Demographic Estimates and Projections Using Multiple Data Sources: A Bayesian Approach

Bryant, John
Statistics New Zealand
Dollan House
Christchurch 8042, New Zealand
E-mail: john.bryant@stats.govt.nz

Graham, Patrick
Bayesian Research
24 Bletsoe Ave
Christchurch 8024, New Zealand
E-mail: patrick.graham@xtra.co.nz

Subnational population estimates and projections are one of the key outputs of national statistical agencies. Local-level information about past and future demographic trends plays a central role in decisions involving billions of dollars of public and private expenditures. User expectations about accuracy, timeliness, and detail are high and rising, even as population censuses—the backbone of subnational population estimation in many countries— come under increasing financial pressure.

Statistical agencies have responded by widening the range of data sources they use. Most of the traditional sources, such as censuses and vital registration systems, were designed specifically for the production of demographic statistics. Most of the newer sources, such as tax returns, school rolls, and building consents, were designed for different purposes entirely. Much ingenuity has been devoted to the exploitation of these "administrative" data sources. However, current methods are highly labour intensive and complex, have difficulty coping with noise in the data, and provide little information about uncertainty.

In this paper we introduce a new Bayesian framework for subnational population estimates and projections. We draw on a growing literature applying Bayesian methods to demographic problems (e.g. Alkema et al., 2008; Bijak and Wiśniowski, 2010; Brierley et al., 2008; Congdon, 2008; Daponte et al., 1997; Lynch and Brown, 2010), as well as the general literature on missing data (e.g. Little and Rubin, 2002). The framework combines the evaluation of data quality, the estimation of historical demographic rates and counts, the projection of future rates and counts, and the assessment of uncertainty. We illustrate these ideas with some early results from a project to carry out subnational population estimation in New Zealand.

A general Bayesian framework for subnational population estimates and projections

We treat subnational population estimates and projections as an attempt to estimate demographic account Q. A demographic account is a set of population counts, births, deaths, and migrations, disaggregated by region, sex, age, and period (Rees, 1979). All values in an account must satisfy an accounting identity: the size of any subgroup at the end of a period must equal its size at the beginning of the period plus entries (e.g. in-migrations) minus exits (e.g. deaths). We treat Q as a latent construct that cannot be observed directly but must instead be inferred from available data X. Taking a Bayesian approach to inference, we compute the posterior distribution $p(Q|X)$.

Using parameter vector θ, $p(Q|X)$ can be expanded to

$$(1) \quad p(Q|X) = \int p(X|Q,\theta)p(Q|\theta)p(\theta)d\theta.$$

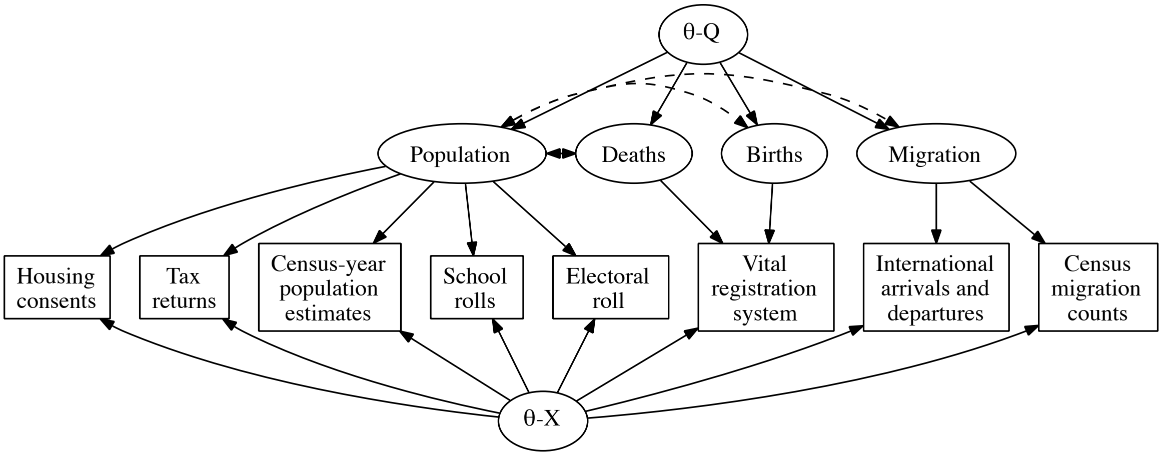

Figure 1: A summary of the model used in our application. The ellipses represent unobserved values, and the rectangles observed values. The solid lines represent probabilitic relationships, and the dashed lines deterministic relationships. Population, births, deaths, and migration nodes together make up the demographic account. The top two layers of the diagram, containing the parameter vector θ_Q and the demographic account, form the system model. The bottom three layers, containing the demographic account, the data sources, and parameter vector θ_X, form the observation model.

The first component of the integrand in (1) is an 'observation model' relating observable data to the state of the demographic system. The second component is a 'system model' governing the evolution of the demographic system. We assume that the observation and system models depend on distinct components of the parameter vector, meaning that $\theta = (\theta_X \theta_Q)$, $p(X|Q,\theta) = p(X|Q,\theta_X)$, $p(Q|\theta) = p(Q|\theta_Q)$, and $p(\theta) = p(\theta_Q)p(\theta_X)$.

Computation of the posterior distribution of the unknowns, $p(Q, \theta_Q, \theta_X|X)$, is carried out using a Gibbs sampler. The sampler alternates between the following full conditional distributions:

(2) $p(Q|X,\theta_Q,\theta_X) \propto p(X|Q,\theta_X)p(Q|\theta_Q)$

(3) $p(\theta_X|X,Q,\theta_Q) \propto p(X|Q,\theta_X)p(\theta_X)$

(4) $p(\theta_Q|X,Q,\theta_X) \propto p(X|Q,\theta_X)p(Q|\theta_Q)p(\theta_Q) \propto p(Q|\theta_Q)p(\theta_Q)$.

Application of the framework to subnational population estimation in New Zealand

We have developed specific system and observation models based on the ideas presented above and used them to carry out subnational population estimation in New Zealand. The models are summarised in Figure 1.

The system model seeks to capture sufficient demographic detail to yield accurate estimates, while still being computationally tractable. We distinguish between internal (domestic) migration and external (international) migration. We use a 'migration pool' specification for internal migration, which is more robust and meaningful than the traditional 'net migration' specification, but requires far fewer parameters than a full 'multiregional' specification (Wilson and Bell, 2004). We estimate separate sub-models for population size, births, deaths, internal in-migration, internal out-migration, external in-migration, and external out-migration. The parameters for each sub-model are distinct. However, values for population size appear in the sub-models for births, deaths, and out-migration in the form of exposure measures. Moreover, any combined draw from the sub-models is given a probability of zero if the draw violates the accounting identity described above. In practice, use of exposure measures and the imposition of the accounting constraint induces substantial dependence between realised values for the demographic series.

Hierarchical Poisson-gamma models (Christiansen and Morris, 1997) are used for each of the demographic series. Every region-sex-age-period cell in every demographic series is given its own expected value parameter. These parameters are in turn given hyperparameters capturing region, sex, age, and period effects. The large number of parameters provides sufficient flexibility that variants of the same specification can be applied to all series. The hierarchical priors provide sufficient shrinkage to protect against over-fitting.

The observation model contains a sub-model for each data source. Each sub-model treats values from the data source as a response and values from the demographic account as a predictor. This reverses the normal situation in population estimation whereby the demographic variable is the response and the data variable is the predictor. This reversal means that limited detail or gaps in the observed data can be dealt with by aggregating or subsetting the demographic account, rather than splitting or extrapolating from the data. Aggregating and subsetting are easier than splitting or extrapolating, so this saves a great deal of work. Consider, for instance, a situation where demographic estimates are required at the level of the district, but immigration data are only available at the level of the province. Under current estimation methods, immigrants would somehow need to be allocated to districts, perhaps by developing some set of splitting factors, or some form of spatial interpolation. Under our model, the district-level in-migration numbers from the demographic account are simply aggregated up to the provincial level before predicting provincial immigration data. Similarly, under current methods, data series that are available for only a few years or age groups require special treatment, but under our methods do not.

The sub-models for the data sources use essentially the same Poisson-gamma specifications as the sub-models for the demographic series. These specifications are sufficiently flexible to allow for the possibility that a data source undercounts some age groups more than others, for instance, or that a data source becomes more accurate over time.

Projections are constructed simultaneously with historical estimates. The system model in fact treats future values for demographic series no differently from historical values. The observation model treats future values for data sources as a form of missing data. All sub-models within the system and observation models use random walks for period effects, to facilitate projections .

Standard methods exist for updating hierarchical Poisson-gamma models within a Gibbs sampler. Considerable experimentation was required, however, to find a feasible method for updating demographic account Q. One difficulty is the sheer size of Q, which typically has tens of thousands or hundreds of thousands of cells. A second difficulty is the constraints imposed by the accounting identity. It is not possible, for instance, to simply draw from each of the sub-models and reject the draw if the accounting constraints are not satisfied, since the chance that the constraints would be satisfied is close to zero. The general solution we adopted was to restrict each update to the combination of (i) a single value, such as deaths in a single region, sex, age group, and period, and (ii) subsequent population counts would be affected by the change in value. A slightly more complicated scheme is needed for internal migration. Many terms in the Metropolis-Hastings ratio cancel. We are careful to take advantage of these cancellations, to speed calculations. Thousands of such updates of the demographic account are carried out for each update of the Poisson-gamma models.

Application to New Zealand data

We have been applying the framework to the problem of subnational estimation and projection

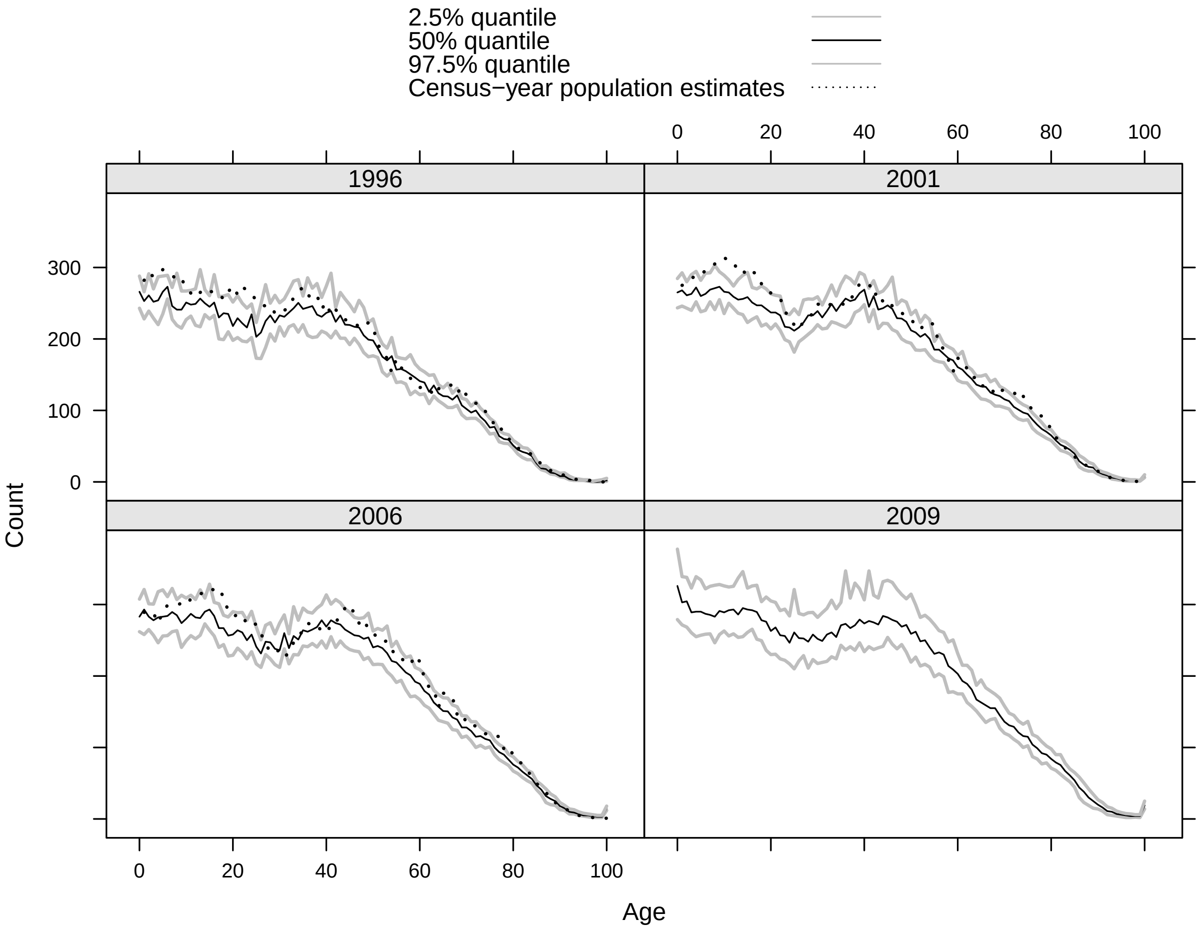

Figure 2: Population by age in a single region, males only, 1996, 2001, 2006, and 2009. The solid black line shows the median estimate and the solid grey lines a 95% credible interval. The dotted line shows (adjusted) census counts.

in New Zealand. The project includes an evaluation of the effect on estimates and projections of the cancellation of the 2011 census, following the Christchurch earthquake. At the time of writing, we have only just developed the software to the point where we can start constructing estimates. The results presented here should therefore be seen only as illustrative.

We have applied the model to a partly synthetic dataset covering 10 regions, 2 sexes, 101 age groups, and 14 years. The data sources are summarised in Figure 1. The results shown here are based on a simulation with two chains, a burn in of 2,500 iterations (where one iteration included 10,000 updates of the demographic account) and a sample of 1,000 iterations, though the sample was thinned, with only 1 in 4 iterations being recorded. This run was, unfortunately, not long enough for the model to converge properly, but deadlines did not permit a longer run. The acceptance rate for updates of the demographic account was 51%.

Figure 2 shows estimated estimated population counts for males, by age, in four different years. The black lines show the median estimates and the grey lines the 2.5% and 97.5% quantiles. The dotted lines show census population counts (adjusted undercount, temporary migrants, and demographic change between the census and reference date). The model has strayed further from the census counts than we would like, but we suspect that this is simply because the model has not had time to converge properly. In any case, the figure does provide a typical example of the output that can be obtained from the model.

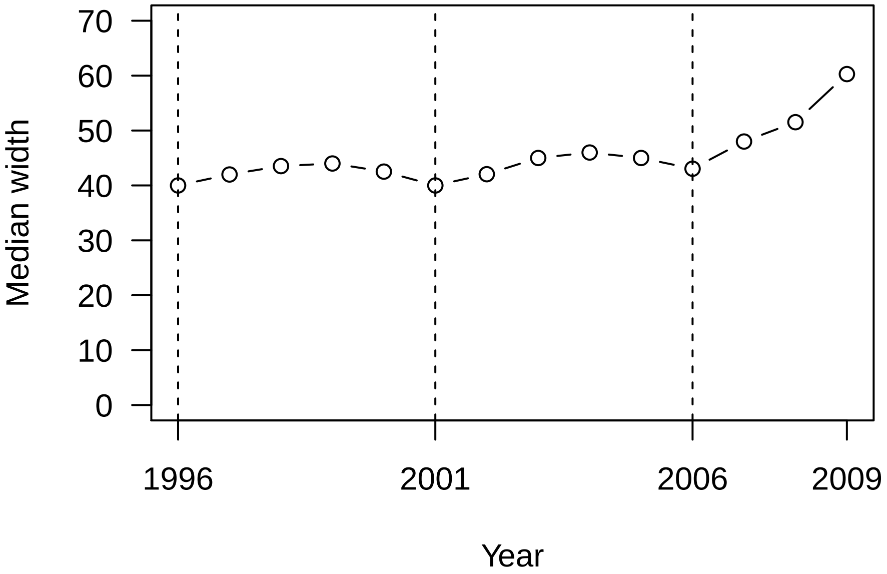

Figure 3: Median width of 95% credible intervals, by year. The vertical lines show census years.

An interesting feature of Figure 2 is the the width of the credible intervals three years out from a census. Figure 3 provides further results on changes in uncertainty over time. It shows the median width of credible intervals by year, aggregating across regions, ages, and sexes. The widths increase and then decrease gently over the two closed census intervals, but grow quickly over the open one. This finding is suggestive, though it will need to be confirmed by more detailed analysis.

Discussion

Although further work is needed to develop and validate the approach described here, we are optimistic that it will provide an attractive alternative to current methods for subnational estimates and projections. Use of a formal statistical model leads to greater transparency, and permits the automation of many tasks that must currently be carried out by hand. The inclusion of measures of uncertainty greatly increases the usefulness of the estimates and projections for decision-making. The structure of the model means that it can exploit numerous data sources, even if these sources produce noisy, irregular data.

REFERENCES

Alkema, L., Raftery, A., Gerland, P., Clark, S., and Pelletier, F. (2008). Estimating the total fertility rate from multiple imperfect data sources and assessing its uncertainty. Technical report, Centre for Statistics and the Social Sciences, University of Washington. Working paper 89.

Bijak, J. and Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected european countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society).*

Brierley, M., Forster, J., McDonald, J., and Smith, P. (2008). Bayesian estimation of migration flows. In Raymer, J. and Willekens, F., editors, *International Migration in Europe: Data, Models and Estimates.* Wiley and Sons.

Christiansen, C. and Morris, C. (1997). Hierarchical poisson regression modelling. *Journal of the American Statistical Association*, 92:618–632.

Congdon, P. (2008). Models for migration age schedules: a bayesian perspective with an application to flows between scotland. In Raymer, J. and Willekens, F., editors, *International Migration in Europe: Data, Models and Estimates*. Wiley and Sons.

Daponte, B., Kadane, J., and Wolfson, L. (1997). Bayesian demography: projecting the iraqi kurdish population, 1977-1990. *Journal of the American Statistical Association*, 92(440):1256–1267.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data, 2nd ed.* Wiley, New York.

Lynch, S. and Brown, J. (2010). Obtaining multistate life table distributions for highly refined subpopulations from cross-sectional data: A bayesian extension of sullivans method. *Demography*, 47(4):1053–1077.

Rees, P. (1979). Regional Population Project Models and Accounting Methods. *Journal of the Royal Statistical Society. Series A (General)*, 142(2):223–255.

Wilson, T. and Bell, M. (2004). Comparative empirical evaluations of internal migration models in subnational population projections. *Journal of Population Research*, 21(2):127–160.

ABSTRACT

The paper describes a Bayesian framework for carrying out subnational population estimates and projections. The new methods allow data from multiple noisy data sources to efficiently combined. They also generate formal measures of uncertainty. The paper provides some illustrative results from software implementing the methods.