Linking Scottish Vital Event Records Using Family Groups

Özgür Akgün[1], Alan Dearle[1], Graham Kirby[1], Eilidh Garrett[2], Tom Dalton[1], Peter Christen[3], Chris Dibben[4], Lee Williamson[4]

[1] School of Computer Science, University of St Andrews
[2] Department of History, University of Essex
[3] Research School of Computer Science, The Australian National University
[4] School of Geosciences, University of Edinburgh

Abstract

The reconstitution of populations through linkage of historical records is a powerful approach to generate longitudinal historical microdata resources of interest to researchers in various fields. Here we consider automated linking of the vital events recorded in the civil registers of birth, death and marriage compiled in Scotland, to bring together the various records associated with the demographic events in the life course of each individual in the population. From the histories, the genealogical structure of the population can then be built up.

Rather than apply standard linkage techniques to link the individuals on the available certificates, we explore an alternative approach, inspired by the family reconstitution techniques adopted by historical demographers, in which the births of siblings are first linked to form family groups, after which intergenerational links between families can be established.

We report a small-scale evaluation of this approach, using two district-level data sets from Scotland in the late 19th century, for which sibling links have already been created by demographers. We show that quality measures of up to 83% can be achieved on these data sets (using F-Measure, a combination of precision and recall). In the future, we intend to compare the results with a standard linkage approach and to investigate how these various methods may be used in a project which aims to link the entire Scottish population from 1856-1973.

Keywords

Scottish vital event records, record linkage, linkage methods, group linkage, population reconstruction, Digitising Scotland

1. Introduction

It is now well over half a century since Louis Henry in France and Tony Wrigley in England published their ground-breaking works on family reconstitution, linking together the entries in the baptism, marriage and burial entries of the villages of Crulai and Colyton respectively (Gautier and Henry 1958; Wrigley 1966). Those conducting reconstitution studies, initially carried out using paper slips (family reconstitution forms) which could be sorted and searched manually, soon realised the advantages of applying computing power to the process of 'nominal record linkage' (Wrigley and Schofield 1973; Wrigley et al. 1997), and the field expanded rapidly. Studies using 'record linkage' diversified to include nominal records from other sources such as censuses, population registers and criminal records, and increasing computer power allowed greater numbers of individuals to be linked (see for example Bloothooft 2015; Ruggles 2002; Goeken et al. 2011).

Recent times have seen increased interest in research to automatically reconstruct (historical) populations from large databases (for an overview see for example Bloothooft et al. 2015). Various novel linkage techniques have been developed that either use sets of manually prepared ground truth data to train supervised machine learning techniques (see for example Antonie et al. 2013; Goeken et al. 2011) or exploit—as we do in this paper—the specific structures in populations such as households and families (Christen et al. 2017; Fu et al. 2014a; Fu et al. 2014b). Bailey et al. (2017) recently compared several automatic linkage methods on historical US records where manually prepared ground truth data is available, and found significant differences in both match rates and error rates between different methods. The main challenges addressed by recent work are linkage quality due to the often low data quality of historical population records that have been scanned and transcribed manually or using OCR, and the scalability of linkage algorithms due to the size of large population databases.

In the British Isles, parish records have several disadvantages; they can be inconsistent and incomplete in the information they record, they tend to record ceremonies—baptisms and burials rather than births and deaths—and they do not necessarily record the demographic events occurring to all members of society. The latter is an increasing problem for historians interested in the later nineteenth century, as many members of society broke away from the main churches, or simply gave up church-going altogether. The civil registers of vital events, introduced by the state in 1837 in England, 1855 in Scotland and 1864 in Ireland, were much more comprehensive in their coverage, and the information was collected in more standardised forms. For legislative and financial reasons these records have seldom been accessible to those wishing to undertake family reconstitution (the few exceptions include, for example, Paddock 1989; Kemmer 1989, 1997; Davies 1992, 1993; Garrett & Davies 2003; Davies & Garrett 2005; Blaikie et al. 2005; Reid et al. 2002, 2006).

Now, however, the Digitising Scotland[1] project (Dibben et al. 2012) has been given permission by National Records of Scotland (NRS) to transcribe the contents of the civil registers of births, marriages and deaths for Scotland covering 1856-1973[2]. This project aims to undertake a family reconstitution exercise which will encompass the entire population of Scotland over these 12 decades: some 14 million births, 11 million deaths and 4 million marriages. This unprecedented project requires the methods developed by demographic historians and computer scientists to be adapted and developed to cope with the numbers of individuals and amount of detail contained in the registers. Here we report preliminary work on developing and evaluating methods to support the national family reconstitution project.

This paper investigates the linkage of 'vital events' recorded in the certificates of births, marriages and deaths to create individual 'event histories' and family pedigrees (Reid et al. 2002). Such event histories potentially provide for a greater understanding of the different conditions under which populations lived, both through time and over space, than those facilitated through aggregated demographic rates such as births or deaths per thousand of the living population at a particular point in time. Modern day demographers prefer 'longitudinal' measures, such as the ones we strive to create, which follow individuals, couples or families over time—allowing researchers to assess who is 'at risk' of experiencing a particular type of event (Reid et al. 2002). Highly linked civil registration data sets also offer the opportunity to explore inter-generation questions of inheritance and change (eg social mobility). High quality family structure or pedigree data allow the separate influences of genetic, epigenetic or environments on important social and biological processes to be explored.

Civil registers very often contain many more details than were usually recorded by clergymen filling out parish registers. In the period of study, the Scottish birth records collected by the General Register Office for Scotland (now NRS) include the following fields (plus some others omitted here for simplicity):

• register entry number, year, registration district number and suffix
• child's forename(s)[3] and surname
• child's sex
• date and place of birth
• mother's forename(s), surname and maiden surname
• father's forename(s) and surname
• parents' date and place of marriage
• father's occupation

Death records include the following fields:

• register entry number, year, registration district number and suffix

---

[1] https://digitisingscotland.ac.uk
[2] The Scottish civil registers are of very high quality, with many more fields than are typically available in historic sources.
[3] Forenames are also known as first or given names.

- deceased's forename(s) and surname
- deceased's sex
- date, place and cause of death
- either the deceased's date of birth or their age at death
- deceased's occupation
- deceased's marital status
- deceased's spouse's name and occupation
- deceased's mother's forename(s), surname and maiden surname
- deceased's father's forename(s) and surname
- whether deceased's father and mother were deceased

Marriage records include the following fields:

- register entry number, year, registration district number and suffix
- groom's forename(s) and surname
- bride's forename(s) and surname
- date and place of marriage
- religious denomination
- either the bride and groom's dates of birth or their ages at marriage
- bride and groom's addresses
- bride and groom's occupations
- bride and groom's previous marital status
- bride and groom's mothers' and fathers' forenames, surnames and maiden surnames
- bride and groom's fathers' occupations
- whether bride and groom's fathers and mothers were deceased

This wealth of detail can make linking records a more certain task, as, when the details are accurately recorded an individual appearing in different records can be more securely identified, but the many pieces of information also mean that there is a greater scope for variation or mis-recording from one record to another.

Every individual observed must have had a birth and a death, although these may not always be included in the available records: the event may not have been recorded for some reason, or it may have occurred outwith the period covered by the records, or the individual may have migrated into or out of Scotland. Some individuals will have married, but not all of them, and some will have married more than once. Some individuals will have had children. The births and deaths of these children, and the marriage and the death of

spouses are all 'events' relevant to an individual's 'time line' or 'life course'.
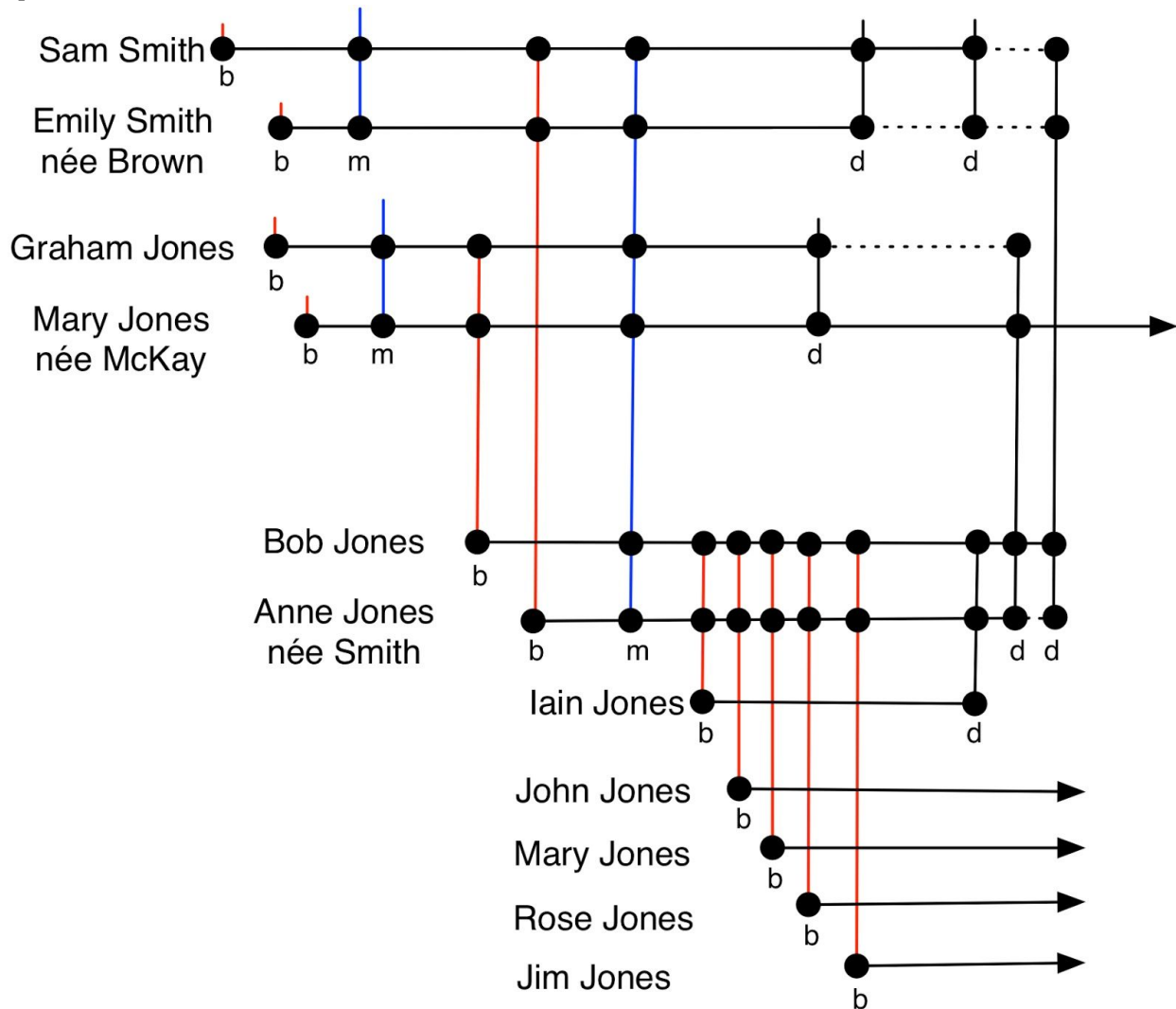

Figure 1: Relationship between vital event records (vertical lines) and individual life courses (horizontal lines).

The relationships between the events that comprise the life course of an individual and those of others may be usefully visualised as shown in Figure 1. Each horizontal line represents the life course of an individual. Thus we can see that Emily Brown was born, married Sam Smith, gave birth to Anne Smith (who herself married and later died), and then died. Anne Smith married Bob Jones and had five children, one of whom, Iain, died before his mother. The vertical lines represent events recorded in the vital event registers, and the dots on the lines represent the individuals that appear in each entry. Thus the birth record of Anne Jones (née Smith) makes reference to three individuals—her father Sam Smith, her mother Emily Smith (née Brown) and herself as the baby. The task of family

linkage may be usefully thought of as reconstructing this diagram from the individual vital event records.

Probabilistic automatic record linkage allows records to be linked through inexact information, taking account of errors[4] or inconsistencies in the original data or in the transcription, or ambiguity where there are alternative possible links. In the most common approach, called entity resolution (Fellegi and Sunter 1969; Christen 2012), records are linked on the basis that they contain information denoting the same underlying entity—in this domain, an individual. For example, the birth record of Mary Macdonald might be linked to the marriage record of Mairi McDonald, after a decision that both records denote the same person, despite the variation in spelling of the names. The way in which the decision is taken will affect whether or not a link is made, and therefore the outcome of any analyses conducted on the linked data.

The Scottish records form a rich data set, in which records could be linked according to many different criteria. We distinguish *individual linkage (*or *entity linkage)*, in which a single individual is identified as appearing in two records, and *family linkage*, in which relationships between the individuals appearing on records are established.

Examples of individual linkage include linking a woman's birth record to her marriage records, or linking a man's birth record to the marriage records of his daughters. The amount of relevant information available on which to make linking decisions varies with the type of link. In these examples there is more common information available in the first case (names of the woman and her parents, year of woman's birth, occupation of woman's father) than the second (man's name only). For this data set there are 64 possible individual linkage types of this nature (not enumerated here for brevity). How these these links may be exploited is the subject of ongoing work.

Examples of family linkage include linking birth records of full siblings (where the records are linked through common parents), and linking the birth records of spouses (where the records are linked via information on their marriage record).

Most applications of (probabilistic) record linkage focus on individual linkage. In contrast, our approach follows the 'family reconstitution' process undertaken by historical demographers. In this paper we focus on automating the specific process of linking together family groups of full siblings, as a component within an overall full linkage process. Future work will investigate which individual linkages are most appropriate, and how to combine them with family linkage.

We evaluate our family linkage approach using two data sets prepared by researchers pursuing previous projects (Reid et al. 2002, 2006), which act as pseudo subsets of the Digitising Scotland data set. One data set contains records of vital events registered on the Isle of Skye, a rural district, while the other contains records from Kilmarnock, an industrial

---

[4] Preliminary evaluation of the transcription quality of the Scottish records, after processing of around 13% of the records, indicates a per-character transcription accuracy rate of 99.5% based on a quality assurance sample of up to 3% of the records. The original records may also contain other errors of various types.

town. Both cover the period 1861-1901. These are different types of communities, with different family structures and name distributions (Reid et al. 2002).

Reid and her colleagues reconstructed their study populations using census records alongside the information included in the civil registers. They therefore had additional information available; this acted as a 'check' for their links, but in some instances could also 'point' them to a link which would not have been made on the basis of the registers alone. They were also concerned with particular aspects of demography, fertility and infant mortality, so their links focused on family groups, rather than undertaking a full reconstitution. They standardised forenames and surnames and identified links using database queries, following the reconstitution 'rules' laid out by Henry and Wrigley. The links made by machine were then verified by clerical inspection; a 'semi-automated' approach. Here we evaluate the extent to which our automated family linkage identified the same sibling links as those identified by Reid et al. (2002).

The family group formation step itself involves two stages. The first stage identifies small groups with parents of high similarity. This uses the traditional field-based similarity calculation between individual records. However, as has recently been shown (Christen, 2016; Fu, 2014a), only performing linkage on individuals may not yield high linkage quality. Because the historic data can be imprecise (names may be rendered in a variety of ways for example), the first stage, which uses a high similarity threshold, may produce family groups that are too small. The second stage considers various combinations of the small groups for possible merging into larger family groups. This merging stage appears to be a promising contribution to the field of population reconstruction, as it is applicable on all data sets that contain information which allows individuals to be grouped into entities such as families or households.

The structure of the paper is as follows. Section 2 describes the data sets used in the evaluation in more detail; while Section 3 lays out the importance of family grouping when aiming for full linkage. Section 4 then describes the method used to identify family groups and Section 5 provides an experimental evaluation. Section 6 outlines some future research directions, and Section 7 concludes the paper.

2. Data Sets

2.1 Isle of Skye Data Set

This data set, prepared by Alice Reid and her colleagues, contains around 17,600 birth records, 12,300 death records, and 2,700 marriage records. In 1861 the population of Skye was approximately 19,600 but by 1901, as a consequence of out-migration, it had shrunk to around 14,600.

Because of their research interests, Reid et al. identified 4,300 family groups of siblings, with a mean family size of 3.9 siblings and a maximum family size of 16. They also made 2,900 links from a birth record to the child's death record. The linkage was performed by

comparing names, marriage dates and places, between birth, marriage and death records, using spreadsheet and database tools. Cross-checking was performed against census records, and the consistency of intervals between births and marriages within individual families was checked.

The name pool in this data set is relatively restricted, however the difficulties that this might pose for linkage are largely offset by the fact that the Scottish civil registers not only record the 'surname and forename of father', but also the 'forename and maiden surname of mother' on each entry, giving additional information on which to link. Furthermore the 'place and date of parents' marriage' is included on each birth certificate, which strengthens links to the certificate recording that marriage. These additional items of information, seldom found in parish registers and not always present in the civil registers of other countries, serve to reduce ambiguity when carrying out record linkage.

Figure 2 shows the distribution of sizes of the identified family groups. It should be noted that these do not correspond directly to final family sizes. For example, a large family in which the first child was born shortly before the end of the observation period will be recorded with size one, since only that child's birth record is present in the data set.
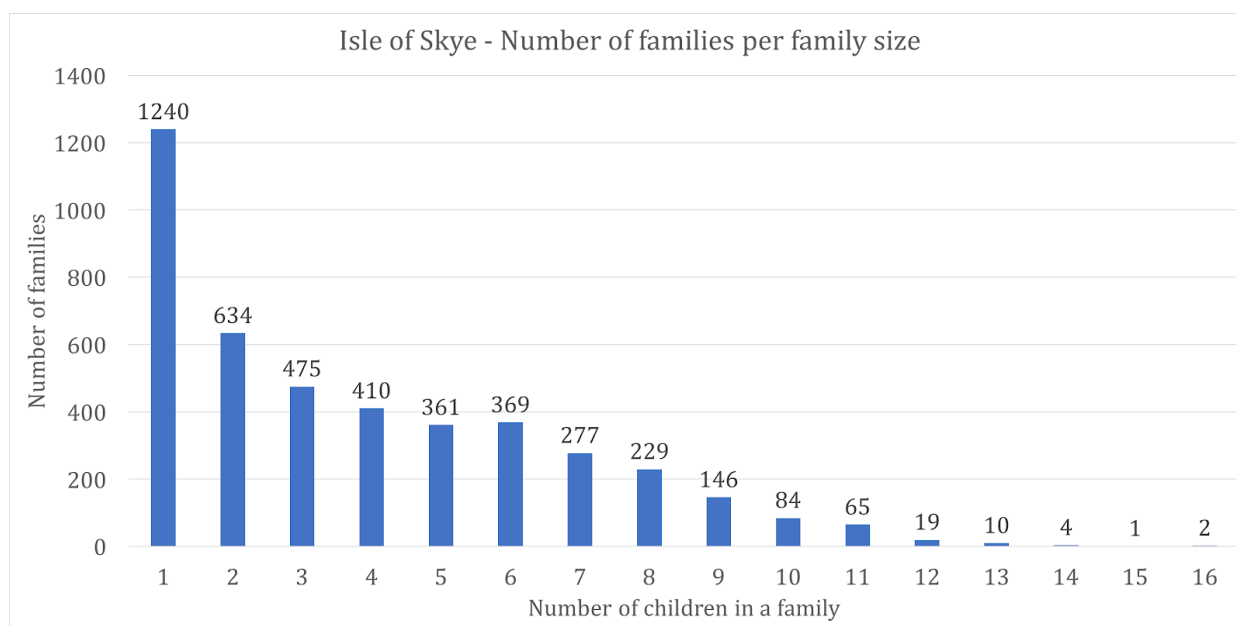

Figure 2: The distribution of observed family sizes in the Isle of Skye data set.

2.2 Kilmarnock Data Set

This data set contains around 38,400 birth records, 23,700 death records, and 8,700 marriage records. Again Reid and her colleagues identified 13,100 family groups of siblings, with a mean family size of 2.8 siblings and a maximum family size of 16. They identified 8,300 links from a birth record to the child's death record. The family groups from Kilmarnock form another 'gold standard', representing an urban population with much

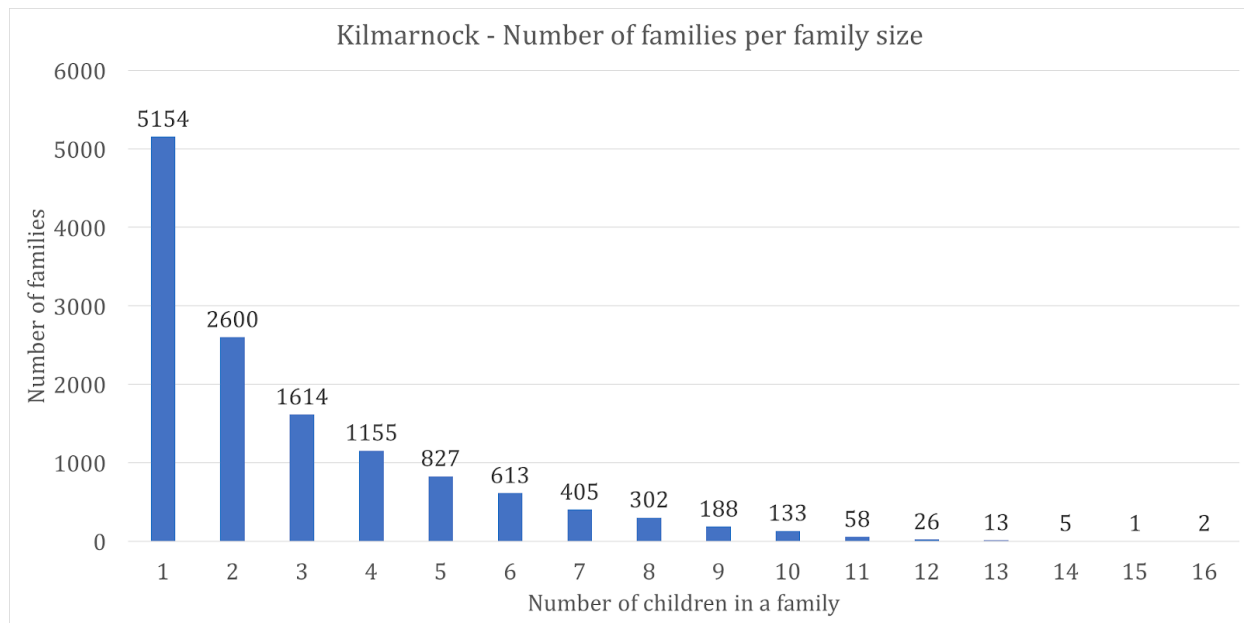population turnover, but a much more varied name pool. The same caveats apply concerning the links made.

Figure 3: The distribution of observed family sizes in the Kilmarnock data set.

3. Full Linkage and the Importance of Family Groups

As shown in Figure 1, links may be established between the three types of Scottish vital event records (births, deaths, and marriages) using the information about the different 'roles' that appear on them. Most birth records contain information about three roles: the baby and the two parents. A death record contains information about four[5] roles: the deceased, their spouse (if any), and the parents of the deceased. Marriage records are the richest of the vital event records, containing information about six roles: the bride, the groom, and the two sets of parents. Recent work explores the complex links that can be made using rich vital event records (Christen 2016), and we plan in future work to further explore the applicability of such advanced linkage techniques.

Table 1 shows a number of roles that an individual may take. For example, a female individual can appear as a baby on a birth record when she is born, the mother on a birth certificate when she bears a child, a bride on a marriage record when she gets married, and so on. The final column gives the number of times that a given individual can occur in a particular role. People are born once and they die once, but all other events can occur multiple times. Each role held by an individual corresponds to an intersection between a horizontal life course line and a vertical record line in Figure 1. In the case of Bob Jones, for

---

[5] In some cases multiple spouses may be listed on a death record.

example, his name is present on his own birth record, his marriage record (to Ann Smith), the birth records of his five children, and on his death record.

Record	Male roles	Female roles	Possible number of occurrences
birth	baby	baby	one
birth	father	mother	multiple
marriage	groom	bride	multiple
marriage	father of bride or groom	mother of bride or groom	multiple
death	deceased	deceased	one
death	husband of deceased	wife of deceased	multiple
death	father of deceased	mother of deceased	multiple

Table 1: Various roles for a given individual

In the light of Table 1, we define *full linkage* in the following way. Given a record that contains an individual in one of the roles from Table 1, a fully linked data set must be able to provide all the other records where this individual appears. Thus given the death certificate of Bob Jones from Figure 1, a fully linked data set would support retrieval of his full individual history: his birth and wedding record, the birth records of his children, and the death record of his wife.

To address some of the possible research questions, full linkage may not be necessary. For example, if a researcher wants to study the changes in family size over time in a certain population, only a statistical summary of those family sizes is needed. In this context, there are several ways of performing "minimal linkage". Minimal linkage establishes complete genealogical relationships for the population, but does not necessarily contain all kinds of links. Linking the parents on birth records to their own birth records is sufficient to generate complete genealogical relationships, assuming there are no missing records and perfectly accurate linkage. Similarly, linking the parents on death records to their own death records would also generate complete genealogical relationships. Depending on data quality (missing records, frequency of errors, etc) one of these may be more accurate than the other. One can also perform multiple minimal linkages to independently establish the relationships using different sources of data, and consider the combination of the two. Depending on the application domain, a union or an intersection of the two may be taken: the former potentially increasing the number of correctly linked certificates (i.e. recall), and the latter potentially increasing the accuracy of the linkage (i.e. precision).

Identifying siblings to form family groups is the focus of this paper. Identifying the sibling relationships does not immediately provide minimal linkage, but it can be an important stepping stone to record linkage. The sibling relationship can be seen as a backbone of linkage: once the sibling relationship is established, the remaining step is to make

intergenerational links to achieve minimal linkage, for example by linking the individuals in a sibling group to their parents.

4. Identifying Family Groups

In order to establish family groups, we aim to find full siblings from the vital event records, i.e. groups of individuals with the same parents. Sibling linkage could be performed across a number of different record types, considering various roles on those records. Here we focus on the six fields shown in Table 2, occurring on the birth, death and marriage records. The names of the parents are common to all records. Birth and marriage records also contain the place and date of the parents' marriage.

Birth	Marriage	Death
mother's forenames	bride's forenames	mother's forenames
mother's maiden name	bride's maiden name	mother's maiden surname
father's forenames	groom's forenames	father's forenames
father's surname	groom's surname	father's surname
parents' date of marriage	date of marriage	
parents' place of marriage	place of marriage	

Table 2: Key fields for identifying family groups

The technique that we employ, *sibling bundling*, involves matching records using a subset of the fields shown in Table 2, resulting in a set of records for each of the sibling groups that has been identified. For example, it is possible to compare birth and death records using the parents' names. Alternatively, birth records may be compared with birth records, using the parents' names and marriage information, as will be described below. This is a major component of our proposed population reconstruction process.

In order to describe the methodology, we initially focus on sibling bundling using only birth records, termed *birth-birth linkage*. In Section 4.3, we describe the generalised algorithm using other types of records. Bundling siblings together allows us to create the family structure that can be used as a backbone for later stages of linkage.

The sibling bundling algorithm has two stages as described in detail in the following two sub-sections. The first, the *family forming* stage, looks at the individual (baby) identified by each birth certificate, and attempts to bundle them together with one or more of their siblings with similar parental information. In the second stage, *family merging*, the algorithm considers whether to merge certain family groups generated by the first stage to form larger families.

4.1. Family Forming Stage

The family forming stage attempts to find siblings with highly similar parental information. In order to do this using birth records, we require a *distance metric* over pairs of records.

This is a function that calculates, for a given record pair, the distance between them in terms of the difference in their field values. For identical records the distance would be zero.

Here we use the six fields from column 1 of Table 2, which may be thought of as defining an imputed marriage record for the parents of the baby on a birth record. The intuition is that if two such imputed marriage records are highly similar, there is a high probability that the birth records are from full siblings. Perhaps counter-intuitively, the names of the babies themselves are not considered in this step, since it is shared parents that lead to birth records being linked (family linkage rather than individual linkage).

There are many possible distance metrics over records, varying in the methods used to compare individual fields, the method used to combine the individual field distances, and the types of cleaning performed on the field values, if any (Christen 2012).

We illustrate using a simple metric, the Levenshtein edit distance (Levenshtein 1966), to compare individual fields, combined using an unweighted sum, without cleaning[6]. Levenshtein distance (LD) measures the number of character deletions, insertions and substitutions required to transform one string into another. Thus we define the Record Distance metric (RD) over two birth records *b1* and *b2* as the sum of the Levenshtein distances between the corresponding record fields:

$$RD(b1, b2) = \sum_{i=1}^{6} LD(b1[i], \; b2[i])$$

where *b1[i]* denotes the i'th field of record *b1*. As an example, consider the following birth records taken from the families shown in Figure 1:

Forename	John	Iain
Surname	Jones	Jones
Date of birth	14/3/71	22/5/73
Mother's forenames	Anne	Ann
Mother's maiden name	Smith	Smythe
Father's forenames	Bob	Bob
Father's surname	Jones	Jones
Parents' date of marriage	16/10/59	15/10/59
Parents' place of marriage	Dollar	Dollar

Table 3: Two example birth records

The distance between these records based on the six parental information fields is:

---

[6] In separate experiments we have explored using date-aware distance functions for comparing dates, and phoneticising names before comparison, without observing any significant improvements in linkage quality.

RD	=	LD("Anne", "Ann") + LD("Smith", "Smythe") + LD("Bob", "Bob") +
		LD("Jones", "Jones") + LD("16/10/59", "15/10/59") +
		LD("Dollar", "Dollar")
	=	1 + 2 + 0 + 0 + 1 + 0
	=	4

Thus in this case the total distance between the two records (for the purpose of sibling bundling) is four, meaning that a total of four character edits would be required to transform all six relevant fields of one record into those of the other. This might be interpreted as the records being highly similar.

To find all potential sibling groups requires comparison of parental information on all birth records. The simplest approach would be to compare each record with every other record, requiring (N x (N-1))/2 comparisons if there are N records to be linked. This is likely to be infeasible for the 14 million birth records of the Scottish population[7]. To reduce the number of comparisons required, most automated linkage approaches use *blocking* (Christen 2012), which partitions records into sets according to an equivalence function defined over a subset of the record attributes, and then only compares pairs of records within each block.

We are exploring an alternative approach to matching records using *similarity search*, with a supporting data structure, the M-tree (Ciaccia 1997). An M-tree stores a collection of records, using a given distance metric, and supports three basic query operations:

- find the nearest neighbour of a given record
- find the *n* nearest neighbours of a given record
- find all neighbours within a distance *d* of a given record

Each of these operations takes a query record as a parameter, and returns one or more of the records stored in the tree. The most significant feature of the M-tree and related similarity search approaches is that they provide much more efficient implementations of these queries than full pair-wise comparison, and are thus usable with much larger data sets, and are less prone to giving rise to false negatives than blocking approaches.

In the family forming stage of the algorithm we use an M-tree of birth records with the distance metric defined above (comparing the parents' information on the birth records). All the birth records are added to the M-tree. Next, we iterate over all birth records, and for each one we retrieve its nearest neighbour record from the M-tree. If the distance between these two records is less than some specified threshold, we link the two records as (potential) siblings by placing them into a family group. If either of the siblings is already in a family group, we do not create a new group, instead the other child in the pair is added to the existing group.

---

[7] There would be around 10^{14} record pairs to compare. If a pair could be compared in one microsecond, for example, the overall process would take around three years.

The setting of an appropriate threshold is a complex issue which we explore below. We expose this as a parameter to the algorithm, the **family_forming_threshold**, representing the maximum Levenshtein distance at which two records are linked as siblings.
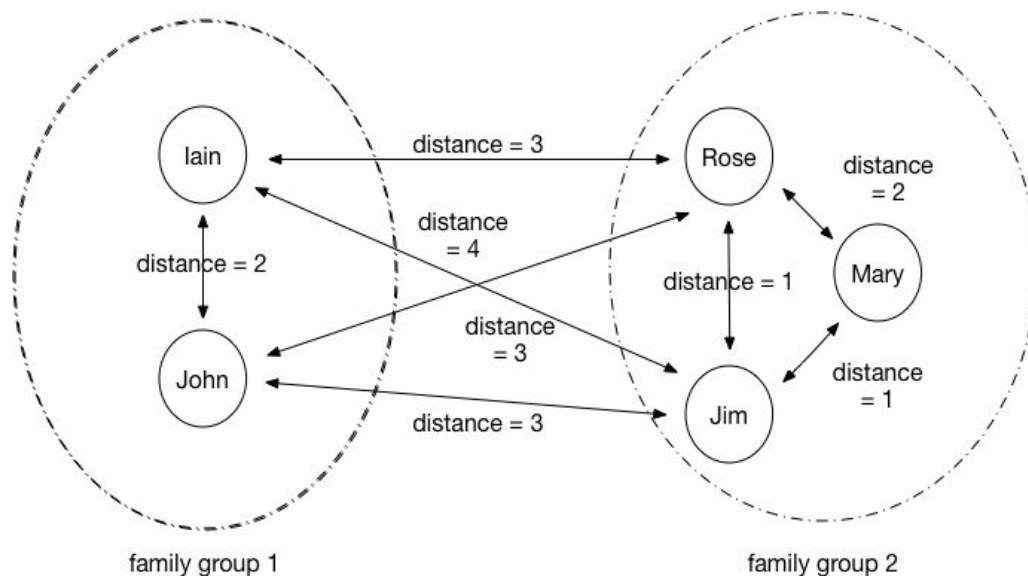

Figure 4: Family groups considered during family merging.

The output of the family formation stage is a set of mappings from birth records to those of their siblings. However, this does not necessarily link complete sibling groups. For example, Figure 4 shows two separate sibling groups. Each record has been linked with its nearest neighbour, but not with the members of the other group. The siblings in each of the two groups in the figure are deemed to have the same parents; it is possible that the two sets of parents are in fact the same couple and the five children are all siblings.

4.2. Family Merging Stage

The family merging stage merges family groups that are close to each other (as defined by the same parental distance metric). For example, in Figure 4 we might merge the family group of Iain and John with the family group of Rose, Mary and Jim.

The family forming stage, with an appropriate threshold, can identify accurate collections of siblings (high precision) with few false positive links. There may, however, be many false negatives (low recall), i.e. many siblings are missing because the information on their records has been entered slightly differently. The family merging stage improves the completeness of the families, thus raising recall.

The family merging process is similar to the family forming process but performs similarity searching over family groups rather than over births. Again, an M-tree is employed, this time storing family groups rather than births, with a distance metric defined over groups.

The first stage of the algorithm is to add all of the family groups formed in the first stage to the new M-tree. Next, for each family we find all the families within a configurable distance *family_merging_threshold*, using a range search. These families are merged into a new larger family group, so long as the combined size would not exceed the parameter *maximum_family_size*, used to constrain the size of merged family groups.

This process is illustrated in Figure 5. The family groups 1, 2, 3 and 4 will be merged into a new family group provided its size is not greater than *maximum_family_size*. Even if this is not the case (the combined family would be too large), individual groups could still be merged with each other later in the merging process.
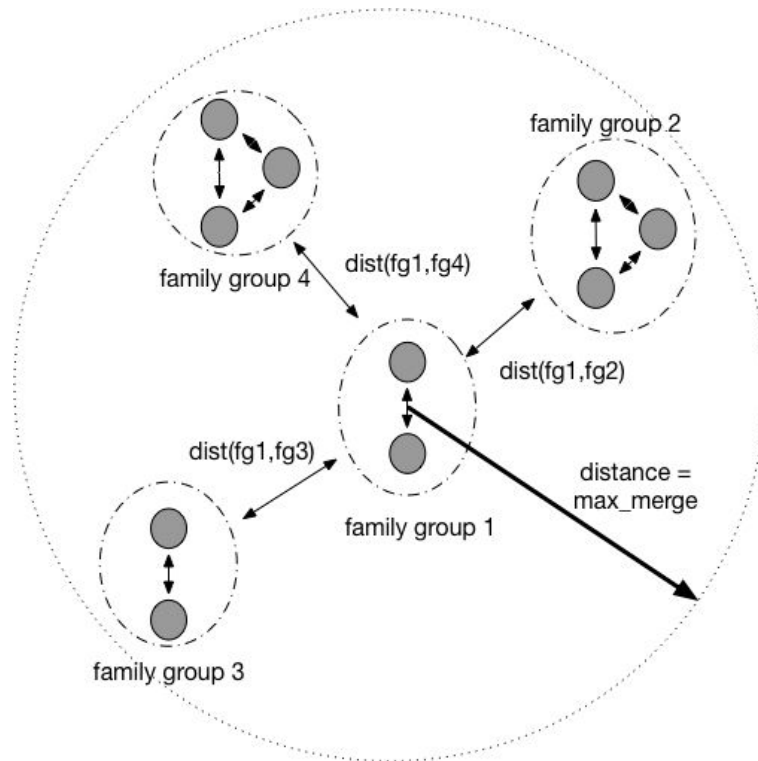

Figure 5: Finding candidate family groups to consider for merging.

There are several options for calculating the distance metric over family groups. For example, the minimum, mean or maximum distances between any record in one family group and any record in the other family group could be used.

We have experimented with these three distance metrics. All involve first calculating the pairwise distances between all pairs of records, one drawn from each family group. *Closest* selects the smallest value as the distance between the families, *Furthest* chooses the largest, and *Mean* takes the mean of all distance values. These correspond to *single link*, *average link* and *complete link* similarity calculations as used in hierarchical clustering (Han 2012). These alternatives are investigated in the experiments described below.

Section 5 examines the determination of appropriate values for the parameters *family_merging_threshold* and *maximum_family_size*.

4.3. Generalised Algorithm Description

In this section, we generalise the *birth-birth* linkage approach to use other types of vital event records. Parental information can be extracted from birth, death and marriage records, thus the linkage algorithm can be parameterised with the source of records, called ***family_hints***. These records are placed in the M-tree and searched to find matches, from which family groups are identified.

The M-tree is searched using parental information extracted from the hints, which are not necessarily of the same type as the records used to populate the tree. The parameter ***candidates*** specifies the set of records which we draw upon to search the M-tree. These candidates are formed together into family groups. The generalised family forming algorithm is shown in Figure 6.

```
Initialise empty M-tree of records
For each record (F) in family_hints:
     PF = Parental information of F
     Add F to M-tree using PF as key
For each record (S) in candidates:
     PS = Parental information of S
     N = Find closest neighbour of PS in M-tree
     If Distance(PS, N) <= family_forming_threshold:
          Place S and N in the same family group
```

Figure 6: Pseudo-code for the generalised family forming algorithm.

Figure 7 shows the algorithm used for merging families formed by the algorithm given in Figure 6.

```
Initialise empty M-tree of family groups
For each family group (G) in formed_families:
     Add G to M-tree
For each family group (G) in formed_families:
     N = Find all neighbour groups of G in M-tree within
         family_merging_threshold
     If merged family size <= maximum_family_size:
          Merge G and N into the same family group
```

Figure 7: Pseudo-code for the generalised family merging algorithm.

The intuition behind being able to select the ***family_hints*** and ***candidates*** is that by altering these parameters, it may be possible to link a greater number of siblings more accurately. For example, if marriage records were missing (due to lost records or unmarried parents), using deaths for the ***family_hints*** might yield higher quality results.

Conversely, if a birth record was missing, using deaths for the **candidates** would allow the identification of an extra child within the relevant family. In a production version of our algorithm, multiple sources of family groups would be used in conjunction, but in this paper we only compare the three options, births, marriages and deaths, separately.

5. Sibling Bundling Experiments

In this section we explore the effects of varying some of the following parameters to the linkage process:

- the distance metric used to compare fields (here fixed as Levenshtein distance)
- the fields used in comparing records (here fixed as the parental information fields)
- the manner in which inter-record distances are derived from individual field distances (here fixed as equally weighted sum)
- *family_forming_threshold*, the distance below which two records are considered to denote siblings (experimental parameter)
- *family_merging_threshold*, the distance below which two family groups are considered for merging (experimental parameter)
- *maximum_family_size*, the maximum family size (experimental parameter)
- *family_distance_method*, the manner in which inter-family distances are derived from individual record distances (experimental parameter)
- *candidates*, the type of record used to identify possible siblings (here fixed as birth records)
- *family_hints*, the type of record used to identify families (experimental parameter)

In our experiments, we sample the following candidate values for these parameters.

Parameter	Values
family_forming_threshold	2, 5, 8, 10 (distance)
family_merging_threshold	2, 5, 8, 10 (distance)
maximum_family_size	8, 20 (family size)
family_distance_method	Closest, Furthest, Mean
family_hints	Birth, Death, Marriage

Table 4: Candidate values for the algorithm parameters

As can be observed in Table 4, the experimental space contains 4x4x2x3x3 = 288 combinations of algorithm configurations. We run all combinations on the demographer-linked data sets from Skye and Kilmarnock. In the rest of this section, we explain the linkage quality measures we use, report our findings for each data set separately, and present a combined analysis.

5.1. Linkage Quality Measures

Precision and recall are common measures for assessing linkage quality (Rijsbergen 1979). Precision is an indicator of soundness: it is defined as the proportion of true-matches within the identified links. Recall is an indicator of completeness: it correlates with the proportion of identified links within the true-matches. If a linkage procedure always finds correct links and makes very few mistakes, its precision will be high. If it identifies a large proportion of the true links correctly, its recall will be high.

It is easy to design a linkage procedure that achieves either high precision or high recall in isolation. A conservative procedure that only identifies links with very high confidence will have high precision, but low recall. Similarly, a risky procedure which identifies almost all potential links as links will achieve high recall, but with low precision; many of the links will not be true-matches. To be useful, a linkage procedure must achieve both good precision and good recall.

The F-Measure combines precision and recall. Achieving a high F-Measure requires achieving relatively high values of precision and recall at the same time. We use the F-Measure in our evaluation, while noting that recent research identifies some problematic aspects with using the F-Measure to compare record linkage procedures at different similarity thresholds (Hand and Christen 2018).

Figure 8 gives the formulas for calculating precision (P), recall (R), and the F-Measure (F), derived from the numbers of true-positives (TP), false-positives (FP), and false-negatives (FN). True-positives are the true-matches (as identified by the demographers) correctly identified by the algorithm as links. False-positives are identified as links by the linkage procedure, but are in fact not true-matches. False-negatives are not identified as links, but are in fact true-matches.

Computationally, counting all the TPs and FPs is very fast. We iterate over each link and count those which are true-matches as TPs and those which are not true-matches as FPs. Counting all the FNs is also fast: we iterate over each true-match and count those which are not identified as links.

$$
\begin{aligned}
\text{Precision (P)} \quad &= \quad TP / (TP + FP) \\
\text{Recall (R)} \quad &= \quad TP / (TP + FN) \\
\text{F-Measure (F)} \quad &= \quad 2{*}P{*}R/(P+R)
\end{aligned}
$$

Figure 8: Linkage Quality Measures

5.2. Comparative Evaluation Nomenclature

Our linkage approach has six parameters (see Table 4). It is non-trivial to work out which parameter values give good overall results when the number of dimensions is large. We perform a dominance analysis to provide a better understanding of the effect of different

parameter values. The following relations between configurations of the algorithm are defined:

- We say a configuration C1 (with resulting F-Measure F1) is **preferable** to another configuration C2 (with resulting F-Measure F2) if F1 is greater or equal to F2.

- We define **pref**(P,A,B) to be the number of configurations in PA that are preferable to the corresponding configuration in PB (i.e. with P=B rather than P=A and the other parameters the same).

- We define **p_ratio**(P,A,B) = pref(P,A,B)/N, i.e. the proportion of configurations in PA that are at least as good as the corresponding configurations in PB.

- We say P=A **dominates** P=B if p_ratio(P,A,B) is 100%. Using this definition, if P=A dominates P=B, we know that we should always choose A over B for the parameter P.

- Finally, we say P=A is **optimal** if it dominates all other values for P.

In order to calculate **pref(P,A,B)** we construct a list PA containing N configurations, in each of which P takes the value A, and the other parameters take certain values. We construct another list PB by copying the configurations in PA, and setting P to the value B in each one. We evaluate the resulting F-Measure for each configuration in PA and PB.

5.3. Evaluation using the Isle of Skye Data Set

We evaluate all methods of performing sibling bundling on the Isle of Skye data set. Different algorithm configurations provide very different linkage quality (relative to the links made by the demographers): we observe F-Measure values as low as 17% and as high as 84%. This means that the selection of appropriate values for our algorithm parameters is crucial to achieve good performance. Figure 9 shows a plot of the distribution of F-Measure values, with the horizontal axis showing the different configurations and the vertical axis showing the F-Measure values.
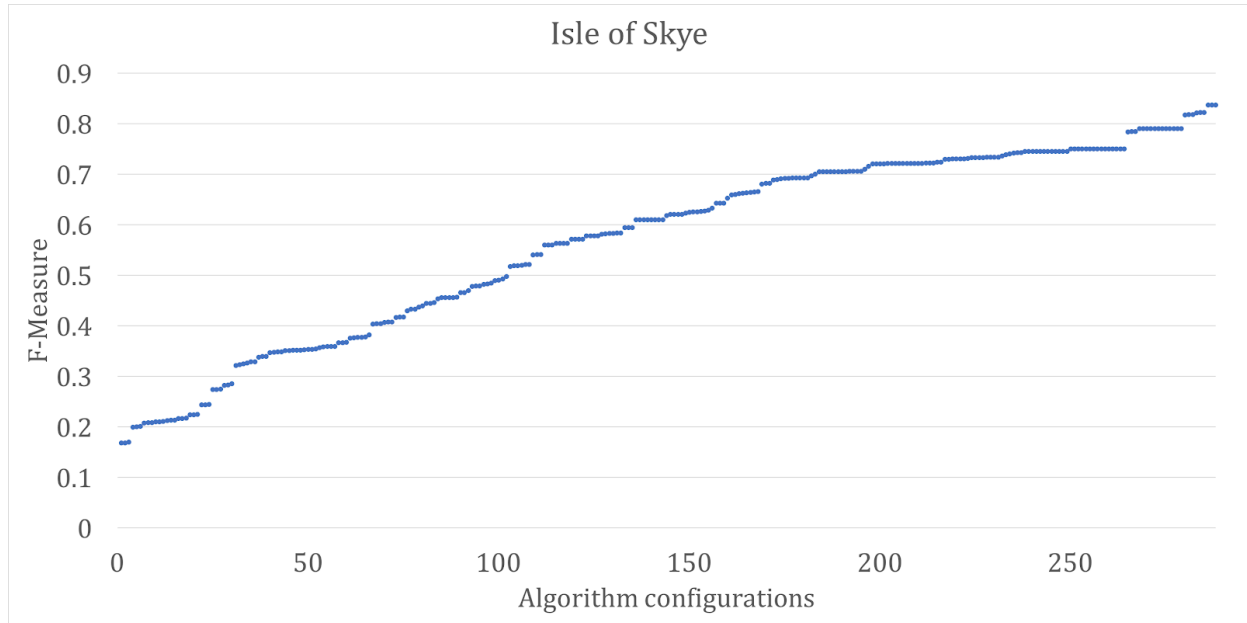
Figure 9: Sorted F-Measure values for the Isle of Skye data set,
for each of the 288 algorithm configurations derived from Table 4.

The `p_ratio` values for Isle of Skye are presented in Tables 5.1-5.5. It is important to note that `p_ratio(P,A,B)` + `p_ratio(P,B,A)` is not necessarily `100%`, indeed it is often greater than `100%`. This is because of the definition of what we consider `preferable`. If two settings, `P=A` and `P=B` provide the same F-Measure values, then `A` is preferable to `B` and also `B` is preferable to `A`. This is not helpful for individual comparisons, but it makes the dominance and optimality analysis possible. Using the definitions above, we make the following observations for the Isle of Skye data set.

- Setting the *family_forming_threshold* to 10 dominates all other values, hence it is optimal (threshold 8 dominates 2 and 5, but not 10).

- Setting the *family_merging_threshold* to small values is better: 2 dominates all the other values, hence it is optimal.

These two observations in combination indicate that being liberal when initially forming the families and being conservative when merging them is the best option for this data set.

- There is not a clear winner between 8 and 20 for the *maximum_family_size* parameter, but 20 is more promising. Since we know that the maximum family size is 16 in our ground truth, we did not initially plan to test larger values for this parameter. However, after observing that size 20 is generally better than size 8, we experimented with an even larger value for this parameter: size 30. We find that size 20 dominates size 30.

- For the *family_hints* parameter, using the death records is always better than using the marriage records. However this does not mean that using the death records is the

best option, indeed in 94% of the cases using the birth records is better. We can say that our results are not fully conclusive with respect to this parameter, but the most promising option is using the birth records.

- For the **family_distance_method** parameter, the *Furthest* option is the optimal. It is worth noting that the **p_ratio** values are very high for every pair. This can be interpreted to mean that the linkage quality does not depend heavily on the choice of this parameter.

	2	5	8	10
2	-	42%	42%	38%
5	96%	-	54%	42%
8	100%	100%	-	43%
10	100%	100%	100%	-

Table 5.1: The dominance table for **family_forming_threshold** (Isle of Skye)

	2	5	8	10
2	-	100%	100%	100%
5	17%	-	100%	100%
8	0%	0%	-	100%
10	0%	0%	0%	-

Table 5.2: The dominance table for **family_merging_threshold** (Isle of Skye)

	8	20
8	-	25%
20	75%	-

Table 5.3: The dominance table for **maximum_family_size** (Isle of Skye)

	Birth	Death	Marriage
Birth	-	94%	97%
Death	12%	-	100%
Marriage	9%	12%	-

Table 5.4: The dominance table for **family_hints** (Isle of Skye)

	Closest	Furthest	Mean
Closest	-	86%	100%
Furthest	100%	-	100%
Mean	100%	91%	-

Table 5.5: The dominance table for *family_distance_method* (Isle of Skye)

5.4. Evaluation using the Kilmarnock Data Set

Our second data set is from Kilmarnock. We run the same combinations of parameter-value assignments for the Kilmarnock data set and present the findings. The rationale for doing this is that if similar configurations of our algorithm perform similarly on these quite different data sets, it would build confidence in those particular parameter value assignments. The `p_ratio` values are presented in Table 6.
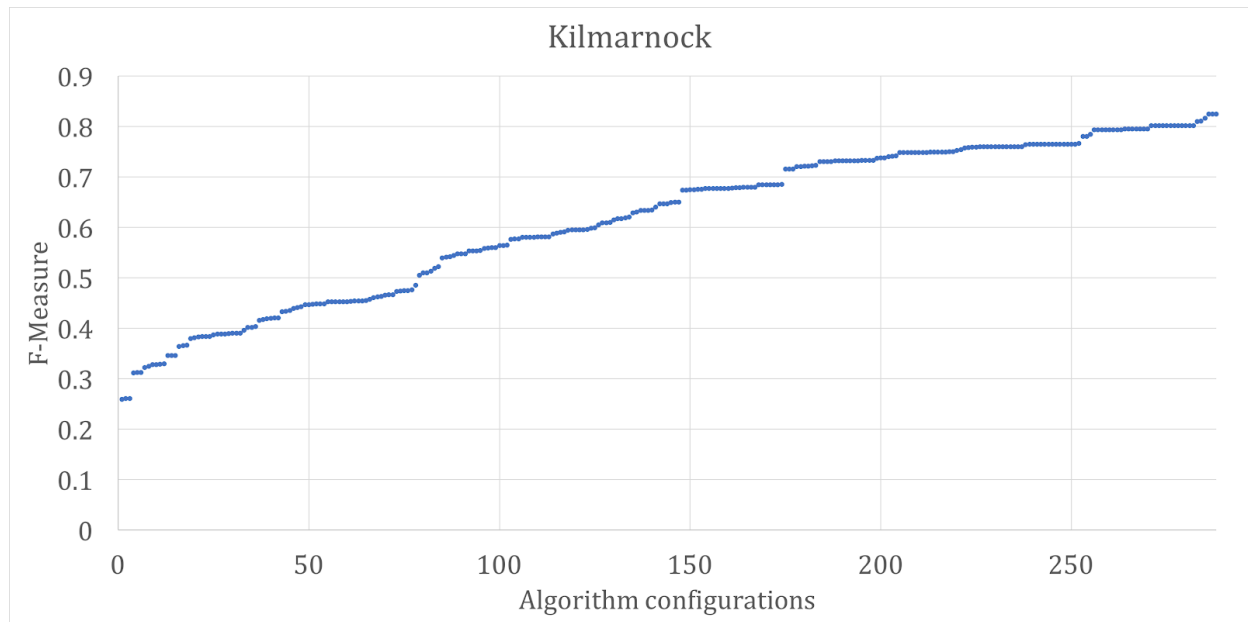

Figure 10: Sorted F-Measure values for the Kilmarnock data set,
for each of the 288 algorithm configurations derived from Table 4.

In the Kilmarnock data set, we have a similar spread of F-Measure values. The lowest is 26%, and the highest is 83%. The dominance analysis is as follows.

- The value of 10 for the *family_forming_threshold* is not optimal for Kilmarnock (in contrast to Isle of Skye). However, it dominates 2, and is preferable to 5 and 8 in the majority of cases.

- Similarly, the value of 2 for the *family_merging_threshold* is not optimal, but it is very close. Both 2 and 5 dominate the larger values, 8 and 10. Between 2 and 5, 2 is

preferable in 88% of the configurations. Using a smaller value for the *family_merging_threshold* gives better results.

- The findings with respect to the values of **maximum_family_size**, **family_hints**, and **family_distance_method** are exactly the same as those in the Isle of Skye data set.

Overall the experimental results between the two data sets are very close to each other.

	2	5	8	10
2	-	33%	42%	40%
5	100%	-	58%	46%
8	96%	92%	-	42%
10	100%	92%	99%	-

Table 6.1: The dominance table for **family_forming_threshold** (Kilmarnock)

	2	5	8	10
2	-	88%	100%	100%
5	86%	-	100%	100%
8	0%	0%	-	100%
10	0%	0%	0%	-

Table 6.2: The dominance table for **family_merging_threshold** (Kilmarnock)

	8	20
8	-	25%
20	83%	-

Table 6.3: The dominance table for **maximum_family_size** (Kilmarnock)

	Birth	Death	Marriage
Birth	-	94%	100%
Death	9%	-	100%
Marriage	0%	0%	-

Table 6.4: The dominance table for **family_hints** (Kilmarnock)

	Closest	Furthest	Mean

Closest	-	86%	97%
Furthest	100%	-	100%
Mean	100%	94%	-

Table 6.5: The dominance table for ***family_distance_method*** (Kilmarnock)

5.5. Combined Analysis

The behaviour of our algorithm with respect to linkage quality on the two data sets is quite consistent: similar configurations perform similarly across the two data sets. This is a desirable feature, since we intend to apply the same algorithm to unseen data sets in the future. It is important to note that this algorithm was not designed for a particular data set, but to be a part of a general purpose population reconstruction system. The algorithm is also completely unsupervised: it does not require any training with labelled data. Therefore it is readily applicable to unseen data sets.

In both data sets we see that ***family_forming_threshold=10*** and ***family_merging_threshold=2*** are very good options (they are optimal for Isle of Skye and close to optimal for Kilmarnock).

We see that 20 is a better option than 8 for ***maximum_family_size*** as well, but not decisively. Choosing the largest value from a range always raises the question of what might happen with larger values. This prompted further research into larger values for this parameter. We added 30 as a third option for this parameter and repeated our experiments. For both data sets, 20 dominates 30 for the ***maximum_family_size*** parameter. This means using 30 as the threshold for maximum family size would reduce linkage quality. We speculate that this is due to the true maximum family size being 16 (<20).

The same configuration yields the best quality on both data sets (*family_forming_threshold=10, family_merging_threshold=2, maximum_family_size=20, family_hints=Death, family_distance_method=Closest*). On the Kilmarnock data set this gives an F-Measure of 82%, corresponding to recall of 71% and precision of 99%. On the Skye data set the F-Measure is 84%, with recall of 78% and precision of 90%.

The raw data from our experiments is available for further analysis by the interested reader (Akgün 2018).

5.6 Comparison Against Traditional Record Linkage

We compare the linkage quality obtained by our algorithm against the linkage quality obtained by individually linked records, which is the more standard approach (Christen 2012, Fellegi and Sunter 1969).

For reference, we compare this algorithm to a standard approach in which for every pair of birth records we calculate their Levenshtein distance (Christen, 2012). If this distance is

below the linkage threshold, we designate the pair as a link. All pairs that are not linked are considered to be non-links.

We conduct experiments using birth records from both of our benchmark data sets: Isle of Skye and Kilmarnock. We use the same distance metric (Levenshtein on parental information fields) and use the same distance thresholds as described in the previous sections. Overall the linkage quality using individually linked records is worse than our proposed method which utilises familial information.
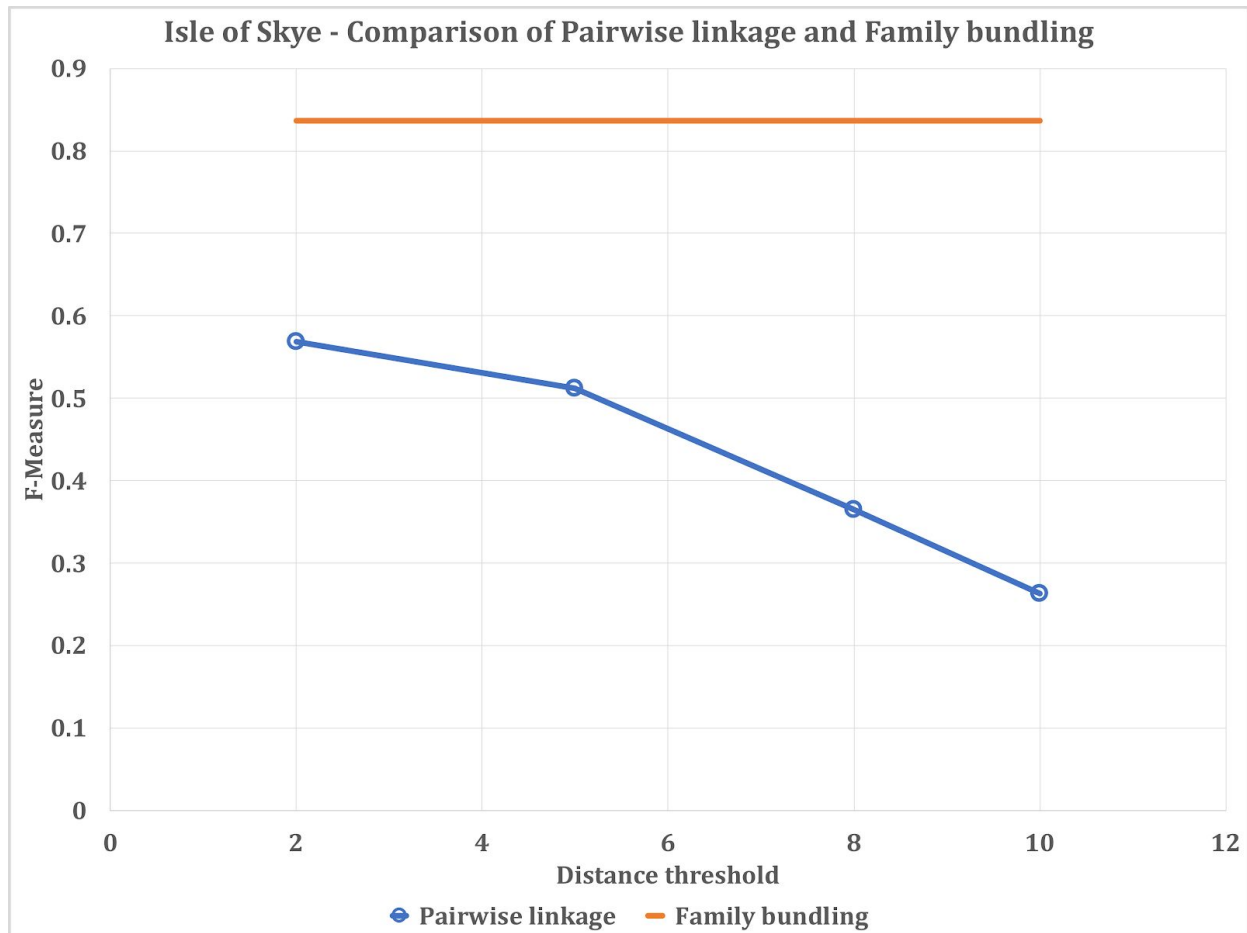

Figure 11: Comparison of the F-Measure values obtained by performing pairwise linkage and family bundling on the Isle of Skye data set. We present the best family bundling configuration as a target line in comparison to several settings of distance threshold for the pairwise linkage algorithm.

In the Isle of Skye data set, the highest F-Measure we observe with family bundling is 84% (see Section 5.3). In comparison, the best F-Measure value we obtain with individually linked records is 57%, with a distance threshold of 2. The other distance thresholds 5, 8, and 10 give 51%, 36%, and 26% F-Measure values, respectively. This means the best

setting of the traditional record linkage algorithm produces linkage results of significantly lower quality in comparison to our proposed method.
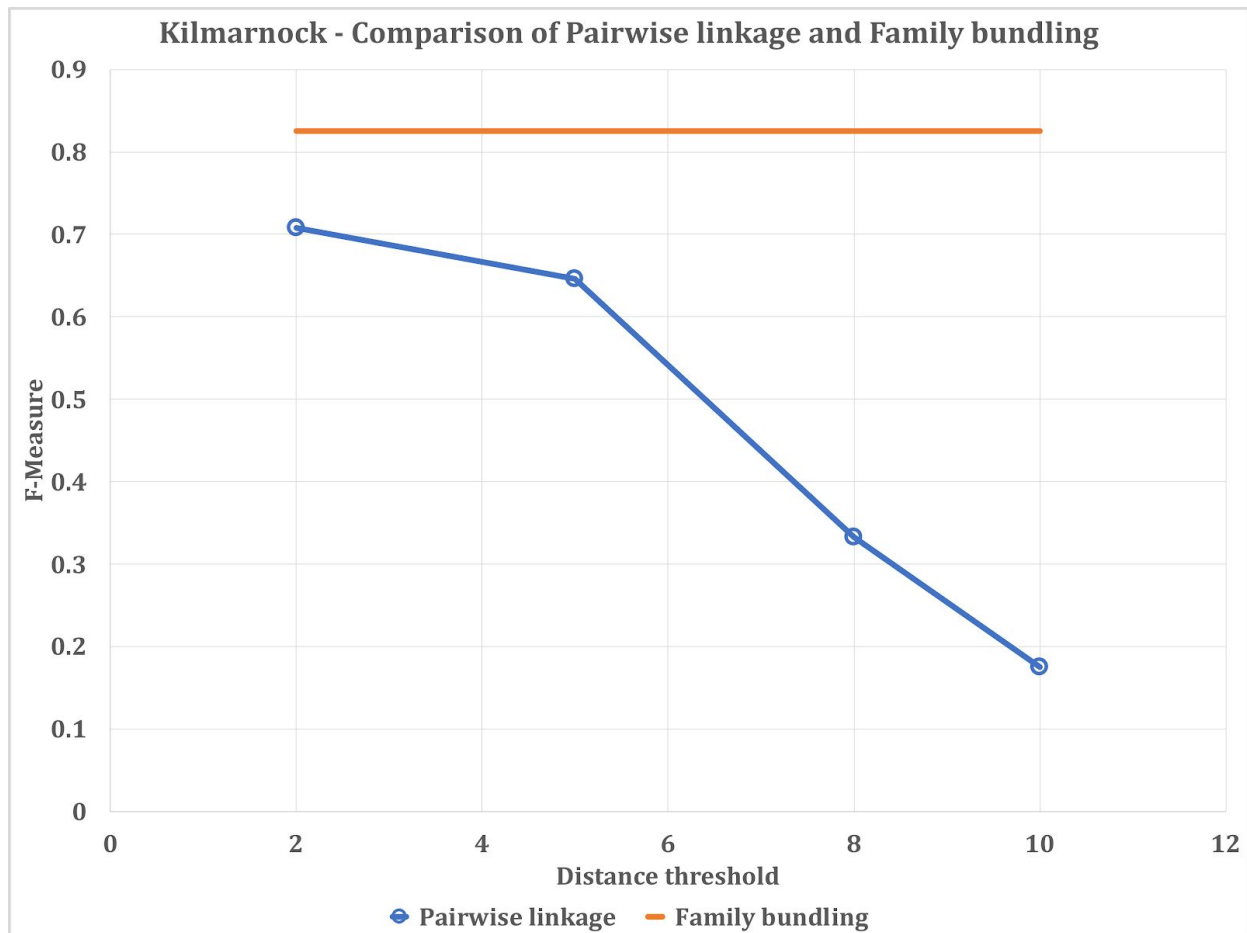

Figure 12: Comparison of the F-Measure values obtained by performing pairwise linkage and family bundling on the Kilmarnock data set. We present the best family bundling configuration as a target line in comparison to several settings of distance threshold for the pairwise linkage algorithm.

In the Kilmarnock data set we observe a similar outcome. The best F-Measure value using our proposed family bundling algorithm is 83% (see Section 5.4), and the F-Measure values obtained with the traditional pairwise distance thresholds 2, 5, 8, and 10 are 71%, 65%, 33%, and 18%, respectively.

Raw results and the source code used for performing individually linked records is available at (Akgün 2018).

6. Future Directions

Sibling bundling is the backbone of family group based population reconstruction, but it does not provide a fully linked data set by itself. For example, babies on birth records are not linked directly to their death certificates, hence we cannot calculate measures such as average life span with respect to that child's place in the family birth order, etc. We identify in Table 1 all the different roles that can be played by an individual on Scottish vital event records. Future research directions include developing ways of making all these other kinds of links between the roles in different types of records. We believe having the family structure at hand will make these tasks easier.

The establishment of intergenerational links, for example between the father on a birth record and the groom on a marriage record, benefits from having the sibling bundles available. Each full sibling must have the same father and the same mother, hence they must be linked to the same marriage record (unless the marriage record is missing). Making the link for one sibling immediately allows making all the links for the other siblings.

For a particular kind of link, either family links, or links between different roles of an individual through their records, there are multiple ways of performing linkage. We show in this paper that there are a large number of configurations of a single algorithm to perform the same kind of linkage. We also show that family linkage can be performed using different types of records (the *family_hints* parameter of our algorithm). The fact that we can perform family linkage using entirely different sets of records means that a data set containing vital event records contains significant redundancy. For example, sibling links can be established using only the birth records, or using only the death records. This redundancy provides opportunities for cross-checking linkage results.

Performing the same linkage tasks using two different methods, and then reasoning about the results in a combined manner can (a) provide higher confidence in the links by only keeping those links that appear in both linkages, and (b) provide a larger number of links by keeping a union of the results of the two linkages. The former corresponds to increasing precision, and the latter corresponds to increasing recall. Depending upon the desired end use of the linked data set, either precision or recall may be more important. Finally, this approach can be generalised to multiple linkages instead of only two. We plan to explore the possibility of increasing linkage quality by performing multiple linkages independently and interpreting the results in combination.

We also plan to investigate approaches to further check the consistency of the obtained links such as temporal limitations with regard to time intervals between births or the ages of parents. We will identify any potentially inconsistent results and use domain experts to manually validate these results, which in turn will provide us with additional training data of ambiguous cases which can be used to improve any future learning based linkage

approaches (Antonie 2013, Christen 2012, Goeken 2011). We will also use domain expertise to manually assess inconsistent results that occur due to homonymy.

We use the M-tree data structure to efficiently find similar data points with respect to a given data point. We intend to further explore the effects of choosing different distance metrics on search efficiency, and to investigate the applicability of other similarity search techniques. More generally, we plan to compare the similarity search approach to existing blocking approaches (Christen 2012), both for quality of the results and computational scalability.

Finally, we plan to further develop ongoing work (Dalton 2018) on generating synthetic genealogical populations, with the intention of producing realistic synthetic sets of vital event records with known ground-truth. We intend to investigate the extent to which evaluating a linkage algorithm using such synthetic data can reach conclusions that are representative of performance on real data.

7. Conclusions

This paper outlines a process for family group based population reconstruction using vital event records. It presents an algorithm to perform one component of the process: sibling bundling. This algorithm has two phases, the first for forming initial family groups accurately (with high precision), and the second for merging these family groups into larger families if groups are sufficiently close to one another.

We also present an experimental study on two small data sets from Scotland, which samples a number of values for the algorithm's six parameters. This shows that choosing appropriate values for the algorithm parameters is crucial for achieving high-quality linkage. We present a dominance analysis to explain the linkage quality achieved by the different configurations. For some of the parameters, such as the distance thresholds between records and the maximum family size threshold, we identify optimal values for the two data sets. For other parameters, such as the distance method used when comparing families and the record type that is used for identifying family groups, our results are less conclusive. These parameters seem to have less of an effect on the linkage quality than the distance thresholds and the maximum family size.

The reconstruction of population structures through linkage of historical records is a powerful approach to generate longitudinal historical microdata resources of interest to researchers in various fields. Family group based methods offer a promising approach to this problem. This paper shows the efficacy of such methods on two real-world data sets. We plan to expand and improve on these methods in the future.

Acknowledgements

References

Akgün, Ö. 2018. *Supplemental Files for 'Linking Vital Event Records Using Family Groups'.* https://github.com/digitisingscotland/hm-familygroups.

Antonie, L., K. Inwood, D. J. Lizotte, and J. A. Ross. 2014. Tracking People Over Time in 19th Century Canada for Longitudinal Analysis. *Machine Learning* 95 (1): 129–46. doi:10.1007/s10994-013-5421-0.

Bailey, M., C. Cole, M. Henderson, and C. Massey. 2017. How Well Do Automated Methods Perform in Historical Samples? Evidence From New Ground Truth. *NBER Working Paper 24019. Cambridge, MA: National Bureau of Economic Research*. doi:10.3386/w24019.

Blaikie, A., E. Garrett, and R. Davies. 2005. Migration, Living Strategies and Illegitimate Childbearing: a Comparison of Two Scottish Settings, 1871-1881. In *Illegitimacy in Britain, 1700-1920*, ed. A. Levene, T. Nutt, and S. Williams, 141–67. Palgrave Macmillan.

Bloothooft, G., P. Christen, K. Mandemakers, and M. Schraagen, eds. 2015. *Population Reconstruction*. Springer International Publishing. doi:10.1007/978-3-319-19884-2.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer.

Christen, P. 2016. Application of Advanced Record Linkage Techniques for Complex Population Reconstruction. *arXiv.org*. 1612.04286.

Christen, V., A. Groß, J. Fisher, Q. Wang, P. Christen, and E. Rahm. 2017. Temporal group linkage and evolution analysis for census data. *International Conference on Extending Database Technology, Venice, Italy*, 620-31.

Ciaccia, P., M. Patella, and P. Zezula. 1997. M-Tree: an Efficient Access Method for Similarity Search in Metric Spaces. *23rd VLDB Conference, Athens, Greece*. Morgan Kaufmann, 426–35.

Dalton, T. 2018. *ValiPop: a Micro-simulation Model for Generating Synthetic Genealogical Populations*. https://stacs-srg.github.io/population-model/.

Davies, H. R. 1992. Automated Record Linkage of Census Enumerators' Books and Registration Data: Obstacles, Challenges and Solutions. *History and Computing* 4: 16–26.

Davies, H. R. 1993. *Nominal Record Linkage of Historical Data: Procedures and Applications in a North Wales Parish*. University of Southampton. http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.238880.

Davies, H. R, and E. Garrett. 2005. More Irish Than the Irish? Nuptiality and Fertility Patterns on the Isle of Skye, 1881-1891. In *Ireland and Scotland: Order and Disorder, 1600-2000*, ed. R. J. Morris and L. Kennedy, 85–100. Edinburgh: John Donald Publishers.

Dibben, C., L. Williamson, and Z. Huang. 2012. *Digitising Scotland*. Economic and Social Research Council. http://gtr.rcuk.ac.uk/projects?ref=ES/K00574X/2.

Fellegi, I. P., and A. B. Sunter. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association 64 (328)*. American Statistical Association: 1183–1210. doi:10.2307/2286061.

Fu, Z., H. M. 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 (2). Edinburgh University Press: 204–25. doi:10.3366/ijhac.2014.0130.

Fu, Z., P. Christen, and J. Zhou. 2014. A Graph Matching Method for Historical Census Household Linkage. *Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taiwan*, 485–96.

Garrett, E., and R. Davies. 2003. Birth Spacing and Infant Mortality on the Isle of Skye, Scotland, in the 1880s: a Comparison with the Town of Ipswich, England. *Local Population Studies* 71: 53–74.

Gautier, E., and L. Henry. 1958. *La Population De Crulai: Paroisse Normande*. Paris: Presses universitaires de France.

Goeken, R., L. Huynh, T. Lenius, and R. Vick. 2011. New Methods of Census Record Linking. *Historical Methods* 44 (1). NIH Public Access: 7–14. doi:10.1080/01615440.2010.517152.

Han, J., M. Kamber, and J. Pei. 2012. *Data Mining: Concepts and Techniques*. Boston: Morgan Kaufmann.

Hand, D., and P. Christen. 2018. A Note on Using the F-Measure for Evaluating Record Linkage Algorithms. *Statistics and Computing*, *28*(3), 539-547. doi:10.1007/s11222-017-9746-6.

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics-Doklady* 10 (8): 707–10.

Kemmer, D. 1990. *Marital Fertility of Edinburgh Professionals in the Later Nineteenth Century*. The University of Edinburgh. http://hdl.handle.net/1842/24036.

Kemmer, D. 1997. Investigating Infant Mortality in Early Twentieth Century Scotland Using Civil Registers: Aberdeen and Dundee Compared. *Journal of Scottish Historical Studies* 17 (1): 1–19. doi:10.3366/sesh.1997.17.1.1.

Paddock, R. J. M. 1989. *Aspects of Illegitimacy in Victorian Dumfriesshire*. University of Edinburgh. http://hdl.handle.net/1842/6883.

Reid, A., R. Davies, and E. Garrett. 2002. Nineteenth-Century Scottish Demography From Linked Censuses and Civil Registers: a 'Sets of Related Individuals' Approach. *International Journal of Humanities and Arts Computing* 14 (1-2): 61–86. doi:10.3366/hac.2002.14.1-2.61.

Reid, A., E. Garrett, R. Davies, and A. Blaikie. 2006. *Scottish Census Enumerators' Books: Skye, Kilmarnock, Rothiemay and Torthorwald, 1861-1901*. doi:10.5255/UKDA-SN-5596-1.

van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd ed. Butterworth-Heinemann. http://www.dcs.gla.ac.uk/Keith/Preface.html.

Ruggles, S. 2002. Linking Historical Censuses: a New Approach. *International Journal of Humanities and Arts Computing* 14 (1-2): 213–24.

Wrigley, E. A. 1966. Family Reconstitution. In *An Introduction to English Historical Demography: From the Sixteenth to the Nineteenth Century*, ed. E. A. Wrigley, 96–159. London: Weidenfeld & Nicolson.

Wrigley, E. A., and R. S. Schofield. 1973. Nominal Record Linkage by Computer and the Logic of Family Reconstitution. In *Identifying People in the Past*, ed. E. A. Wrigley, 64–101. Edward Arnold.

Wrigley, E. A., R. S. Davies, J. E. Oeppen, and R. S. Schofield. 1997. *English Population History From Family Reconstitution 1580-1837*. Cambridge University Press. doi:10.1017/CBO9780511660344.