Genealogical Data Linkage: Approaches to creating multiple linkage solutions with indications of provenance and uncertainty

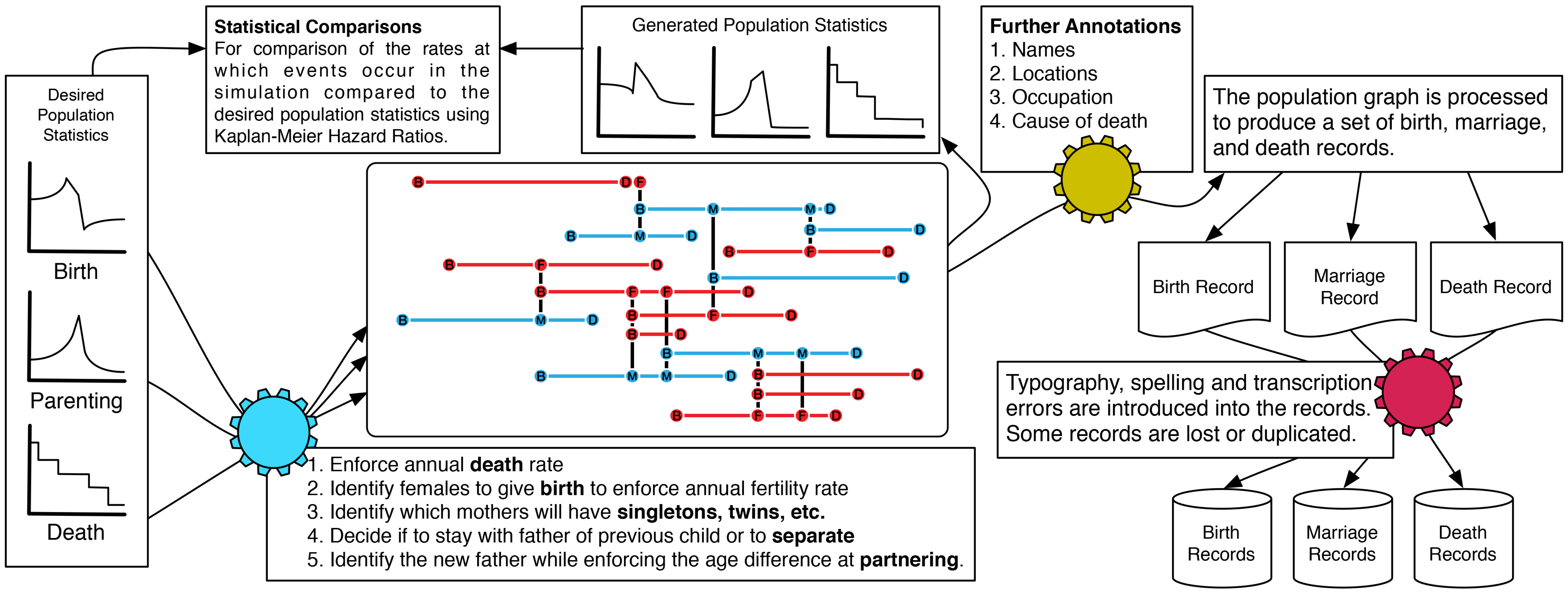
Tom S. Dalton, Alan Dearle, Graham N. C. Kirby, Özgür Akgün

Overview & Aims

The research domain is probabilistic data linkage (PDL) with a specific focus on population reconstruction. There is a range of established PDL approaches that will be used and potentially extended. The end goal of the project is to link large-scale genealogical datasets and maintain multiple linkages with associated provenance and metadata.

Evaluating the success of any data linkage approach at scale presents an issue. To evaluate linkage a ground truth is needed. This is impossible to attain for real-world large-scale genealogical datasets. Therefore, we need to be able to generate synthetic large-scale populations that are statistically similar to real world populations.

The population generation algorithm should take a set of statistical distributions as its input and produce a population with statistics sufficiently close to the input. The algorithm should model a range of social and biological attributes, including parentage, migration, marital status, and occupation. Issues that need to be solved include the handling of conflicts between input distributions, the modeling and set-up of the simulation, and the scaling of the pairing approach within the algorithm.



The application of the statistical inputs to the population requires especially careful thought regarding how the population is initialised and also the order in which statistical distributions are enforced to avoid conflicts.

For the population being initialised at the beginning of the simulation, we require information from before the start time to fill in the output event records for the people alive at the start time - e.g. the maiden name of the mother on their birth record.

In the population model, we are foremost interested in the biological events that allow us to build a family tree structure as shown above.

Once we have confirmed the biological structure to be statistical accurate we then annotate the structure with social and geographical information that will populate the output records.

The output from the population simulation will be in the form of statutory event records, see below.

Often the transcription and reading of these can lead to error and omission in the data. Therefore, as they are the likely source of input data for our data linkage algorithms we need to be able to mimic these kinds of errors which exist in the real-world dataset.

Age	Males		
x	m_x	q_x	l_x
0	0.004360	0.004351	100000.0
1	0.000287	0.000287	99564.9
2	0.000156	0.000156	99536.4
3	0.000134	0.000134	99520.9
4	0.000114	0.000114	99507.6
5	0.000139	0.000139	99496.2
6	0.000153	0.000153	99482.4
7	0.000108	0.000108	99467.2
8	0.000110	0.000110	99456.4
9	0.000135	0.000135	99445.5
10	0.000060	0.000060	99432.0
11	0.000127	0.000127	99426.1
12	0.000090	0.000090	99413.4
13	0.000076	0.000076	99404.5
14	0.000117	0.000117	99396.9
15	0.000095	0.000095	99385.3
16	0.000270	0.000270	99375.9
17	0.000544	0.000543	99349.0

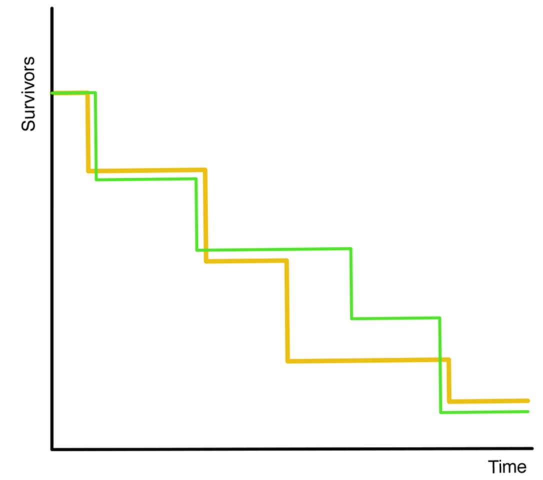
Statistical Input Example

The population model takes a range of statistical distributions as input. The life table, to the left, is used to control rates of death in the population. There is a life table for each year or time period.

The table provides the probability (q_x) of a person of a given age dying before their next birthday. It also provides the rate (m_x) at which people of a given age die in each year and also the number of survivors (l_x) from the starting population.

Related Research

Research of a similar nature can be seen in areas including, microsimulation of travel behaviour (Guo & Bhat, 2008); probabilistic data generation (Pudjijono & Christen, 2009); and data access control (Nowok et al., 2015). This work differs significantly, however, as it allows synthetic longitudinal data to be generated across many generations.



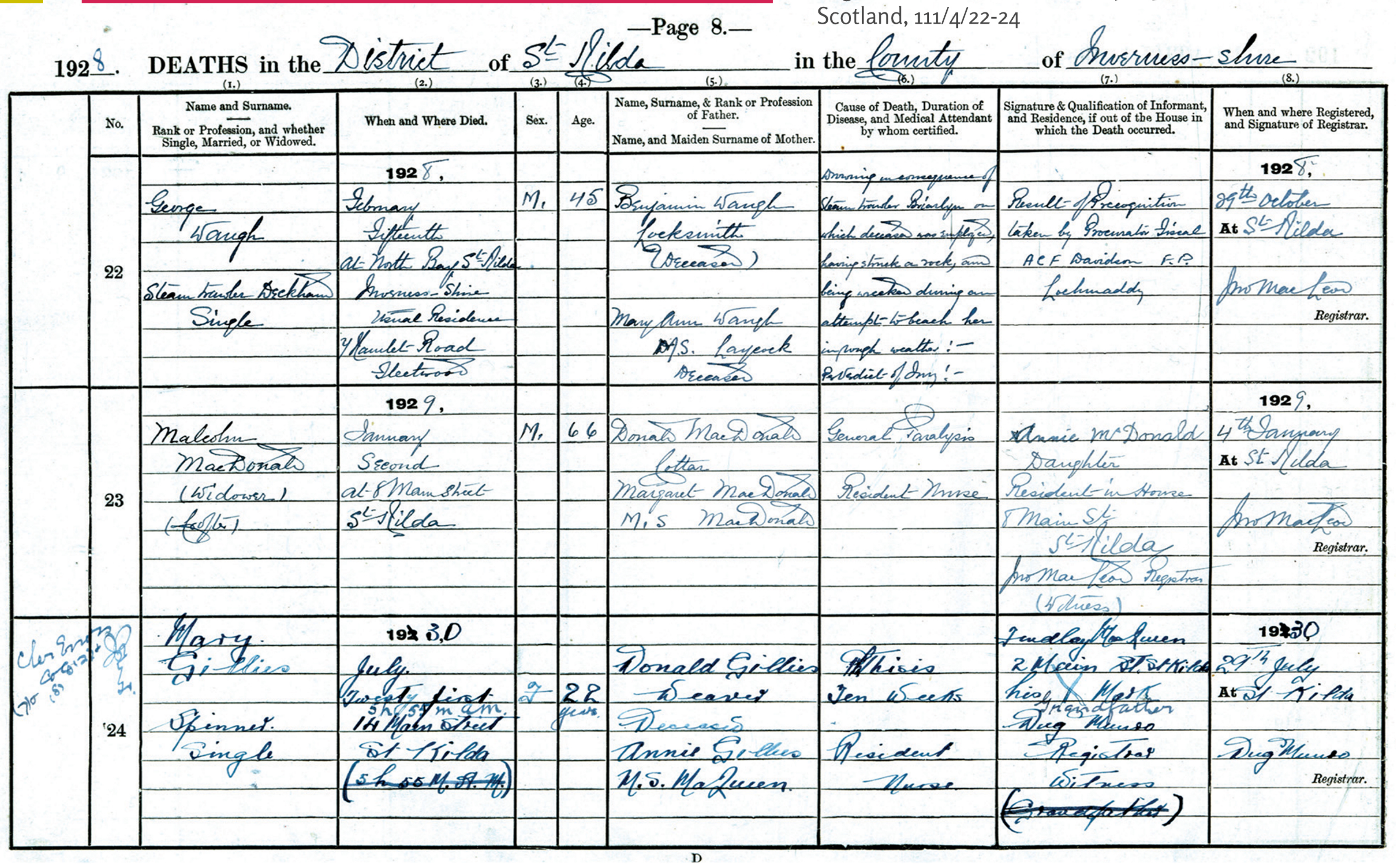
Statistical Comparisons

We are able to consider most of the biological events in our model as things that happen to a group of individuals at certain points in time. This allows us to be able to derive a survival table which is similar to the l_x column in the statistical input example.

We are able to plot these survival tables as graphs (see above) for the statistical input distributions and the generated population.

The Kaplan-Meier method (Kaplan & Meier, 1958) allows us to compare these curves and identify if the difference is statistically significant. By using this approach we are able to assert if the generated population statistics match up with the desired population statistics.

Register of Deaths, St Kilda, 1921-30, National Records of Scotland, 111/4/22-24



References

Guo, J., & Bhat, C. (2008). Population synthesis for microsimulating travel behavior. Transportation Research Record: Journal of the Transportation Research Board.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

Nowok, B., Raab, G. M., & Dibben, C. (2015). synthpop: Bespoke creation of synthetic data in R.

Pudjijono, A., & Christen, P. (2009). Accurate synthetic generation of realistic personal information. In Proceedings of the 13th pacific-asia conference on advances in knowledge discovery and data mining.