Sources and methods

Recent developments in making family reconstitutions

Gill Newton

Family reconstitution is a set of rules for linking historical parish records of baptism, marriage and burial into nuclear family groups, and a set of methods for measuring the demographic characteristics of these reconstituted families without bias. The aim is to arrive at a sample of the inhabitants of a parish for which it is possible to control for migration, so that the size of the population at risk for any given demographic measure is known. The purpose of this note is to outline the evolution of techniques for making (as opposed to analysing) family reconstitutions at the Cambridge Group for the History of Population and Social Structure, and to draw attention to recent methodological work on applying these techniques to large urban parishes.

The French demographers Michel Fleury and Louis Henry first developed family reconstitution into a method that was suitable for demographic analysis, and applied it to historical data.[1] Fleury and Henry were primarily interested in fertility. Tony Wrigley adapted the technique for use with English historical parish records, with a broader and more equal focus on nuptiality and mortality as well as fertility.[2] His seminal family reconstitution of the parish of Colyton in Devon was followed by a long-running programme of research into English historical demography at the Cambridge Group for the History of Population and Social Structure.[3] Dozens of family reconstitutions of English parishes would be undertaken, and 26 of them were eventually selected as being of sufficient quality for rigorous analysis.[4]

1 M. Fleury and L. Henry, *Manuel de dépouillement et d'exploitation de l'état civil ancien* (Paris, 1956, 3rd edition 1985).

2 E.A. Wrigley, 'Family reconstitution', in E.A. Wrigley ed., *An introduction to English historical demography from the sixteenth to the nineteenth century* (London, 1966), 96–159.

3 For a description of the technique of family reconstitution based on Colyton see E.A. Wrigley, 'Some problems of family reconstitution using English parish register material', *Proceedings of the 3rd international economic history conference, Munich, 1965. Section VII, demography and economy* (Paris, 1972), 199–221. This sets out a desiderata of parish characteristics for family reconstitution, including an itemised account of typical detail to be found within a parish register and their relative importance. This is partially reiterated in Wrigley, 'Family reconstitution' (1966), here at 102–111.

4 For more information on the 26 parish family reconstitutions see E.A. Wrigley, R.S. Davies, J.E. Oeppen and R.S. Schofield, *English population history from family reconstitution 1580–1837* (Cambridge, 1997).

It is important to distinguish between the method of making a family reconstitution and the means of analysing it, and to appreciate that both aspects involve substantial work. The Cambridge Group family reconstitutions were analysed by computer, but the record linkage that makes each reconstitution was carried out by hand on paper forms, using the techniques set out in Wrigley, 'Family reconstitution' (footnote 2). Some of those reading this will remember the local historian volunteers who so generously donated their time to this cause, without whom it would not have been possible to proceed. The amount of effort may readily be appreciated when contemplating the paper family reconstitution forms that fill several large filing cupboards in the Cambridge Group's archive, each form completed by hand and itself summarising information from numerous handwritten baptism, burial and marriage slips.

The time-consuming nature of making family reconstitutions must have been apparent to all involved. Efforts were soon underway to write a computer program or programs that could automate the process. In 1973 Roger Schofield and Tony Wrigley published an article on the topic, and on the wider logic of record linkage in family reconstitution.[5] The bulk of this article detailed methods for resolving conflicting possible matches between records, but it also succinctly set out a sequence of operations necessary to link historical records into a family reconstitution, and outlined the biological and heuristic constraints that should be applied. These constraints include, for example, the minimum interval between one birth and the next to the same woman, or the probable age of menarche.[6]

Wrigley and Schofield's article described a theoretical basis for automating record linkage based on Newcombe and Kennedy's theories of rule-based data matching.[7] Much of this concerns decision-making between competing matches in record linkage, which is a complex task, but one well-suited to a mathematically definite solution. Other aspects of the process that were less readily representable in mathematical (and hence programmatic) terms received less attention, although Wrigley and Schofield touched on the standardisation of input data (normalisation), especially names. In the case of an irregular and changeable data format such as parish registers, both the types of identifying data present and the way in which they are expressed will vary considerably. This makes normalisation a much more significant part of the necessary work than was perhaps appreciated at the time.

5 See E.A. Wrigley and R.S. Schofield, 'Nominal record linkage by computer and the logic of family reconstiution', in E.A. Wrigley ed., *Identifying people in the past* (London, 1973).

6 These rules were set out more fully in Appendix 4 of Wrigley et al., *English population history from family reconstitution*.

7 The rules-based approach applied to vital records is described in H.B. Newcombe, J.M. Kennedy, S.J. Axford and A.P. James, 'Automatic linkage of vital records', *Science*, 130 (1959), 954–59. A generalised mathematical model for record linkage was defined in I.P. Fellegi, and A.B. Sunter, 'A theory for record linkage', *Journal of the American Statistical Association*, 64 (1969), 1183–1210.

The variability of English parish registers implies that a large amount of work is necessary to arrive at a fully rule-based solution, if a rule is to be defined that can accommodate every possible matching circumstance. Once a set of rules has been devised, it is important to be able to test that they work. It is difficult to evaluate whether record linkage has been correctly and comprehensively achieved unless we know what we were expecting to see in the output, so it is useful to have test datasets that have been reconstituted by hand for the purpose of comparison. In this respect there were limitations to what was available. All of the Cambridge Group's family reconstitutions were made machine readable for the purpose of demographic analysis by computer, but the baptisms, burials and marriages that underpin them had not been separately required and were therefore not similarly available in machine-readable form. One test dataset for the efficacy of algorithms under development was obtained by making part of the already reconstituted Colyton parish registers machine-readable.

The next step towards proof of the concept of automated creation of family reconstitutions was to be a computerised reconstitution of the very large London parish of Clerkenwell.[8] This formed part of the intended thesis of Amanda Copley, a doctoral student at the Cambridge Group. The choice of this parish as a project for automated record linkage was partly determined by expediency, since with an overall average of 215 baptisms per year between 1565 and 1753 (and up to 400 per year in the first half of the eighteenth century) it was deemed too large to reconstitute by hand. This much larger record set presented a number of new challenges to automated record linkage. The extent and format of identifying information in the Clerkenwell baptisms, burials and marriages was quite dissimilar in several respects to those in the Colyton test dataset.

One difficulty that proved particularly intractable concerned names. Preliminary record linkage attempts revealed limitations in the name standardisation approach described by Wrigley and Schofield in 1973, where it was envisaged that the bulk of the work will be done by algorithm and subsequent manual amendments would be a relatively small task. This method was insufficiently discriminatory when used on a large suburban London parish, and a great deal of manual intervention looked likely to be needed. Work on Clerkenwell was brought to a halt by the sad and untimely death of Amanda Copley in 1988. Subsequently, there was a shift away from pursuing fully automated record linkage as an end in itself, due to changes in the Cambridge Group's funding circumstances.

A new approach in constructing family reconstitutions was made possible by the advent of desktop computers with powerful software. This meant family reconstitutions could be constructed more efficiently, without necessarily having to define an exhaustive set of rules for record linkage at the outset. A computer-assisted rather than computer-determined

8 Clerkenwell actually comprises two parishes from 1723 onwards. St James Clerkenwell was the mother parish from which St John Clerkenwell was created in 1723, and records from both parishes were gathered for the family reconstitution.

method could be used, combining the sequence of operations specified for manual record linkage with the sorting and cross-comparison capacity of databases. Thus, a relational database can be used to store baptism, burial and baptism records together with associated lookup tables for names and other normalised variables (such as occupation, ages and so forth) , and SQL queries assist in sorting and joining records that potentially involved the same individuals.[9] Manual intervention sifts the good links from the bad, and may recover other possible links not envisaged in the lookup tables. Microsoft Access databases have been used in this manner for a number of research projects, including the family reconstitution of a cluster of Oxfordshire parishes.

In 2004 research that entailed the creation of family reconstitutions for several London parishes began at the Cambridge Group, in collaboration with the University of London. Small-scale London family reconstitution studies had been constructed by manual methods, but no large suburban parish had successfully been reconstituted by any method, and part of this research was intended to revisit the suburban parish of Clerkenwell, whose records were originally gathered by Amanda Copley, as well as the even larger suburban parish of St Botolph Aldgate.

Urban parishes present special problems for family reconstitution. The process of record linkage is much more fraught with uncertainty than in the rural communities and market towns which formed the basis of the Cambridge Group's original 26 parish sample. The problems in making use of urban family reconstitutions for demographic analysis concern how to mitigate the effect of much more frequent and widespread migration, but there is a necessary precursor to this that must be solved when constructing the reconstitution: how to identify individuals with reasonable confidence in the first place. With a large parish that is growing substantially through in-migration and which is only part of a much larger conurbation, there is an elevated risk that two separate records of a person of the same name and with complementary characteristics will not be the same individual. Any parish is only a sample drawn from the total population of London (the size of which is known only approximately), and the magnitude of the risk of mistaken identity is difficult to define in absolute terms. However, clearly the faster a parish is growing and the larger its population, the greater the risk of it containing two or more unrelated individuals resident at the same time who shared the same characteristics. Suburban parishes are particularly susceptible to this problem because they are among the largest and fastest growing single units of settlement in England in the sixteenth and seventeenth century, with many thousands of inhabitants.

To recap, the database-assisted method of record linkage described above was used successfully for both a cluster of small parishes in Cheapside in the centre of London and

9 SQL, or Standard Query Language, is the standard database command language. In the sense it is used here, a query can be thought of as an instruction to retrieve and combine copies of records stored within the database according to criteria specified by the user.

for a revisiting of the suburban parish of Clerkenwell. During the Clerkenwell family reconstitution it became clear that choosing links at random between equally probable competing matches introduced undesirable amounts of 'noise' in the results. Two important modifications to the overall method were therefore introduced during this research. The first concerned improved methods for name standardisation, on which so much of the success of historical record linkage depends. The second was the development of new data structures that would enable the persistence of the link between parish register and reconstituted families beyond the period of initial construction, so that in the completed database a user interface allows the inspection of the parish register data as input that forms the basis for each reconstituted family. This forms a useful exploratory tool during record linkage as well.

A family reconstitution for the eastern London suburb of Aldgate demanded a more fully automated solution, Aldgate being considerably more populous even than Clerkenwell, with an overall average of 330 baptisms per year between 1558 and 1710 (and up to 600 per year by the first decade of the eighteenth century). A new programmatic method was deployed, which developed a way of mitigating the challenge of correctly determining identity in a large and rapidly growing parish. In outline, the risk of mistaken identity can be counterbalanced by taking advantage of the much larger pool of individuals from which a sample of reconstituted families may be drawn. The volume of the input data makes it viable to select more stringently in forming this sample, and yet to end up with a sufficient number of reconstituted families for analysis.

To reduce uncertainty, in any family reconstitution it is best to discard potential matches between records where there are equally probable competing matches, at least when subjecting the reconstituted families to analysis. Effectively, this is an *ad hoc* means of removing from consideration those with the most popular names, or other ambiguous characteristics. Such an approach restricts the number of useable reconstituted families, but it also reduces the 'noise' that erroneous links might otherwise make, and increases the confidence that can be placed in the final result. Taking this discarding of conflicting matches a step further, in the most populous parishes, the *existence* of conflicting matches between two records based on (standardised) names and dates alone (and age and occupational information where given) may be taken as an indicator of unacceptably high risk of mistaken identity. In such circumstances, all families entailing matches where there are conflicting possibilities can be discarded immediately, without further scrutiny of the relative merits of those matches.

It must be stressed that this new automated approach is certainly not the only or necessarily the best means of making a family reconstitution in every circumstance. Different parishes vary enormously in size, rate of growth and the richness and completeness of their records, and the choice of method should reflect that as well as pragmatic considerations such as the time available. Reconstituting programmatically by all of the methods discussed still depends on substantial work by hand to create the

necessary normalised input, involving considerable pre-processing of names, dates, occupations, residences and ages into suitably standardised yet flexible forms.

There are further possibilities for automated methods that require less manual normalisation of data, and which do not depend on predefined rules for each circumstance. Peter Kitson at the Cambridge Group is developing a method that starts by representing all possible links between couples that can be made using standardised forenames alone. This is gradually pared down into the most viable links and hence reconstituted families. It is hoped that probabilistic Bayesian reasoning will eventually enable the steps in this decision-making process to be undertaken without manual intervention.[10] This promises to be a good approach for small to moderately large parishes.

To summarise, this note has described the making of family reconstitutions for demographic analysis over several decades at the Cambridge Group for the History of Population and Social Structure. During this time the circumstances in which family reconstitution may be applied has evolved. The adaptation of automated or partially automated family reconstitution to parishes with different characteristics has made apparent the crucial role of normalisation or data standardisation in the process. A number of alternate strategies for handling the record linkage steps on which family reconstitution depends have also been developed. The most appropriate method in a particular circumstance will depend on the characteristics of the parish (or parishes) in question. and in particular, its size, migration and rate of population growth, and the richness and completeness of the baptism, burial and marriage registers.

For further information on methods that have been used in the construction of large urban family reconstitutions for several parishes in early modern London, see the forthcoming working paper by Gill Newton entitled 'Family reconstitution in an urban context: some methods and observations', to be published by the Centre for Metropolitan History in connection with the ESRC-funded 'Life in the suburbs: health, domesticity and status in early modern London' research project. A draft version of this paper is available online at http://www.geog.cam.ac.uk/people/newton/UrbanFamilyReconstitution.pdf

10 The use of probabilistic reasoning in record linkage forms part of most record linkage strategies, but the use of Bayesian networks initially formed a separate strand of enquiry. Record linkage for the U.S. census draws on both traditions, and an overview of the Bayesian approach and record linkage strategies in general is presented in: W.E. Winkler, 'Methods for record linkage and Bayesian networks' [2002], unpublished paper available from http://www.cs.cmu.edu/~wcohen/matching/WinklerAsa02.pdf