; Hernández & Stolfo, 1998; Gu *et al.*, 2003; Christen, 2012).

Typographical and transcription errors are common to many types of records, but in the case of genealogical data, the problem of a person's name or an event date being entered differently in two separate files is compounded by the fact that names and dates may have been taken from different sources and recorded by researchers with different methods of transcription. For example, although there is a nickname tag <NICK> in the GEDCOM format, it is quite common to find the nickname in brackets alongside the first name, whilst either abbreviated or entire first names may be used. Also, a date may be entered as 'before 1876' or '>1654' or 'between 1701 and 1709', causing obvious difficulties for date matching.

Whilst various methods may be used to identify duplicates in selected datasets, it is important to determine to what extent the research question allows for refinement of the data, before a particular method is chosen. If, for example, it is acceptable to only select individuals with an accurate date of birth (DOB) then this makes the task of removing duplicates easier, because this is unique to more individuals than year of birth (YOB). But, crucially, a record linking method that requires date of birth accurate to the day, cannot determine if there are duplicates among those records that do not have this accuracy of detail. That data must, therefore, be excluded, resulting in a reduced size of dataset. The goal for extracting useful datasets with duplicates removed should be to obtain the broadest amount of data, whilst gaining the maximum accuracy in identification of duplicates.

The main focus of this paper is to detail an innovative group-linking method, which takes into account the relationship structures in genealogical data. The purpose of this is to increase the amount of data that can be extracted from a genealogical database for analysis, with reasonable confidence that duplicates or linkages between genealogies have been identified. Notably, there are similarities with the data mining approach developed by Ivie *et al*. (2007), which also makes use of family relationships to identify whether more than one pedigree belongs to a single individual.

Group-linking methods have been described for various other types of data, e.g. census records (Fu *et al.*, 2011) and publication records (Bhattacharya & Getoor, 2007). The idea is that you can better understand whether two separate records refer to the same entity if you understand the relationships with other entities. So, if we know that Joe Bloggs has co-authored a number of books with James Smith, this raises our confidence that Joe Bloggs co-wrote the book with the author record: 'J. Bloggs and J. Smith'.

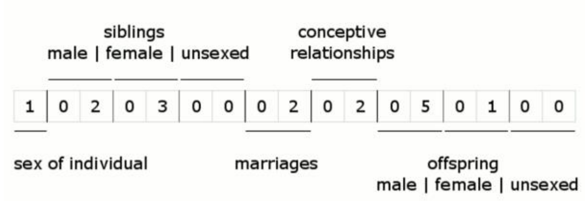

Figure 1: group-linking string

To identify matching family structures, numerical strings based on the sex of the individual, the number and sex of siblings, marriages, conceptive relationships and number and sex of offspring are used. Figure 1 illustrates how the family structure string is divided into sections. The focus individual is trivalent, being either male (1), female (0) or sex-unknown (-1). The sibling section of the string contains 6 digits and holds 3 values - brothers, sisters and sex-unknown (all starting at 00). The marriage section contains number of marriages, the conceptive relationships section contains number of relationships from which children were produced; finally, the offspring section includes all of an individual's offspring (irrespective of spouse) and the sex of those offspring.

3.5 Test of group-linking method

To test the method of group-linking based on family structure, a single screened GEDCOM file was imported into the database. In summary, this file contained:

- 2802 individuals
- 971 marriages/partnerships
- 681 conceptive relationships (where the same man and woman produced children)
- 1913 conceptions

Although we know that this file contains no duplicates, we see that that different methods for selecting data will, to varying degrees, incorrectly identify duplicates and will be applicable to a varying extent of the data. For example, in a typical GEDCOM file, a method based purely on

Data linking criteria	**Unique records (a)**		**Data coverage (b)**		**Linkage power (a*b)/100**
	n	**%**	**n**	**%**	
1. DOB (date of birth)	996	95.7	1040	37.1	35.5
2. YOB (year of birth)	46	3.3	1357	48.4	1.6
3. Surname	391	14	2790	99.5	13.9
3. Surname and DOB	1011	97.3	1039	37	36
4. Surname and YOB	1072	79	1356	48.3	38.2
5. Fam. struc.	419	14.9	2802	100	14.9
6. Fam. struc. and DOB	1026	98.6	1040	37.1	36.6
7. Fam. struc. and YOB	1237	91.1	1357	48.4	44.1
8. Fam. struc. and surname	1892	67.8	2790	99.5	67.5
9. Fam. struc., surname and YOB	1330	98	1356	48.3	47.4

Table 1: results of group-linking test, showing the number of unique records identified by the test and the extent of data that was accessible to the method (i.e. where the required fields did not contain null values)

YOB will cover more of the data than one based on DOB, but will be of little use because it has a much higher chance of incorrectly identifying duplicates. Incorporating the group-linking method, we wished to test which methods would (a) provide the highest level of data coverage, and (b) identify the highest proportion of unique records within that covered data. The combination of data coverage and uniqueness is taken into account to give an estimation of linkage power (Table 1).

It is seen from these tests of linkage methods that the most powerful is the combination of family structure and surname (8). This method has the highest data coverage in combination with unique records identified. However, within the data that it does cover, the percentage of unique records that it finds is much lower than methods 6,7 and 9. The combination of family structure, surname and YOB (9) finds a relatively high number of unique records, whilst also showing high data coverage.

The incorporation of the group-linking method is predicted to be effective in areas with high homogeneity of surnames, because it provides an additional means of distinguishing between individuals, whilst it also provides a way to extract older records where YOB is available, but DOB is less frequently found. It also provides a way to deal with the problem of nicknames, abbreviated first names, middle names and names such as 'baby boy SMITH', or 'unnamed SMITH', because the first name field does not have to be used.

It is interesting that family structure and YOB (7) has quite high linkage power, because this provides an alternative use of names for matching, which may be used to identify linkages, e.g. where there may be mixed use of married and maiden names between genealogies.

4 Discussion

In this paper, I have presented some of the methods being used to screen, correct and aggregate GEDCOM files for the purpose of historical and biological research. It is recognised that there are problems with the GEDCOM format, but we see that systematic screening of files for known types of error can eliminate much of the poor quality research that exists. It must also be recognised that there are biases in the way genealogical research is conducted, which results for example in over-representation of patrilineages (Zhao, 2001) or recording of apparent rather than accurate family linkages. However, there are also significant biases associated with censuses and other forms of historical data, (e.g. Steckel, 1991). As such, analyses with these types of data should always look to control for biases, such as under-recording of females or rounding of dates to the decade, and should seek to corroborate findings with other data sources where this is possible.

A method has been outlined for identifying duplicates within a genealogical database, which may be used to link separate genealogies. It makes use of the structure of an individual's family relationships as a group-linking pattern and combines this with surname, DOB or YOB. A significant advantage of this method is that it can attribute uniqueness for a high percentage of records with a high level of data coverage, without relying on first names (which are problematic in genealogical data) or an accurate DOB (which tends to exclude many individuals in older genealogical records).

The main difficulty associated with the method is that individuals in different genealogical files will appear to have a different family structure if their family is complete in one file and partially complete in the other, so will not be identified as a match. Moreover, whilst there is the possibility to extend the family structure string to include parental sibships, adoptive parents, half siblings, offspring of siblings, etc., which will raise the power of the method to identify unique records, this risks missing matches due to differing levels of completeness in separate genealogies.

In fact, the issue of how family structures are distributed throughout the dataset is important for optimising the method. For those variables where the percentage of unique records identified is low, but data coverage is high, e.g. family structure (5), YOB (2) and surname (3), the probability of finding incorrect duplicates is high. However, there is the possibility to refine the method using probabilistic record linkage, in which value specific frequencies for the variable in question (i.e. family structure) is calculated from the whole database, which is then used to weight the probability of a match between any pair (Fellegi & Sunter, 1969).

The methods put forward here are part of an attempt to recognise that there are many hundreds of thousands of hours of work tied up in the large number of GEDCOM files that have been produced over the last twenty years or so, and that we have the ability to build on this resource in order to reconstruct historical populations. In the short-term this will be useful for current research projects, in the long-term it is likely that researchers will benefit hugely from large openly accessible genealogical data repositories on the web, in which GEDCOM or other file formats are aggregated and meta-data is available that serves to flag up potential errors with individual records, so that users and search engines can obtain levels of confidence about the correctness and completeness of the genealogical data that they find.

References

Bhattacharya, I. & Getoor, L. 2007. Query-time entity resolution. *Journal of Artificial Inteligence Research.* **30**: 621–657.

Christen, P. 2012. *Data Matching*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Church of Jesus Christ of Latter-day Saints, The. 1996. *The GEDCOM Standard Release 5.5*. http://homepages.rootsweb.ancestry.com/~pmcbride/gedcom/55gctoc.htm.

Fellegi, I.P., Sunter, A.B. 1969. A theory for record linkage. *Journal of the American Statistical Association* **64** (328): 1183–1210.

Fu, Z., Christen, P. & Boot, M. 2011. A supervised learning and group linking method for historical census household linkage. In: *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121*, pp. 153–162. Australian Computer Society, Inc.

Gellatly, C. 2009. Trends in Population Sex Ratios May be Explained by Changes in the Frequencies of Polymorphic Alleles of a Sex Ratio Gene. *Evolutionary Biology.* **36**: 190–200.

George, R., Petry, F.E., Buckles, B.P. & Srikanth, R. 1998. Fuzzy database systems-challenges and opportunities of a new era. *International Journal of Intelligent Systems.* **11**: 649–659.

Gu, L., Baxter, R., Vickers, D. & Rainsford, C. 2003. *Record Linkage: Current Practice and Future Directions*. Technical Report 03/83, CSIRO Math. and Information Sciences.

Hernández, M.A. & Stolfo, S.J. 1998. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery.* **2**: 9–37.

Ivie, S., Pixton, B. & Giraud-Carrier, C. 2007. *Metric-Based Data Mining Model for Genealogical Record Linkage.* Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on.

Newcombe, H., Kennedy, J., Axford, S. & James, A. 1959. Automatic linkage of vital records. *Science*. **130** (3381): 954–959.

Post, W., van Poppel, F., van Imhoff, E. & Kruse, K. 1997. Reconstructing the Extended Kin-network in the Netherlands with Genealogical Data:

Methods, Problems, and Results. *Population Studies*. **51** (3): 263-278

Steckel, R.H. 1991. The Quality of Census Data for Historical Inquiry: A Research Agenda. *Social Science History*. **15**: 579–599.

United Nations. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. United Nations publication.

Zhao, Z. 2001. Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies.* **55**: 181–193. Routledge.