USING BAYESIAN MODELS TO FORECAST INTERNATIONAL ARRIVALS AND DEPARTURES BY AGE, SEX, AND REGION IN NEW ZEALAND

John Bryant and Kirsten Nissen

Statistics New Zealand

Abstract

Statistics New Zealand prepares regular population projections at the subnational level. Like other statistical agencies, Statistics NZ currently uses a deterministic approach to subnational projections. However, we are currently developing an alternative, more statistical, approach. In this paper, we describe two Bayesian hierarchical models for estimating and forecasting international in-migration and out-migration rates, disaggregated by age, sex, and territorial authority. The model uses international arrivals and departures data for 1991-2013, and produces forecasts for 2014-2039. Special attention is given to the ability of the model to estimate migration rates for small population domains. The modelled approach provides a coherent and integrated measure of uncertainty at the detailed demographic level. Examples of estimated and forecasted migration rates are presented. We conclude with a discussion of the challenges and limitations of the data and model, and directions for future developments.

Introduction

Subnational population projections are one of the key demographic outputs produced by statistical agencies, including Statistics New Zealand. Subnational population projections are, for instance, an essential input to the planning of education, health, and infrastructure.

Population projections require future values for fertility, mortality, domestic (internal) migration, and international (external) migration. In this paper we focus on international migration. Compared with most countries, New Zealand has unusually good data on international migration, in the form of the arrival and departure cards filled out by virtually everyone leaving or entering the country. However, relative to the size of the population, New Zealand's international migration flows are also unusually large and volatile.

Statistics NZ produces national-level population projections that are 'stochastic' or 'probabilistic' (Dunstan 2011). These projection s are generated by (i) using statistical models to generate a sample of future fertility, mortality, and migration rates, and (ii) combining these with an initial population to create a large sample of future populations. These samples can be analysed much like standard samples, and provide an intuitive guide to future values, and uncertainty about these values.

Cameron and Poot (2010) have produced stochastic population projections for the Waikato region. However, stochastic population projections for multiple areas within a country have not, to our knowledge, been produced anywhere in the world. Like other statistics agencies, Statistics NZ produces subnational population projections that are deterministic. These projections are generated by creating a small number of stylised scenarios, using a mixture of extrapolation and expert judgment. The heavy reliance on expert judgment means that the projections are resource intensive, and difficult to reproduce. The use of deterministic scenarios, rather than probability distributions, makes the projections difficult to interpret.

One of the reasons for using deterministic methods for subnational projections is that statistical models can break down when the input data are sparse, as they are for demographic events that are disaggregated by age, sex, and geographical area. Generating sensible estimates from sparse data is a problem of small area estimation (Rao 2003). In our modelling, we combine ideas from the small area estimation literature with ideas from the time series literature. In particular, we use hierarchical Bayesian models (Gelman and Hill 2007) and dynamic linear models (Prado and West 2010). Notable examples of Bayesian migration forecasting at the national level include Bijak and Wiśniowsky (2010) and Azose and Raftery (2013).

In this paper we describe models for estimating and forecasting international migration that are currently under development in New Zealand. We describe the data sources and model specifications, and then present some illustrative results. The paper concludes by summarising some areas for future work.

Data and Methods

Data

Our main data are counts of "permanent and long-term" arrivals and departures to and from New Zealand in the period 1991-2013. A permanent and long-term arrival or departure is a border crossing that entails a change in usual residence, as opposed to a business trip or a holiday. The counts are disaggregated by 5-year age-group, sex, region, calendar year, and citizenship. The information on citizenship is not of interest in itself; as discussed below, it is used when imputing for non-response in the region variable.

The region variable records the territorial authority that the passenger is travelling to or from. Territorial authorities are the most important subnational administrative unit in New Zealand. In 2010 there were 73 territorial authorities in the country, giving an average population size of 60,000, though the smallest (the Chatham Islands) had a population of less than 1,000. During 2010, the seven territorial authorities within greater Auckland were amalgamated into a single unit. Since 2010, arrivals and departures for the new amalgamated unit have been coded to "Auckland", with no further geographical detail. In the modelling presented in this paper, we use the post-2010 classification with 67 territorial authorities. In the Discussion we look at possibilities for generating estimates for areas within greater Auckland.

Altogether, 7.5% of individual emigration records have no regional information, either because the respondent did not provide it, or because the response could not be coded. The percent of missing values rose sharply around the year 2000, with an average of 2.2% in the 1990s, and 9.8% in 2000-2013.

In addition to the migration data, we use population estimates disaggregated by age, sex, region, and year, constructed at Statistics New Zealand as a customized tabulation.

Methods

The first task is to impute missing values for the region variable. We use multiple imputation, as described by Rubin (2004). We fit a log-linear model to the data with complete responses on the region variable. The model includes region, age, sex, year, country of citizenship (grouped by world region), and all second and third order interactions between these terms. The citizenship variable is included because it is strongly correlated with non-response and with region within the country. Having fit the model, we use it to predict region responses for the records with missing values. We do three sets of predictions, and aggregate over citizenship, yielding three filled-in datasets with the variables region, age, sex, and year. We go through this process twice: once for arrivals and once for departures.

We estimate separate models for arrivals and departures. Each model has the form

$$y_{asrt} \sim \text{Poisson}(\theta_{asrt} n_{asrt}) \quad (1)$$

$$\log \theta_{asrt} \sim \text{N}(x_{asrt}\beta, \sigma^2). \quad (2)$$

Value y_{asrt} in Equation (1) is a count of permanent and long term arrivals or departures for region r during June year t by people in 5-year age group a and sex s. Value n_{asrt} is the mid-year population. Parameter θ_{asrt} is the underlying immigration or emigration rate. The goal of the modelling is to estimate and forecastθ_{asrt}.

Including n_{arst} in the model for departures is uncontroversial: it is an estimate of person-years of exposure to the risk of emigrating. No such population-at-risk interpretation is possible with the model for arrivals. However, including n_{arst} in the model greatly improves the precision of the estimates. Value θ_{asrt} in the model for arrivals can be interpreted as an 'admission rate' (Rees 1986: 139).

Equation (2) gives the prior model for the (log) rates. Vector β contains an intercept, age, sex, region, and time main effects, an age-sex interaction, an age-region interaction, and a region-time interaction. Vector x_{asrt} is a row from design matrix X. Matrix X is composed of 1s and 0s, and is constructed so that each cell $asrt$ receives the appropriate combination of main effects and interactions.

The intercept and sex terms within β are given non-informative uniform priors; in econometric terminology, they are treated as fixed effects. Region main effects are assumed to be drawn from a common normal distribution, that is,

$$\beta_r^{reg} \sim \text{N}(0, \sigma_{reg}^2).$$

Under this prior, the regions effects are shrunk towards a common mean. This increases precision in return for some bias (Gelman and Hill 2007). Smaller values of σ_{reg}^2 imply greater shrinkage. Parameter σ_{reg} is given a non-informative uniform prior. Age main effects, and the age-sex and age-region interactions, are treated analogously to the region main effects.

The time effect is modelled as a random walk with noise,

$$\beta_t^{time} \sim \mathrm{N}\left(\gamma_t^{time}, \tau_{time}^2\right) \tag{3}$$

$$\gamma_t^{time} \sim \mathrm{N}(\gamma_{t-1}^{time}, \omega_{time}^2) \tag{4}$$

(Prado and West 2010). This specification allows for the possibility of shocks that affect the overall migration level only in year t and for shocks that permanently raise or lower the expected level. The strength of temporary shocks is measured by τ_{time}^2 and the strength of permanent shocks by ω_{time}^2. Both these standard deviation parameters are given non-informative uniform priors.

Region-time interactions are modelled using modified versions of (3)-(4), with one independent time series for each region,

$$\beta_{rt}^{reg:time} \sim \mathrm{N}\left(\gamma_{r,t}^{reg:time}, \tau_{reg:time}^2\right) \tag{5}$$

$$\gamma_{rt}^{reg:time} \sim \mathrm{N}\left(\gamma_{r,t-1}^{reg:time}, \omega_{reg:time}^2\right). \tag{6}$$

In addition to region-time interactions, we experimented with region-age-time interactions, where each region-age combination had its own time series. However, these models were very slow to converge, suggesting that there was not enough information in the data to identify all the terms.

Inference is carried out with Markov chain Monte Carlo (MCMC) methods, using software written in C and R. Gibbs samplers for the immigration and emigration models are run with six independent chains. In the case of the immigration model, the burnin is 25,000 iterations, production is 25,000 iterations, and the thinning ratio is 1:100, yielding 6×25,000/100 = 1,500 iterations for posterior inference. The emigration model is slower to converge, so a burnin of 40,000 iterations, and production of 40,000 iterations is used instead.

The model fitting is carried out once for each of the three filled in datasets, and the resulting posterior samples are pooled. Working with three alternative datasets, rather than a single dataset, means that some of the uncertainty created by the imputation process is reflected in the posterior distribution (Rubin 2004).

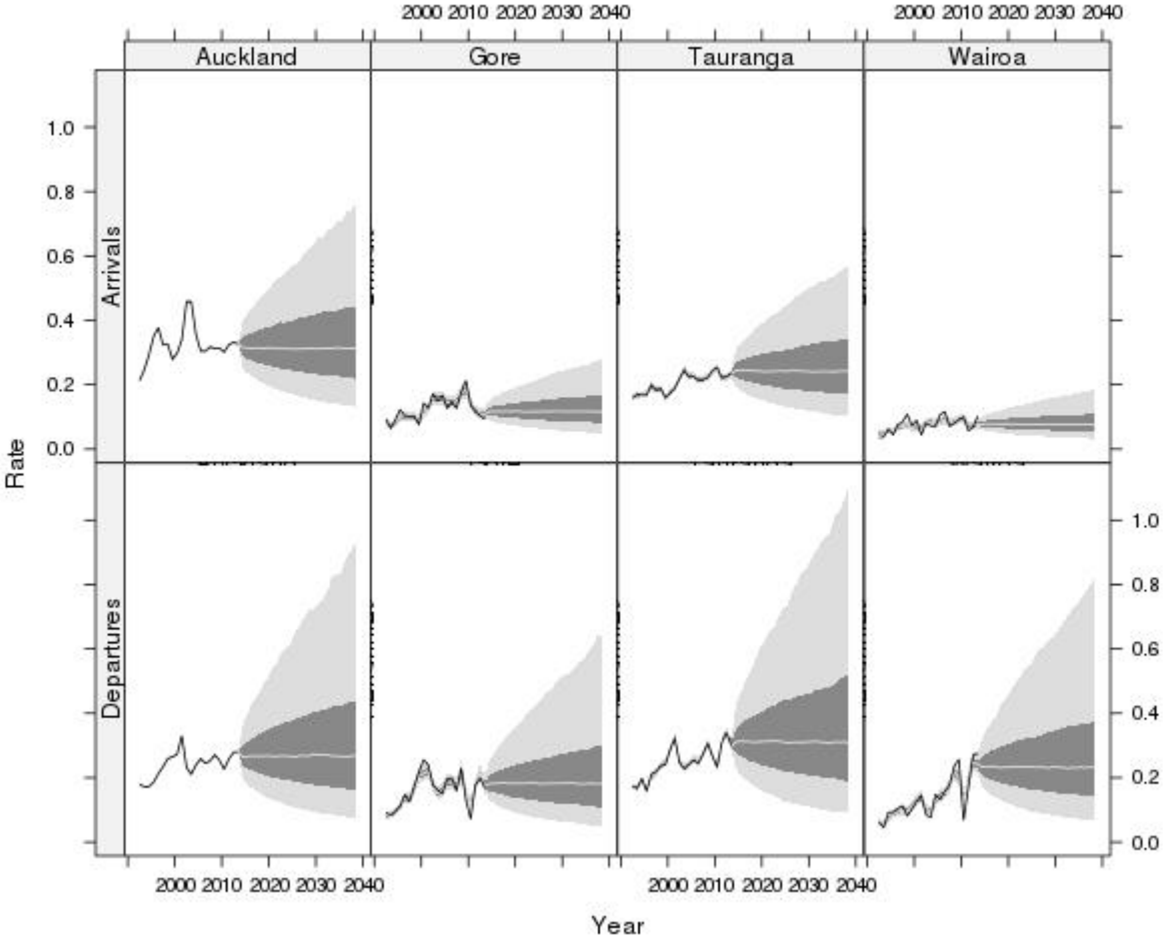
When summarising results from the model we use, among other measures, the 'total arrival rate' and 'total departure rate', defined as $\varphi_{srt} = \sum_a \theta_{asrt}$, where θ_{asrt} is the age-sex-region-time-specific arrival rate or departure rate. The total arrival and departure rates are intended to be simple ways of summarising migration rates that are not affected by changes in age structure, though they can be interpreted, analogously to the total fertility rate, as the average number of arrivals or departures a person would make over his or her lifetime, given prevailing rates.

Results

The hierarchical model produces estimates and forecasts for thousands of parameters. In this section, we present a small sample of the output, to give a sense of the present capabilities and limitations of the model.

Figure 1 shows estimates and forecasts of total arrival and departure rates in four territorial authorities, for females. The largest of these territorial authorities (Auckland) had a population in 2013 of 1.5 million; the smallest (Wairoa) had a population of 8,100. The light grey shaded areas show 95% credible intervals. The 95% credible intervals give a range of values that, under the model, have a 95% chance of containing the true migration rate. The dark grey shaded areas show 50% credible intervals. The light grey lines in the middle show the medians of the posterior distributions. Posterior medians or means are the most commonly-used point estimates in Bayesian analyses. The black lines show direct estimates, that is, the observed number of migrations divided by the mid-year population.

Figure 1 Total arrival rates and total departure rates for females in four selected territorial authorities



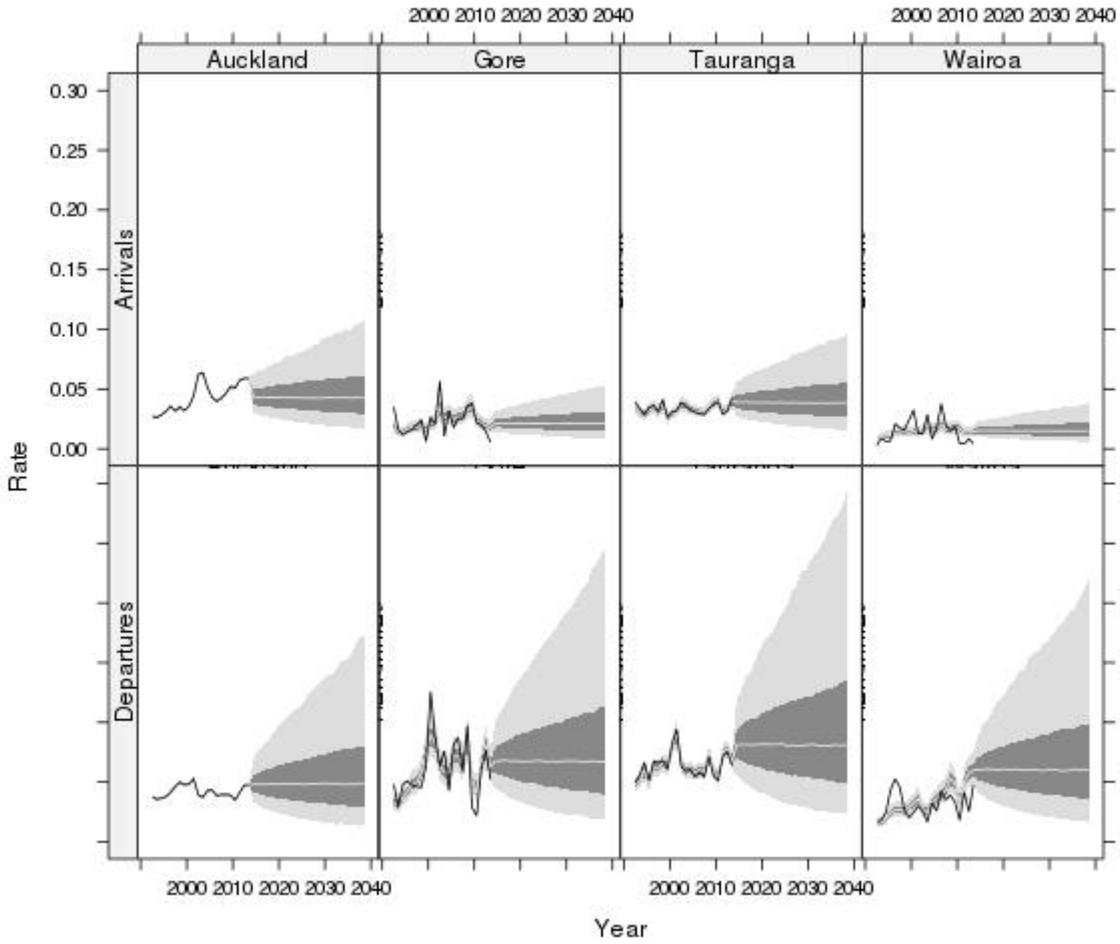
Note – The 'total arrival rate' and 'total departure rate' are defined in the Methods section. The light grey shading represents 95% credible intervals; the dark grey shading represents 50% credible intervals, and the white lines represent posterior medians. The black lines are direct estimates.

The modelled estimates for Auckland, where there are abundant observations, are virtually identical to the direct estimates, with extremely narrow credible intervals. In contrast, the estimates for Wairoa are somewhat smoother than the direct estimates, with visible credible intervals. This is typical behaviour for hierarchical Bayesian models. All modelled estimates are compromises between the direct estimates and the prior model. As sample sizes become smaller, the prior model receives more weight, and, reflecting the extra uncertainty, the credible intervals become wider (Gelman and Hill 2007).

The credible intervals for the forecasts of departure rates are wider than the credible intervals for arrival rates, implying that forecasts of departure rates are more uncertain. The greater uncertainty reflects the greater annual volatility of departure rates. For instance, the posterior median for the ω_{time}^2 term from Equation 6 is almost twice as large in the model for departures as it is in the model for arrivals. Uncertainty grows rapidly during the early years of the forecasts, which is consistent with the existence of large short-term fluctuations in the historical data. Further out into the forecast horizon, uncertainty continues to grow, albeit more slowly, which is sensible behaviour.

One aspect of the model that is working well is the region-time interactions. The evidence for this is the smooth transition between the estimates and forecasts. Without the interaction, each region would jump to its historical mean in the first year of the forecast: for instance, the departure rate for Wairoa would drop to about half its 2013 level. Instead, there is a very slight movement towards the historical mean. This limited reversion to the mean is appropriate, given the annual variability of the series, as measured by the $\tau_{reg:time}^2$ term in (5).

Figure 2 Estimated and forecasted migration rates for females aged 20-24 in four selected territorial authorities

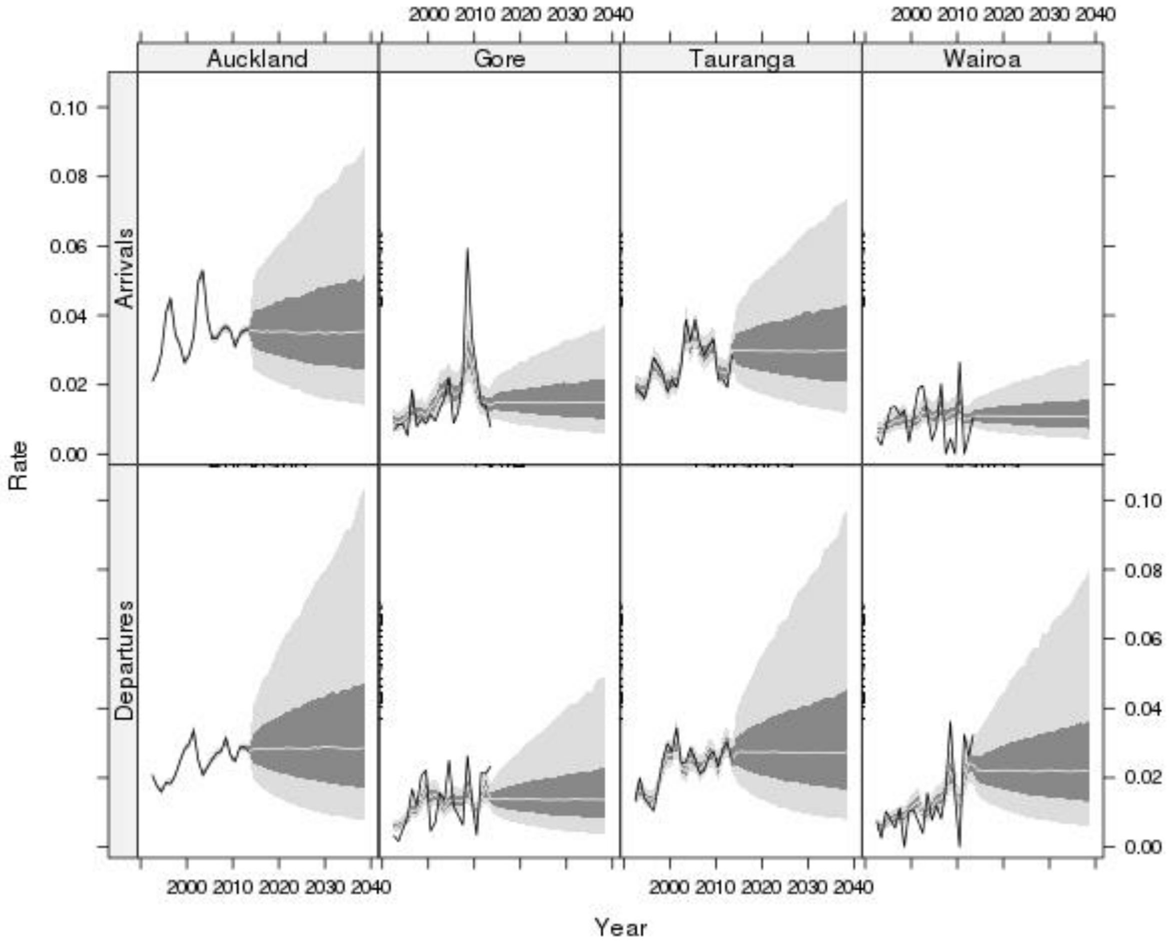


Note – The light grey shading represents 95% credible intervals; the dark grey shading represents 50% credible intervals, and the white lines represent posterior medians. The black lines are direct estimates.

Figure 2 shows arrival and departure rates for females aged 20-24. The number of events is smaller than for the total arrival and departure rates, so the direct estimates are more volatile. The smaller number of events also induces the model to place more weight on the prior, which leads to greater smoothing, and produces wider credible intervals for the estimated rates in Gore and Wairoa. This is again sensible behaviour.

One feature of the model which is less satisfactory is its treatment of arrival rates for Auckland. There is a sudden dip in the median value for the arrival rate in the first year of the projection. The share of 20-24 year olds in overall arrivals has been rising in Auckland, but without a region-age-time interaction, this trend is not incorporated into the forecasts.

Figure 3 Estimated and forecasted arrival and departure rates for males aged 30-34 in four selected territorial authorities

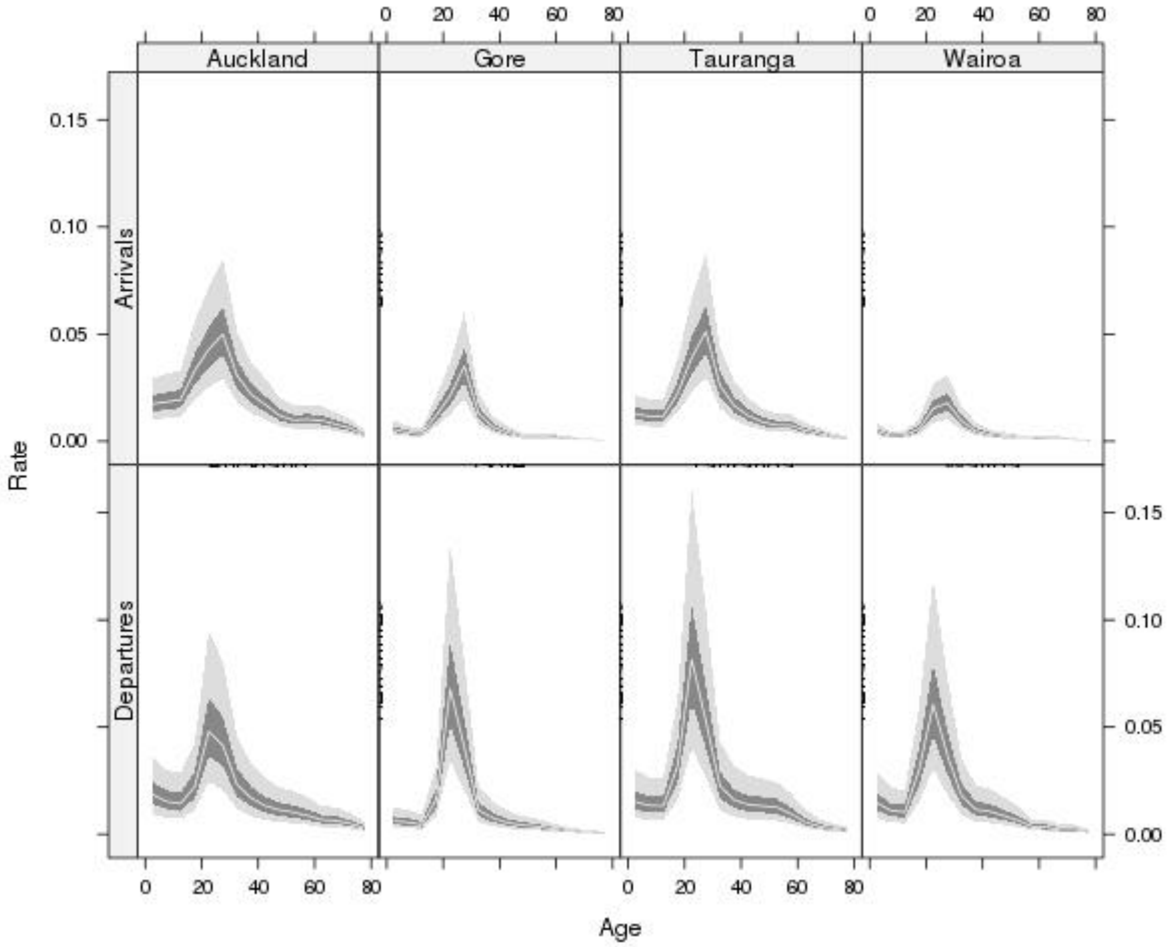


Note – The light grey shading represents 95% credible intervals; the dark grey shading represents 50% credible intervals, and the white lines represent posterior medians. The black lines are direct estimates.

Figure 3 shows the corresponding estimates and forecasts for males aged 30-34. The vertical scale differs from Figure 1, reflecting the lower migration rates at ages 30-34. The results follow a similar pattern to those of Figure 2, though the direct estimates are even more volatile than in Figure 2, due to the even smaller numbers of events.

Finally, Figure 4 shows the whole forecasted age distribution for females in 2020. Because the model does not contain an age-region-time interaction, the future age profiles preserve the historical age profiles. The age profiles differ between arrivals and departures and across the four territorial authorities. For instance, peak age for departures is 20-24, while for arrivals it is 25-29.

Figure 4 Forecasted arrival and departure rates for females in 2020 in four selected territorial authorities



Note – The light grey shading represents 95% credible intervals; the dark grey shading represents 50% credible intervals, and the white lines represent posterior medians.

Discussion Speeding up of comp-utation

Our experience so far suggests that estimating and forecasting external migration at the age-sex-region level through formal statistical models is difficult but possible. We do, nevertheless, have a long list of extensions for future work. Some high priority extensions are as follows.

Speeding up computations. Although the time critical code is written in C, and designed to be efficient, many calculations must still be run overnight, or longer. This places practical constraints on model building that we would like to overcome. Possibilities for speeding up computations include the use of Hamiltonian Monte Carlo methods (Neale 2011), and the use of weakly informative priors (Gelman 2006).

Model for time effect. The prior for the time effect has a strong effect on forecasts. We would like to test our current prior (Equations (3) and (4)) further, and to experiment with alternatives.

Age-region-time interactions. As apparent in Figures 2 and 3, including age-region-time interactions would improve the quality of the forecasts for some regions. It is not computationally feasible to use the flexible priors used for modelling age-region or region-time priors. We plan to investigate the use of more restricted priors.

Joint modelling of arrivals and departures. Arrivals tend to generate departures, and vice versa. Including these effects in a model is likely to lead to improved forecasts of net migration. We have experimented with joint models of arrivals and departures, and intend to continue with this work in future. The main challenges are computational.

Assessing calibration. A model is well calibrated if a 95% credible interval contains the true value 95% of the time, a 50% credible interval contains the true value 50% of the time, and so on. The standard method for assessing the calibration of a forecasting model is to withhold some data from the model and see if the model can predict the withheld values. We intend to run these sorts of tests in future.

At the same time as we improve the models for international migration, we will also be working on models for other components of population change. Our experience so far suggests that fertility and mortality in New Zealand are relatively easy to model, as they have been stable in recent decades or have changed in regular ways. Internal migration has also been surprisingly stable, but is more difficult to model, because it is more complex, and because only a few periods of data are available. We therefore expect that much of our modelling efforts will be concentrated on migration.

References

Abel, G. J., Bijak, J., Forster, J. J., Raymer, J., and Smith, P. W. F. (2010). What Do Bayesian Methods Offer Population Forecasters*?* Centre for Population Change Working Paper 6/2010, Economic & Social Research Council, UK.

Azose, J. J. and Raftery, A. E. (2013). Bayesian Probabilistic Forecasting of International Migration Rates, Department of Statistics, University of Washington (http://arxiv.org/pdf/1310.7148.pdf)

Bijak, J. and Wiśniowsky, 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)*, 173(4): 775–796.

Cameron, M. P., and Poot, J. (2010). *A Stochastic Sub-national Population Methodology with an Application to the Waikato Region of New Zealand* Population Studies Centre Discussion Paper, University of Waikato, NZ.

Dunstan, K (2011). Experimental stochastic population projections for New Zealand: 2009(base)–2011 (Statistics New Zealand Working Paper No 11–01).

Gelman, Andew. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis. 1(3): 515-534.

Neal, Radform M. 2011. MCMC using Hamiltonian dynamics. In Brooks, Steve and Gelman, Andrew and Jones, Galin and Meng, Xiao-Li, Handbook of Markov Chain Monte Carlo. CRC Press.

Prado, Raquel and West, Mike. 2010. Time Series: Modelling, Computation, and Inference. Boca Raton: CRC Press.

Rees, Philip, 1985, “Choices in the construction of regional population projections”, in Population structures and models: developments in spatial demography, edited by Robert Woods and Philip Rees. Boston: George Allen and Unwin.

Rubin, Donald B., 2004, Multiple Imputation for Nonresponse in Surveys, New York: John Wiley and Sons.