A Bayesian method for deriving population statistics from multiple imperfect data sources

John Bryant
Statistics New Zealand, Christchurch, New Zealand john.bryant@stats.govt.nz

Patrick Graham
Statistics New Zealand, Christchurch, New Zealand patrick.graham@stats.govt.nz

Regional population counts, disaggregated by age, sex, and other variables, are a key output for national statistical agencies. In the absence of a complete population register, population statistics must be assembled from multiple imperfect data sources. The process of assembling these statistics is typically informal and labour-intensive. Statistics New Zealand has been developing a new statistical framework that allows the process to be formalised and automated.

At the heart of the new framework is a demographic account. A demographic account is an internally-consistent description of population, births, deaths, and migration, disaggregated by variables such as age, sex, region, and ethnicity. Regularities within the demographic account are captured by a set of sub-models. The relationships between the account and the data sources are captured by a further set of sub-models. The framework is fully Bayesian, and inference is carried out using Markov chain Monte Carlo methods. The framework is modular and flexible. A prototype of the framework has been constructed and tested, and software for the full model is currently under development.

We expect that the new methods will provide more accurate estimates than current methods. Moreover, unlike current methods, the new methods provide formal measures of uncertainty. This increases the feasibility of carrying out censuses less regularly, or of relying entirely on administrative data.

Because the new methods are formal and explicit, they can be used in simulation experiments. For instance, by withholding census data from the model, it is possible to assess how changes to the census would affect the accuracy of population statistics.

The methods do not yield individual-level datasets. However, they could provide weights to be used with individual-level datasets, such as datasets that had been constructed by linking administrative records. This would provide one means of addressing the serious problems of over-coverage and under-coverage in administrative data.

Key Words: Population statistics, Bayesian, administrative data, statistical modelling, uncertainty