# Evaluation of Simulation Model

The code for the simulation model, summary inputs, supporting analysis and raw data for all plots are available on Github (Dalton 2017). The focus of our evaluation is to demonstrate the scalability and correct behaviour of the generator. The inputs used for the were hand constructed and draw on publically available Scottish and English population statistics for the period.

* 1. Failure rates

Within the 45 synthetic populations generated for the evaluation we saw no failures on account of an insufficient number of available people for specific required roles within the simulation. In development we have seen this is not always the case when artificial input statistics are used. However, our evaluation populations, given their use of reasonably and inter-compatible input statistics, demonstrate that for reasonable inputs the population generator will be able to create a full population reliably.

* 1. Scalability

The 45 synthetic populations we have produced are of five sizes, approximately 0.1, 2, 4, 8.5 and 17 million individuals in total size. Below we show plots of runtimes against population size and validity against population size.

Plot 1 demonstrates the computational scalability of *ValiSim* to large populations. The time taken to generate the contingency table is by far the largest time cost. It should also be noted that our experiments indicate the simulation runtime to be near linear. The value used for the contingency table runtime is the average run time when the expected frequencies have been pre-calculated. The same pre-calculated frequencies can be used for populations of the same size and using the same input statistics. The time to run the categorical analysis using R is independent of population size and also negligible (4.1 seconds for valid populations and 10.2 seconds for invalid populations).

Plot 2 shows the runtimes to generate the contingency table of expected frequencies. Although these are the large time cost within the simulation model we only needed to create one contingency table of expected frequencies for each size of population.

Plot 3 shows the proportion of valid populations at each population size. It can be seen that *ValiSim* performs well for populations up to four million in total size, however beyond this threshold the statistical tests for deaths and partnering fail. At this scale, we still produce genealogical reasonable populations but that do not fully comply with the input statistics.

Closer evaluation of the produced synthetic populations shows that a slight, but systematic, under achieving of the required statistics occurs. This is within the bounds of the statistical analysis for smaller populations. As populations increase in size the under achieving of statistics causes the deviation of the synthetic population from the model of the expected frequencies to increase so that the statistical tests begin to fail. We aim in the future to modify *ValiSim* to address this under achieving of required statistics. However, the populations of less than four million people are valid and useful for detailed linkage evaluation.

It may also be necessary to consider other statistical validation approaches for the larger size of populations. Due to our requirement for non-determinism we expect some deviation from the expected frequencies which will scale with population size. However, the test performed for each contingency table relies on a model with a fixed number of degrees of freedom independent of population size. This means eventually our expectation for variation due to non-determinism will overpower the number of degrees of freedom in the model and result in the failure of the underlying chi-squared test.





