

Chapter 10

Bayesian Disaggregated Forecasts: Internal Migration in Iceland



Junni L. Zhang and John Bryant

10.1 Introduction

Ministries of Finance want national-level population forecasts. Almost all other users of population forecasts, from local councils, to market analysts, to planners of roads, supermarkets, and hospitals, want local-level forecasts.

Constructing local-level population forecasts is not easy. The most difficult part is estimating historical trends for demographic rates that can be extrapolated into the future. Fertility, mortality, and migration rates vary across subnational areas in ways that can be difficult to model. The age profiles of migrants coming to university towns, for instance, are dramatically different from the age profiles of migrants coming to rural areas (Wilson 2010). Moreover, the more finely a population is disaggregated, the smaller the number of observations that are available for each combination of classifying variables such as age, sex, and region. Random variation starts to dominate, and the underlying propensities become lost in the noise.

Traditional demographic techniques, which were designed for national-level datasets, are poorly suited to estimation and forecasting with sparse data. The most traditional demographic approach to estimating rates is to simply divide the number of observed events by the population at risk, and to do so separately for each

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-3-030-42472-5_10) contains supplementary material, which is available to authorized users.

J. L. Zhang (✉)
National School of Development, Peking University, Beijing, China
e-mail: junnizhang@163.com

J. Bryant
Bayesian Demography Ltd, Christchurch, New Zealand

© The Author(s) 2020
S. Mazzucco, N. Keilman (eds.), *Developments in Demographic Forecasting*,
The Springer Series on Demographic Methods and Population Analysis 49,
https://doi.org/10.1007/978-3-030-42472-5_10

combination of the classifying variables. When most cells have small numbers of events, however, estimates obtained by considering each cell separately are erratic and unreliable.

In response to these problems, demographers turn to some form of smoothing or modelling. Estimates for each cell are informed by data for neighbouring cells, and perhaps also by information about overall patterns. The classic method for smoothing migration rates, for instance, is model migration schedules (Rogers and Castro 1981). These allow demographers to construct typical age profiles for migration by specifying only a handful of parameters. More recent alternatives include splines, or other types of general-purpose statistical smoothing techniques. A second general approach is to use log-linear models, which provide parsimonious ways of representing the main patterns in the data (van Imhoff et al. 1997; Raymer and Rogers 2007; Rogers et al. 2010).

Demographic estimation and forecasting models based on model life tables, splines, or log-linear models have had many successes. But even these start to break down when cell counts become very small (Bernard and Bell 2015; Baffour and Raymer 2019). Standard log-linear models, for instance, cannot handle cell counts of zero.

As statisticians have long recognized, the ability to extract complex patterns from sparse datasets is a particular strength of Bayesian statistical methods (Gelman et al. 2014). Bayesian methods are, accordingly, becoming increasingly popular among demographers carrying out subnational estimates and forecasts (Lynch and Brown 2010; Schmertmann et al. 2013; Bijak and Bryant 2016; Alexander et al. 2017; Bryant and Zhang 2018). There are, of course, limits to how much can be inferred from any given dataset, even with the best available methods. However, Bayesian analyses also yield detailed measures of uncertainty, which can be used to inform users about these limits.

In this chapter, we present Bayesian forecasts for one particular component of local-level population change: internal migration, i.e., changes of residence within national boundaries. Getting internal migration right is essential to local-level forecasting, as internal migration is typically the biggest source of population change for small geographical units.

To illustrate the ability of Bayesian methods to cope with sparse data, we have chosen an extreme case: Iceland. The population of Iceland in 2018 was 348,450. Once the internal migration data for Iceland are disaggregated by sex, single-year-of-age, 8 regions of origin, 8 regions of destination, and calendar year, 66% of cells have values of zero. Using single years of age and calendar years, rather than, say, aggregating to 5-year units, increases sparsity. However, it reflects user needs. Consumers of population forecasts often want forecasts for particular years, or for age groups such as school ages that cannot be constructed from 5-year age-time blocks.

We begin the chapter with a review of the Icelandic data and migration trends. We then present a baseline model that tries to capture these trends in a parsimonious way. We subject the baseline model to some model checking, using ‘replicate data’ techniques. Based on these checks, we construct a revised, slightly more

complicated model. We use held-back data to choose between the baseline and revised models. We then present forecasts from the best-performing of the two models.

Our recent book *Bayesian Demographic Estimation and Forecasting* (BDEF) (Bryant and Zhang 2018) also includes a chapter on internal migration in Iceland. However, the BDEF model uses confidentialised data, and has a component to account for the confidentialisation process, which is the main focus of that chapter. The BDEF component dealing with demographic rates is also simpler than the one presented here, and is not subjected to model testing or model comparison.

10.2 Data

Our first dataset is counts of internal migrations by region of origin, region of destination, single year of age (up to age 80+), sex, and calendar year. The data were obtained from the Statistics Iceland website.¹ The Statistics Iceland website states that the data come from the Register of Migration Data, and that a person is considered to have moved between regions if the person has stayed in the new region for at least one month. Altogether, the migration dataset has 181,440 cells.

These 181,440 cells do not include ‘structural zeros’, that is, cells where the count is zero by definition. In our case, since our definition of migration requires a change of region, a cell is a structural zero if the region of origin for the cell equals the region of destination. The figure of 66% of cells equalling zero cited above also does not include structural zeros. Among the non-zero cells, the median value is 2, and the maximum is 34.

To provide a feel for the sparsity of the data, Fig. 10.1 shows migration counts for three selected regions for a single year. The age profiles are jagged, and flows not involving the Capital Region are tiny, with most age groups having counts of zero.

In addition to migration counts, we also use a dataset giving resident population counts at 1 January of each year. These counts are disaggregated by region, age, sex, and year. The data were also obtained from the Statistics Iceland website.² The largest region in Iceland, Capital, had a population in 2018 of 222,484, and the smallest, Westfjords, had a population of 6,994.

We divide the data into a training set and a test set. The training set covers the years 1999–2008 and the test set covers the years 2009–2018. As we discuss below, we build our models using the training set, and choose the best model based on performance in the test set, before using the combined training and test sets to construct our final forecasts.

¹Table *Internal migration between regions by sex and age 1986–2017—Division into municipalities as of 1 January 2018*, downloaded on 19 March 2019.

²Table *Population by municipality, age and sex 1998–2018—Division into municipalities as of 1 January 2018*, downloaded on 19 March 2019.

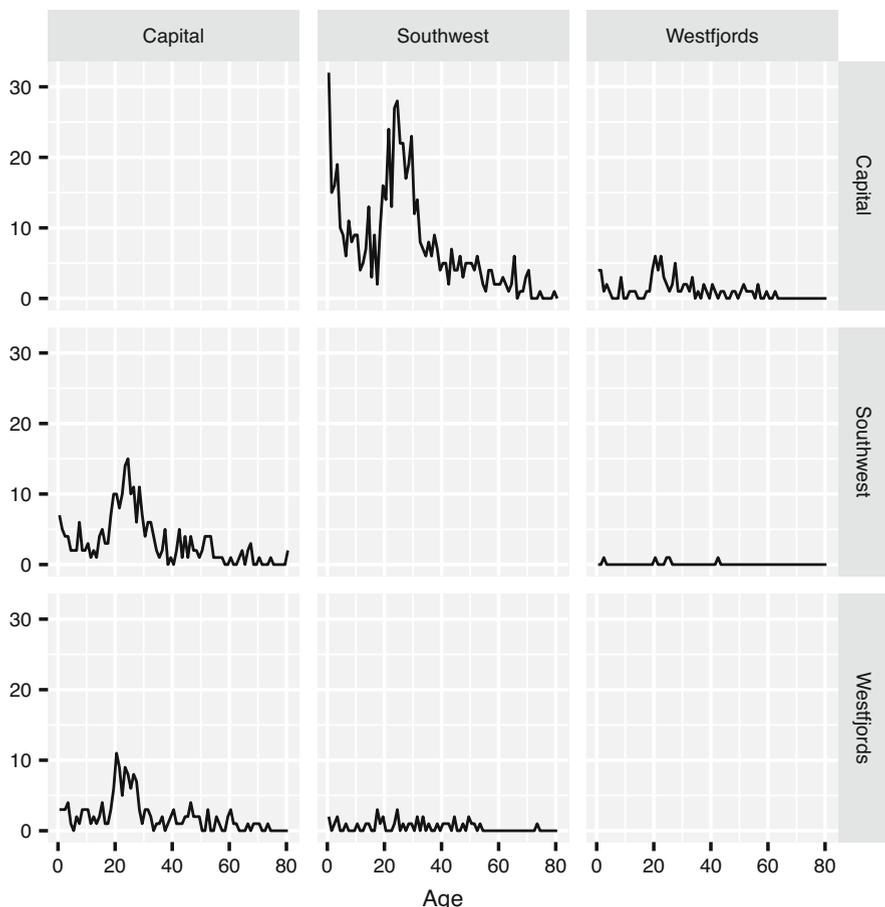


Fig. 10.1 Number of migrations of females in 2008, for three selected regions. Each row shows an origin region and each column shows a destination region: for example, row 2, column 1 shows migration from Southwest to Capital

10.3 Empirical Patterns

We begin by looking a little more closely at the data, starting with regional populations. Figure 10.2 shows regional population counts by age in 2008. Although the age profiles are broadly similar across regions, there are some important differences at the young adult ages. From about age 20, age profiles in most regions bend downwards. In Capital Region, however, the profile bends upwards. Even without seeing the migration data, we might suspect that young people are migrating from other regions into Capital Region.

Figure 10.3 shows direct estimates of migration rates by age, for each combination of origin and destination. We use the term ‘direct estimate’ to mean estimates

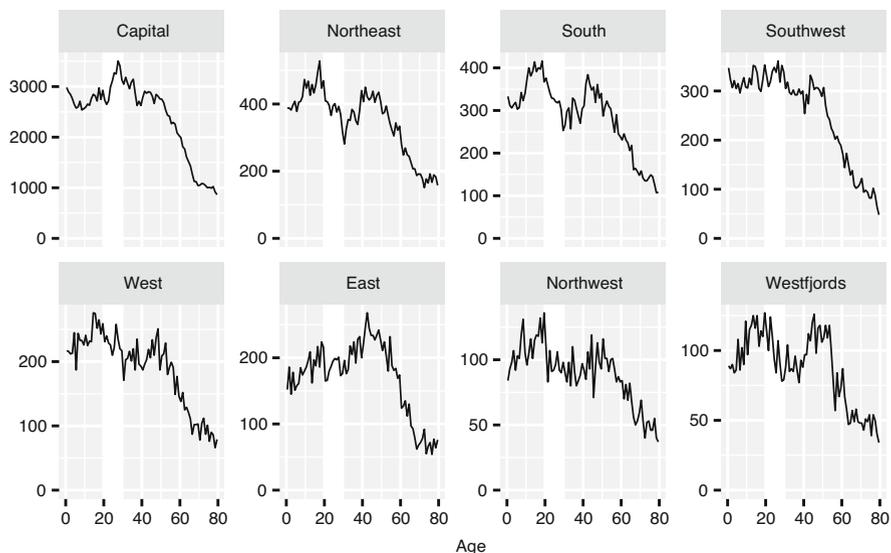


Fig. 10.2 Population aged 0–79, by single year of age and region in 2008. The regions are arranged by population size, from top left to bottom right. Each panel has a different vertical scale. The white vertical strips show ages 20–29

obtained by dividing the number of events in a given cell by the population at risk for that cell, as opposed estimates obtained from a statistical technique that pools information across cells. The estimated rates vary by two orders of magnitude across age and region, so, for clarity, we display them on a log scale. Comparing across columns, we can see that age-specific migration rates for migration into Capital Region have a more pronounced peak at the young adult ages than age-specific rates for migration into other regions. This is consistent with the observation that Capital region has proportionally more young people than other regions.

One sort of difference not readily apparent in Fig. 10.3, however, is sex differences. Females and males in Iceland seem to have very similar migration patterns.

Figure 10.4 displays a different aspect of the data, showing trends in migration between regions, for all age-sex groups combined. Once again, the rates are shown on a log scale. Migration rates into Capital Region, in the first column, are much higher than migration rates into any other region. There are hints of upward or downward trends, most notably for migration between Northwest and East regions, though in many cases it is difficult to be sure because of random variation in the rates.

Finally, Fig. 10.5 gives the age profiles for migration in 1999 and 2008. There appears to have been a slight shift in the age profile between these two years, particularly in the young adult ages.

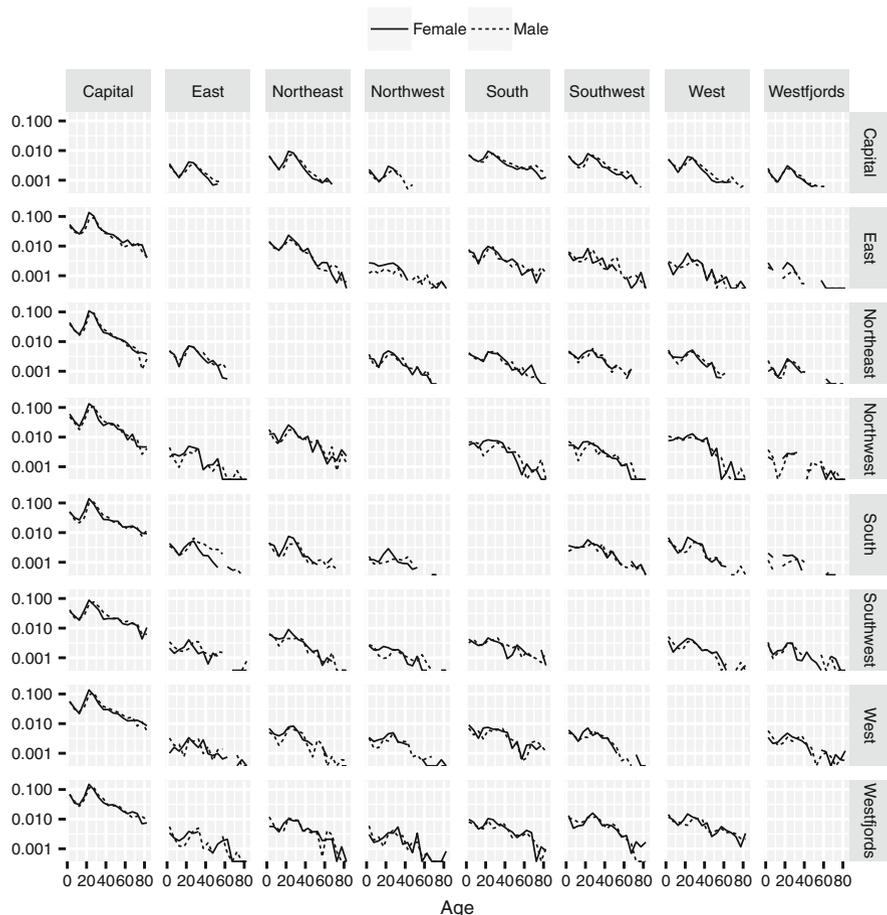


Fig. 10.3 Direct estimates of migration rates, by region of origin, region of destination, age, and sex, 1999–2008. Each row represents one origin region and each column represents one destination region. The rates are shown on a log scale. To reduce variability, the figure uses 5-year age groups, and uses average migration rates over the entire period 1999–2008

10.4 Baseline Model

10.4.1 Counts and Rates

Our baseline model tries to capture the main patterns in the migration data as simply as possible. Let y_{ijast} denote migrations between regions i and j by people in age group a and sex s during period t . As noted above, we define $y_{ijast} \equiv 0$ whenever $i = j$. Let p_{iast} denote the number of people at the start of period t in the population of region i , age group a and sex s . Let w_{iast} denote the number of person-years lived

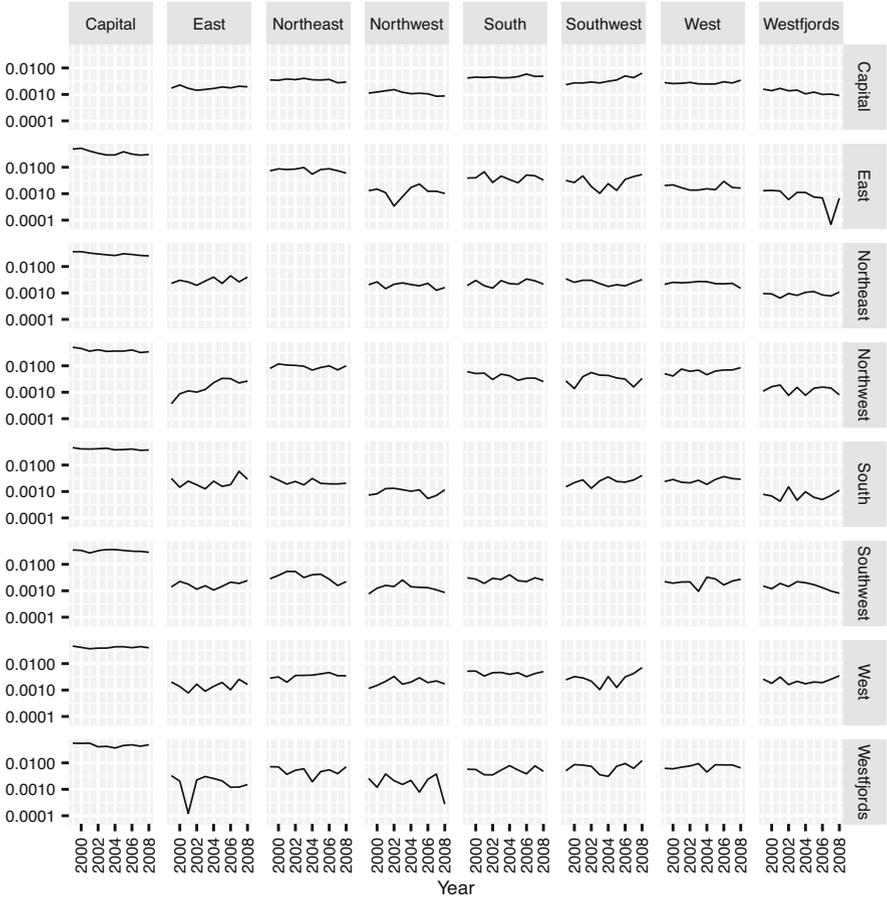


Fig. 10.4 Direct estimates of migration rates, by region of origin, region of destination, and time, 1999–2008. Each row represents one origin region and each column represents one destination region. The rates are shown on a log scale

during period t for the population of region i , age group a and sex s . Demographers commonly approximate the number of person-years lived using

$$\frac{\text{initial population} + \text{final population}}{2} \times \text{length of period},$$

which gives $w_{iast} = (p_{iast} + p_{i,a,s,t+1})/2$. We assume that, within each cell, migration counts follow a Poisson distribution,

$$y_{ijast} \sim \text{Poisson}(\gamma_{ijast} w_{iast}), \tag{10.1}$$

where γ_{ijast} is the underlying migration rate.

Fig. 10.5 Direct estimates of migration rates, by age, in 1999 and 2008

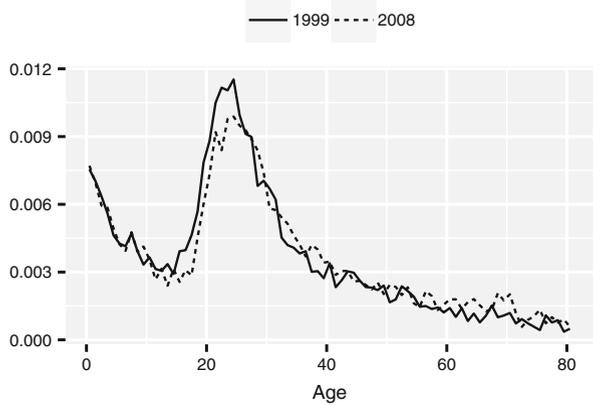


Table 10.1 Priors for main effects and interactions, baseline model

Term	Prior
(Intercept)	Exchangeable with known variance
region_orig	Exchangeable with covariates
region_dest	Exchangeable with covariates
age	Local trend
sex	Exchangeable with known variance
time	Local level
region_orig:region_dest	Exchangeable
region_dest:age	Exchangeable with covariates
age:time	Local level

Equation (10.1) allows for the fact that, for a given migration rate and exposure, the actual number of migrations is a random quantity. Standard log-linear models have no equivalent to Equation (10.1). This omission does not matter when cell counts are large, and variation due to the randomness of individual events is minor relative to variation due to differences in rates and exposures. However, ignoring random variation becomes problematic when cell counts are small. One consequence is the inability of such models to deal with cell counts of zero.

The migration rates γ_{ijast} are modelled using

$$\log \gamma_{ijast} \sim N(x_{ijast}\beta, \sigma^2). \tag{10.2}$$

Vector β contains a combination of main effects and interactions, which are listed in Table 10.1. Vector x_{ijast} , which is composed of 0s and 1s, assigns the appropriate elements of β to each value for γ_{ijast} .

A main effect is a predicted difference for one variable that remains constant across values for all the remaining variables. In our model, for instance, a sex main effect is a female-male difference that remains constant across all possible combinations of region, age, and time. An interaction is a predicted difference that

varies across values for one or more other variables. The age-destination interaction in our model, for instance, measures the way that migration age profiles vary across regions of destination.

An important feature of Equation (10.2) is that $x_{ijast}\beta$, the value for cell $ijast$ assembled from the various elements of β , is the *expected* value for $\log \gamma_{ijast}$, not the *actual* value. The fact that Equation (10.2) uses a probability distribution implies that actual values differ in general from expected values. The typical size of the difference between actual and expected values is governed by the parameter σ . The smaller the value of σ , the tighter the fit. The parameter σ is estimated as part of the overall model-fitting process.

In models like that of Equations (10.1)–(10.2), the final estimate for each γ_{ijast} is a compromise between the predicted value calculated from $x_{ijast}\beta$ and the direct estimate calculated from y_{ijast} and w_{iast} . All else equal, the more observations there are for cell $ijast$, that is, the higher the values of y_{ijast} and w_{iast} , the closer the final value will be to the direct estimate. Models like that of Equations (10.1)–(10.2) perform a sort of local smoothing. Estimates are pulled towards the model predictions in cells where counts are small, but are left more-or-less unchanged in cells where counts are large. This is a sensible and effective way to smooth.

Effective smoothing is essential to demographic forecasting. A good forecast is one that carries forward into the future genuine, long-lasting features of the demographic series, and leaves out transient features or random noise.

10.4.2 Priors

Each main effect and interaction in β is given a prior distribution. In a Bayesian analysis, a prior distribution is a way of representing information about the system being modelled, beyond what is contained in the main datasets (Bryant and Zhang 2018, pp. 88–92). In our case, prior distributions allow us to encode some qualitative features of migration rates, beyond what is contained in the y_{ijast} and w_{iast} .

The prior for the sex effect β_s^{sex} , for instance, is

$$\beta_s^{\text{sex}} \sim N(0, 1). \quad (10.3)$$

This prior implies that, on a log scale, we expect female-male differences to be values like 0.1, -0.5 , or 1.1, but not values like -18 or 400. This prior understates our actual knowledge. A difference of 0.1 on a log scale corresponds to a difference of about 10% on the original scale, which is about as large as we would expect to see for sex differences in Icelandic migration rates. In Bayesian terminology, our prior for the sex effect is ‘weakly informative’. It places a constraint on the range of values that a parameter can take, but only a soft constraint. However, even a soft constraint can greatly speed up computations, and help the model distinguish between random noise and genuine differences.

The priors for the region-of-origin effect, region-of-destination effect, and origin-destination interaction all have the same basic form as the prior for sex. In the case of the region priors, however, the standard deviation parameter is estimated from the data rather than specified in advance. Values for two sexes do not provide enough information to estimate a standard deviation, but values for eight regions do. In addition, the priors for origin and destination include two covariates. The first covariate takes a value of 1 if the region is Capital, and 0 otherwise. The second covariate equals the log of population counts in 2008. By including these covariates, we are allowing for the fact that the Capital region is not like the other regions of Iceland, and that, as emphasised by gravity models of migration (Anderson 2011), migration rates tend to vary systematically with the population size of the origin and destination regions. In principle, we could refine the predictions by allowing the covariate to change over time, as regional population changed. However, this would greatly complicate the forecasting process, and regional population sizes are in any case relatively stable.

The time effect has a local level model (Prado and West 2010, ch. 4),

$$\beta_t^{\text{time}} \sim N(\alpha_t^{\text{time}}, \tau_{\text{time}}^2) \quad (10.4)$$

$$\alpha_t^{\text{time}} \sim N(\alpha_{t-1}^{\text{time}}, \omega_{\text{time}}^2). \quad (10.5)$$

A local level model is a generalisation of a random walk. Like a random walk, it allows for random shifts in the long-term mean of the series, but unlike a random walk, it also allows for one-off departures from this mean. The size of the long-term shifts is governed by ω_{time} , and the size of the one-off departures is governed by τ_{time} . The ω_{time} and τ_{time} parameters are both estimated from the data.

By using a local level model, we are ruling out the possibility of a long-term upward or downward trend in overall migration rates. This assumption is based on inspection of the Iceland data, as shown, for instance, in Fig. 10.4.

Age effects are modelled using a local trend model (Prado and West 2010, ch. 4),

$$\beta_a^{\text{age}} \sim N(\alpha_a^{\text{age}}, \tau_{\text{age}}^2) \quad (10.6)$$

$$\alpha_t^{\text{age}} \sim N(\alpha_{t-1}^{\text{age}} + \delta_{t-1}^{\text{age}}, \omega_{\text{age}}^2) \quad (10.7)$$

$$\delta_t^{\text{age}} \sim N(\delta_{t-1}^{\text{age}}, \varsigma_{\text{age}}^2). \quad (10.8)$$

A local trend model, through the parameter δ , allows for a persistent upward or downward trend. However, because the δ can vary, the size and direction of the trend can change. A local trend model thus allows for the fact that migration age profiles bend upwards through the teens and early twenties, and downwards after that.

Applying time-series models to age effects is an long-standing practice in statistical demography (e.g. Alho and Spencer (2005, pp. 281–282) or Congdon

(2008)). Time series models are based on the principle that neighbouring units are more highly correlated than distant units, an idea which is just as valid for age groups as it is for time periods.

The prior for the age-destination interaction has the same structure as the origin-destination interaction, in that it uses a normal distribution with a standard deviation that is estimated from the data. The prior also includes a covariate, the log of the 2008 population in each combination of age and destination. The prior for the age-time interaction uses a separate local level model for each age group, sharing the same $\tau_{\text{age:time}}$ and $\omega_{\text{age:time}}$ across age groups.

All standard deviation parameters that are not specified in advance are given priors constructed from half- t distributions. Half- t distributions are restricted to non-negative values, and favour values near 0. In all cases, we use distributions with 7 degrees of freedom. In our experience, results are generally insensitive to the exact choice of degrees of freedom, but a value of 7 provides a good tradeoff between robustness and speed of convergence. (See Sect. 10.4.4 for a discussion of model convergence.) We use scale parameters of 1 for σ and the main effects, and 0.5 for interactions. In doing so, we are implying that we expect interactions to be smaller than main effects (Gelman et al. 2008). All the priors for the standard deviations are, nevertheless, relatively weak. The *Prior Choice Recommendations* page³ on the website for the Bayesian modelling language *Stan* discusses the advantages and disadvantages of the half- t prior and other priors.

10.4.3 Model Output

As with most Bayesian analyses, the output from the modelling is a sample from the posterior distribution for the unknown quantities. In our case, the unknown quantities are the γ_{ijast} , the standard deviation σ , the main effects and interactions, that is, β^{time} , $\beta^{\text{age:time}}$, and so on, and the parameters for each of the priors distributions.

We can use summaries of the posterior sample to describe the posterior distribution, in much the same way that a survey statistician uses summaries of a sample survey to describe the population. Thus if sample values for a particular rate are 0.0021, 0.0032, ..., 0.0019, and if the 50%, 2.5%, and 97.5% quantiles for these values are 0.0025, 0.0018, and 0.0030, then we can use 0.0025 as a point estimate for the rate and (0.0018, 0.0030) as a 95% 'credible interval'. Under the assumptions of the model, a 95% credible interval for a parameter has a 95% probability of containing the true value for that parameter.

³<https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>

10.4.4 Calculations

Estimates for the parameters in the model are obtained using computational methods known as Markov chain Monte Carlo (MCMC) (Gelman et al. 2014). Essentially, we start with an approximate answer, and then use a Gibbs sampler (Gelman et al. 2014, ch. 11) to cycle through the following steps:

- Draw values for the migration rates γ_{ijast} , conditional on y_{ijast} , w_{iast} , and the current values for all parameters other than the γ_{ijast} .
- Draw the main effects and interactions β , conditional on the γ_{ijast} , and all other parameters.
- Draw values for the remaining parameters, conditional on the γ_{ijast} and β .

The output from this process is a series of draws from the posterior distribution.

The techniques used to draw values for each set of parameters vary according to the conditional distribution of those parameters. Values for β , for instance, are drawn straight from normal distributions. Values for the γ_{ijast} , in contrast, are obtained through a Metropolis-Hastings step, in which new values are proposed and then accepted with probabilities that depend on the proposal distribution and on the posterior probabilities of the current and proposed values (Gelman et al. 2014, ch. 11). Values for standard deviation terms are drawn using a technique called slice sampling (Neal 2003).

We use multiple sets of starting values, and construct an independent chain starting from each set. Using multiple chains in this way can allow generation of more draws for the same amount of time, since the chains can be run in parallel on a multicore computer. It also provides a way of seeing whether the calculations are working as intended. If all is well, the chains should all converge to the same distribution of values. Depending on the quality of the initial approximate answers, it may take some time before this convergence occurs. Values generated during this initial burn-in period are discarded. Non-convergence across the chains can be detected using a statistic generally referred to as ‘R-hat’ (Gelman et al. 2014, p. 285). A value for an R-hat much above 1 indicates non-convergence.

In a model with as many parameters as ours, it is not feasible to calculate R-hats for all parameters. Instead, when a vector of parameters has more than 25 elements, we sample 25 elements and calculate R-hats only for those. We consider the model to have converged when the maximum of all observed R-hats is less than 1.1. By this point, R-hats for most of the cells we are monitoring are usually indistinguishable from 1.

For each model, we use 4 independent chains, each with a burnin of 15,000 iterations, and production of 15,000. We retain 1 out of every 60 iterations, yielding a sample of $4 \times 15,000 \div 60 = 1,000$ draws from the posterior distribution.

The calculations are all carried out using our own open source R packages **dembase** and **demest**. The R packages make use of C code for the most computationally-intensive part of the estimation. The packages can be downloaded from github.com/statisticsnz/R. All code for the Iceland migration example is available at: github.com/bayesiandemography/iceland_migration.

10.5 Model Checking Using Replicate Data

While building a model, we inevitably make many simplifications. Before we can trust the output from the model, we need to verify that, despite these simplifications, the model is still able to capture the substantively important features of the data. One effective way to check for important omissions in a model is to generate replicate data (Gelman et al. 2014, ch. 6). We illustrate with the example of regional time trends.

Our baseline model has a single, shared time trend. In other words, all region-to-region flows are assumed to shift upwards or downwards by the same percentage from year to year. If this assumption is too strong, it could materially affect forecasted values for future migration flows, which is an outcome of central importance to users of the migration forecasts.

Some region-to-region variation in time trends is indeed visible in Fig. 10.4. But, given the small numbers of observations, it is possible that these variations are random noise, and that the data are in fact compatible with the assumption of a single time trend.

To assess the compatibility of the data and the assumption of single time trend, we generate 19 synthetic or ‘replicate’ datasets, using our baseline model. We then compare the one actual dataset with the 19 replicate ones, to see if the actual dataset looks distinctive or out-of-place. If it does, we conclude that the single time trend assumption is too strong.

We generate a replicate dataset by randomly selecting a draw from the posterior sample, plugging the γ_{ijast} from that draw into Equation (10.1), and obtaining a set of simulated y_{ijast} . Repeating this process 19 times yields 19 replicate datasets. We could then, in principle, make 19 new versions of Fig. 10.4 and compare these with the original Fig. 10.4. Instead, we work with summary values. We fit a straight line to each of the $8 \times 7 = 56$ time series of origin-destination migration rates—in other words, to time series like those shown in each panel of Fig. 10.4. We then see whether the distribution of these slopes is similar across the actual and replicate datasets.

Figure 10.6 shows the results from these calculations. The actual dataset is clearly different from the replicate datasets. The baseline model fails to reproduce the observed variability in regional time trends.

10.6 Revised Model

In response to the results from the replicate data test, we construct a revised version of our model that, in addition to all the terms in the baseline model, includes an interaction between origin and time, and an interaction between destination and time. The priors for these interactions have the same structure as the age-time interaction in the baseline model. Each region has its own local level model, but standard deviation terms are shared across regions.

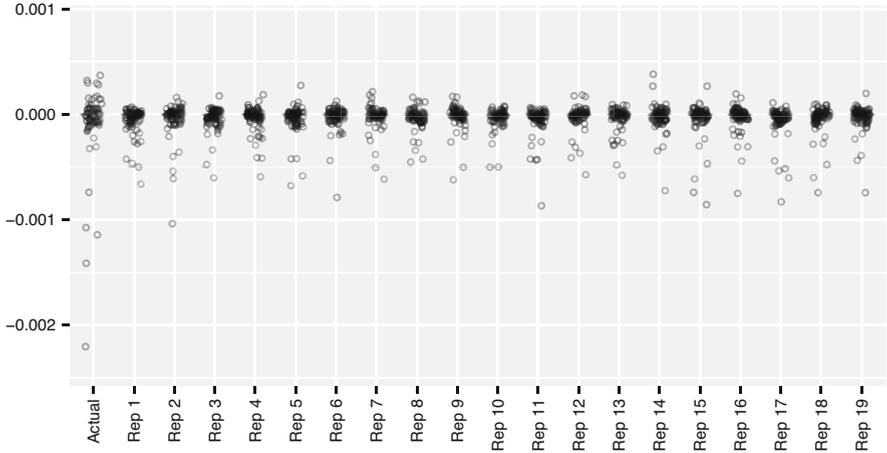


Fig. 10.6 Results of model checking for the baseline model. Using replicate data to test the ability of the baseline model to describe regional time series. Each point shows the slope from a straight line fitted to a time series for migration between a particular origin and destination. There are 56 points in each set. The first set of slopes are obtained from the actual model, and the remaining 19 are obtained from replicate datasets generated from the baseline model

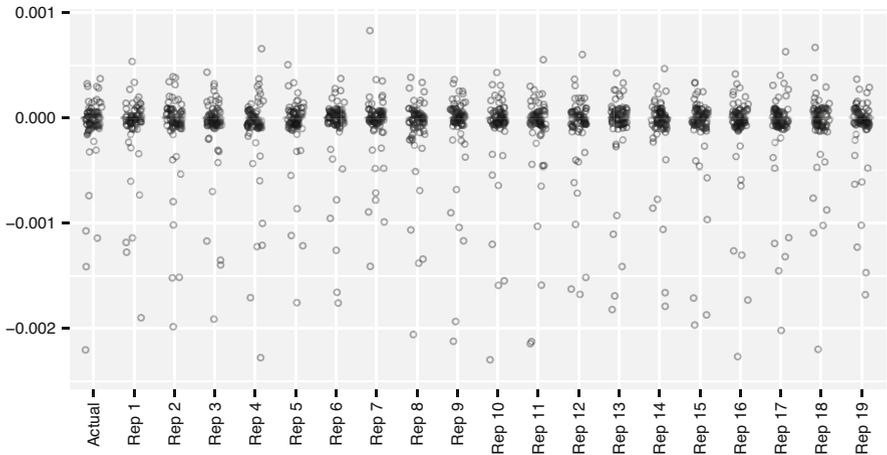


Fig. 10.7 Results of model checking for the revised model. An updated version of Fig. 10.6, using replicate data generated from the revised model rather than the baseline model

Figure 10.7 shows the results from applying the replicate data test to the revised model. The revised model performs much better than the baseline model. The distribution of slopes from the actual data is indistinguishable from the distributions generated under the replicate datasets.

In a full-scale analysis, we would repeat the test-and-revise process a few more times. For instance, we might use replicate data to test whether the data were

consistent with the assumption of no overall trend upwards or downwards. If it turned out that the assumption was clearly violated, then we would extend the model accordingly.

10.7 Forecasts

Our forecasts use exactly the same set of assumptions as our estimates. Indeed, from a Bayesian point of view, there is no sharp distinction between forecasting and estimation. Forecasting is just estimation with missing data (Bryant and Zhang 2018).

We construct the forecasts by extending forward in time each draw from the posterior sample. With the baseline model, the process for extending the s th draw is as follows.

1. Plug values $\tau_{\text{time}}^{(s)}$ and $\omega_{\text{time}}^{(s)}$ into Equations (10.4) and (10.5), and then apply the equations iteratively to the end of the forecast period. This yields a forecasted set of time effects.
2. Plug values $\tau_{\text{age:time}}^{(s)}$ and $\omega_{\text{age:time}}^{(s)}$ into the prior model for age-time interactions, and iterate to obtain forecasted series of age-time interactions.
3. Plug the forecasted time effects, the forecasted age-time interactions, the non-time-varying elements of $\beta^{(s)}$, and $\sigma^{(s)}$ into Equation (10.2). Use Equation (10.2) to generate future values for the γ_{ijast} .

Carrying out these steps for $s = 1, \dots, S$ yields a posterior distribution for migration rates for future years, which can be summarised and manipulated just like any other posterior distribution. Because the forecasts use the same sample of parameter values as the estimates, all the parameter uncertainty in the estimates propagates through into the forecasts.

10.8 Model Choice Using Held-Back Data

We have two models: a baseline model that does not include region-time interactions, and a revised model that does. At first sight, it might seem obvious that we should use our revised model for forecasting, since the replicate data checks imply that region-time interactions are needed to accurately reproduce the historical data. However, while replicate data checks can suggest directions for model improvement, they cannot provide definitive answers on which models will yield the best forecasts. Complex models that do a better job of explaining historical trends do not necessarily do a better job of predicting future values (Shmueli 2010). We use tests based on held-back data to make the final decision on which model to use.

Model choice using held-back data proceeds as follows:

1. Split the data into a training set and a test set.
2. Use the training set to make forecasts about values in the test set—one forecast for each model.
3. By comparing the forecasted values for the test set with the actual values, evaluate the performance of the models.
4. Based on the comparisons, choose a best model.

As noted above, our training data set consists of data for the years 1999–2008, and the test set consists of data for the years 2009–2018.

As well as providing a way of choosing a model, held-back data tests also give a sense of how the models will perform in practice. For instance, if, when measured against the test dataset, 80% credible intervals from a model only contain the true values only 50% of the time, then we would expect that the model to be overly optimistic in other settings as well.

The test data yields direct estimates of migration rates. We must be careful that the forecasted rates from our model are comparable to the direct estimates, in that they also reflect the randomness of the individual events. To do this, we take the forecasted γ_{ijast} , plug them into Equation (10.1), and use Poisson draws to obtain forecasted migration counts. Dividing the forecasted migration counts by exposures gives us the rates that we need.

Our first performance measure is median absolute error. This measure is constructed from the absolute differences between point forecasts and actual value from the test dataset. We obtain point forecasts by taking the medians of the posterior samples of the rates. The second measure is the proportion of values from the test dataset that lie within the credible intervals. We use 80% credible intervals for performance measurement, so ideally 80% or more of the test values should lie within our intervals. The third measure is the median width of the credible intervals: for the same coverage level, the narrower the intervals the better. We take medians of the absolute errors and of the intervals widths, rather than means, because both measures are highly skewed, with many small values and a few large values.

Ideally, we would like to make our comparisons at the lowest level of aggregation, that is, to compare forecasted rates classified by origin, destination, age, sex, and time with test-set rates classified in the same way. Unfortunately, with such sparse data, it is difficult to form credible intervals with the required degree of coverage, since a difference in migration counts of 1 or 2 can imply very large differences in coverage. We instead work with rates classified only by origin, destination, and time, which are considerably less lumpy.

When assessing the performance of the models, we do, however, distinguish between flows out of Capital Region and flows out of other regions. The population at risk of migration is so much larger for Capital Region than for other regions that the job of estimating and predicting migration is much easier. We would therefore expect model performance to differ between Capital Region and elsewhere.

Table 10.2 summarises the performance of the two models. The baseline and revised models have similar levels of accuracy, as measured by median absolute

Table 10.2 Comparison of performance of baseline and revised models, using 80 percent credible intervals

	Median abs. error	Median width	Coverage
Baseline: Capital	0.00044	0.00069	0.23
Baseline: Other	0.00075	0.00133	0.47
Revised: Capital	0.00056	0.00186	0.71
Revised: Other	0.00083	0.00238	0.73

error. Credible intervals from the baseline model are much narrower than credible intervals from the revised model. However, as can be seen in the third column of Table 10.2, the credible intervals from the baseline model are *too* narrow: they contain the true value far less than 80% of the time. The credible intervals from the revised model are much better calibrated, though not perfectly so.

Both models give more accurate predictions for flows from the Capital Region than for flows from other regions. This is not surprising: predictions for the Capital Region are based on more observations than the predictions for the other regions.

Forecasts from the revised model are less accurate than forecasts from the baseline model. However, the revised model is much better calibrated than the baseline model in that its actual coverage rate comes much closer to the nominal rate. We therefore base our forecasts on the revised model.

10.9 Estimates and Forecasts from the Revised Model

We look now at estimates and forecasts from the revised model. The estimates and forecasts are all based on data for the entire period 1999–2018. Figure 10.8 shows estimates of migration rates γ_{ijast} for females in 2018. As well as the modelled estimates, the figure also shows direct estimates, though, unlike the modelled estimates, the direct estimates are aggregated to 5-year age groups, to reduce variability.

Showing the direct estimates alongside the modelled estimates in Fig. 10.8 is a form of reality check on the modelled estimates. If the direct estimates departed in some systematic way from the modelled estimates, then we would suspect that the model had missed out an important feature of the data.

We should not, however, expect 95% of the direct estimates to lie within the 95% credible intervals for the γ_{ijast} . The direct estimates contain all of the original random variability in y_{ijast} . The model tries, as much as possible, to strip away this random variability.

Figure 10.8 illustrates the effects of the smoothing process discussed in Sect. 10.4.1, whereby the modelled estimates stay close to the direct estimates for flows involving Capital Region, where data are plentiful, and rely on predicted values from Equation 10.2 for other flows where data are scarce. This is typical behaviour for Bayesian hierarchical models. To obtain a sensible estimate of

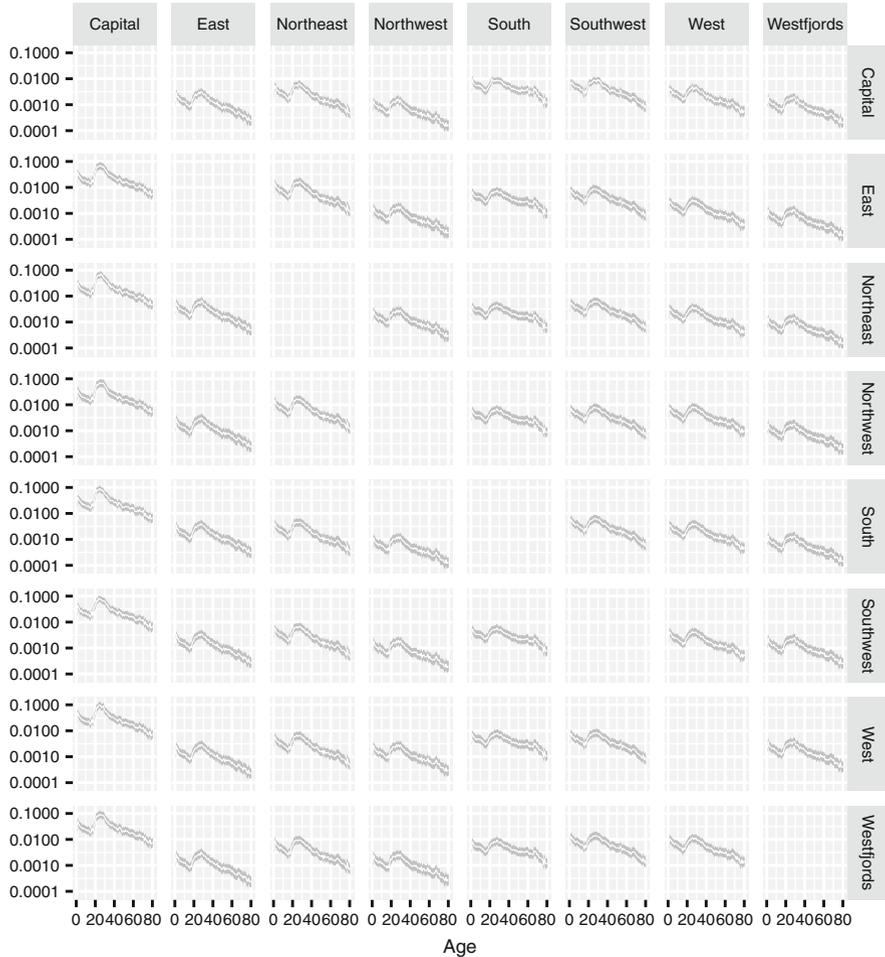


Fig. 10.8 Modelled and direct estimates of migration rates γ_{ijast} , by region of origin (rows), region of destination (columns), and age, for females in 2018. The modelled estimates come from the revised model, using data from the combined training and test datasets. The estimates are shown on a log scale. The grey bands represent 95% credible intervals and the white lines represent posterior medians, for single years of age. The dots represent direct estimates for 5-year age groups. The black dots represent estimates greater than 0, and the grey dots at the bottom of each panel represent estimates equal to 0, which are undefined on a log scale. As discussed in the text, we would not expect 95% of the direct estimates to lie within the 95% credible intervals for the γ_{ijast}

γ_{ijast} for each cell, the model not only uses information coming from the direct estimate for that cell, but also borrows information from all other cells. When data are plentiful such that the direct estimate is reliable, information from the direct estimate outweighs information from other cells. When data are scarce such that the direct estimate is unreliable, information from other cells receives a larger

weight. For instance, in the panel for flows from East to Northwest, the direct estimate is nonzero for age group 25–29, and is zero for all other age groups. It is highly unlikely that the true underlying migration rates follow such an extreme age profile. The rates estimated by combining the East-Northwest data with information borrowed from other cells are much more plausible.

As can be seen by comparing across columns, the age profiles for modelled migration rates differ across destinations. The profile for Capital Region has a sharper peak at the young adult ages than the profile for East Region, for instance, which in turn has sharper peak than South Region. These differences would be difficult to see using direct estimates alone.

The overall level of migration also differs substantially from flow to flow, though this is partly obscured by the use of a log scale.

Figure 10.9 shows estimates and forecasts of migration rates into Capital region for females in selected single-year age groups. As is apparent in the figure, there is substantial uncertainty about underlying migration rates for young adults, even for years where data are available. Uncertainty does, nevertheless, grow further out into the forecast period.

Although, within each age group, migration rates are similar across regions, there are nevertheless differences. Migration rates appear to be higher for young adults from Westfjords, for instance, than they are for young adults from the Northeast.

With Fig. 10.10 we shift from the largest region of Iceland to the smallest. The vertical scale for Fig. 10.10 covers a much smaller range than the vertical scale for Fig. 10.9. People are much less likely to migrate to Westfjords Region than they are to Capital Region.

The data available for directly measuring migration into Westfjords are accordingly very limited. Between 1999 and 2018, for instance, there was not a single case of a 10-year-old migrating from Northwest Region to Westfjords. The model, nevertheless, yields estimates and forecasts that are intuitively reasonable. It implies, for instance, that underlying propensity for 10-year-olds in Northwest Region to migrate to Westfjords has been low, and will continue to be low, but is not zero. The model also virtually ignores the apparent spikes in migration rates suggested by the direct estimates. The model's behaviour in such cases is sensible, given the small counts that give rise to these spikes.

Switching from the training dataset for 1999–2008 to the full dataset for 1999–2018 produces only small differences in estimates for the same years. Figure 10.11 shows some representative examples. There do not appear to have been any major shifts in migration trends between the training period and the test period.

10.10 Discussion

It is still common in demography departments and statistics agencies to encounter rules of thumb stating that demographic rates cannot be calculated unless every cell in a table has, say, at least 5 observations, or at least 30 observations. In this

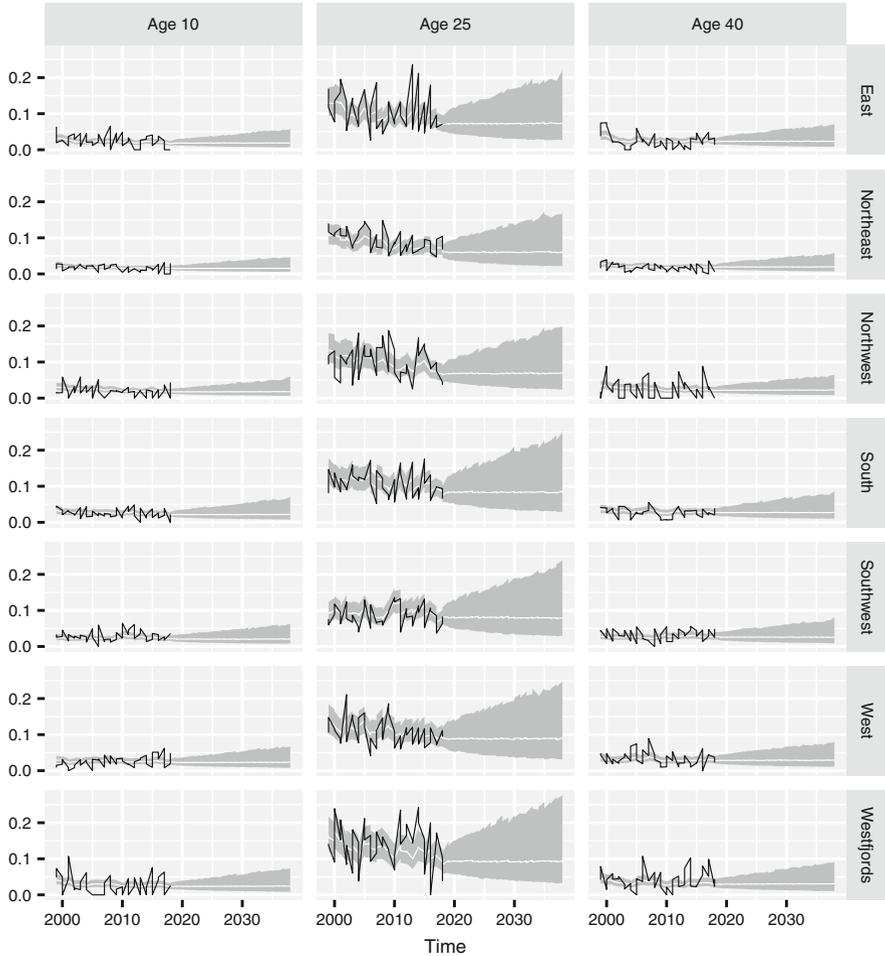


Fig. 10.9 Migration rates to Capital Region from other regions, for females in selected single-year age groups. The grey bands represent 95% credible intervals and the white lines represent posterior medians. The black lines represent direct estimates

chapter, we have broken all such rules. Of the 181,440 cells in our migration dataset, only 11,298 have 5 or more observations, and only 9 have 30 or more. And yet, while there is scope for further checking and refinement, the held-back data tests suggest that our revised model is already attaining respectable levels of accuracy and coverage. Moreover, using credible intervals or other uncertainty measures, consumers of the forecasts can be given guidance on how much trust to place in the rates, including rates calculated from small counts.

The availability of new methods for estimating and forecasting with sparse, complicated datasets, such as the methods we present in this paper, should prompt demographers and statisticians to rethink conventional rules of thumb about what

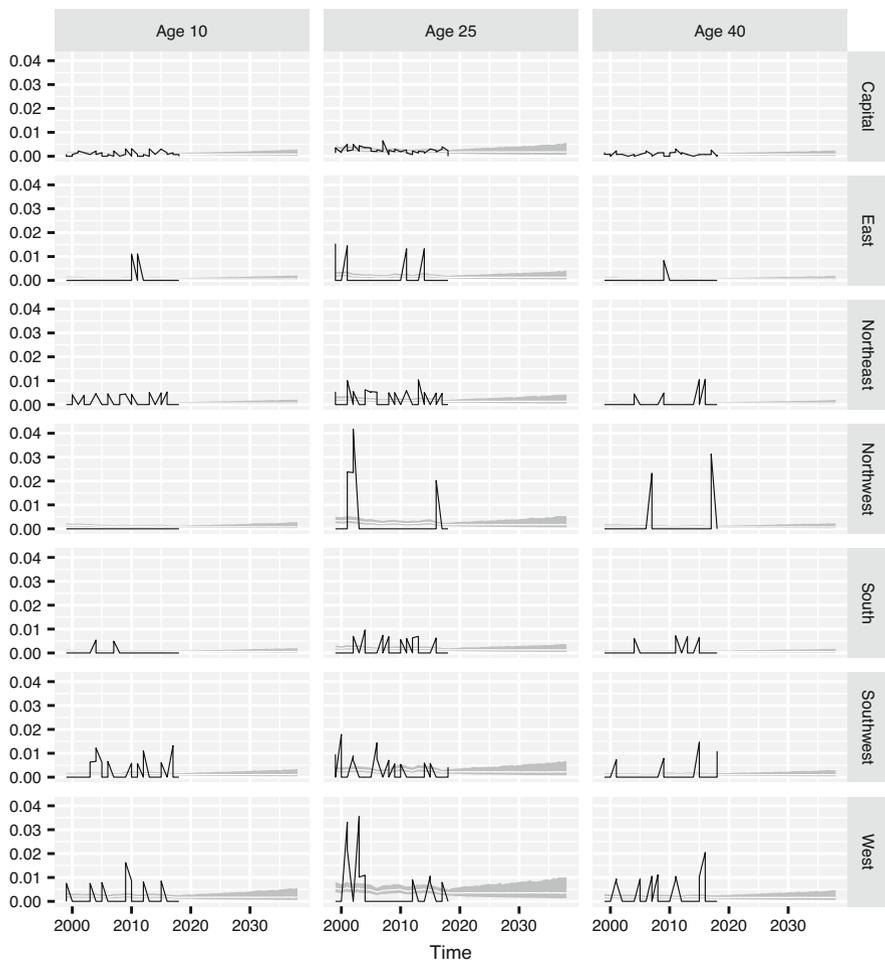


Fig. 10.10 Migration rates to Westfjords from other regions, for females in selected single-year age groups. The grey bands represent 95% credible intervals and the white lines represent posterior medians. The black lines represent direct estimates

is achievable in demographic forecasting. Users of demographer forecasts are demanding ever-more detail. Demographers and statisticians increasingly have the tools to meet these demands.

Of the remaining obstacles to the use of methods like the ones in this chapter, perhaps the most important is computation. Running all of the calculations in this chapter currently takes around 18 hours on a desktop computer. With these sorts of computation times, scaling up from 8 regions to 80 or 800 is difficult.

Speeding up computations is, however, a solvable problem. Our experience over several years is that improving algorithms and code yields steady improvements in speed, and we still have a long list of additional modifications to try. Moreover, the

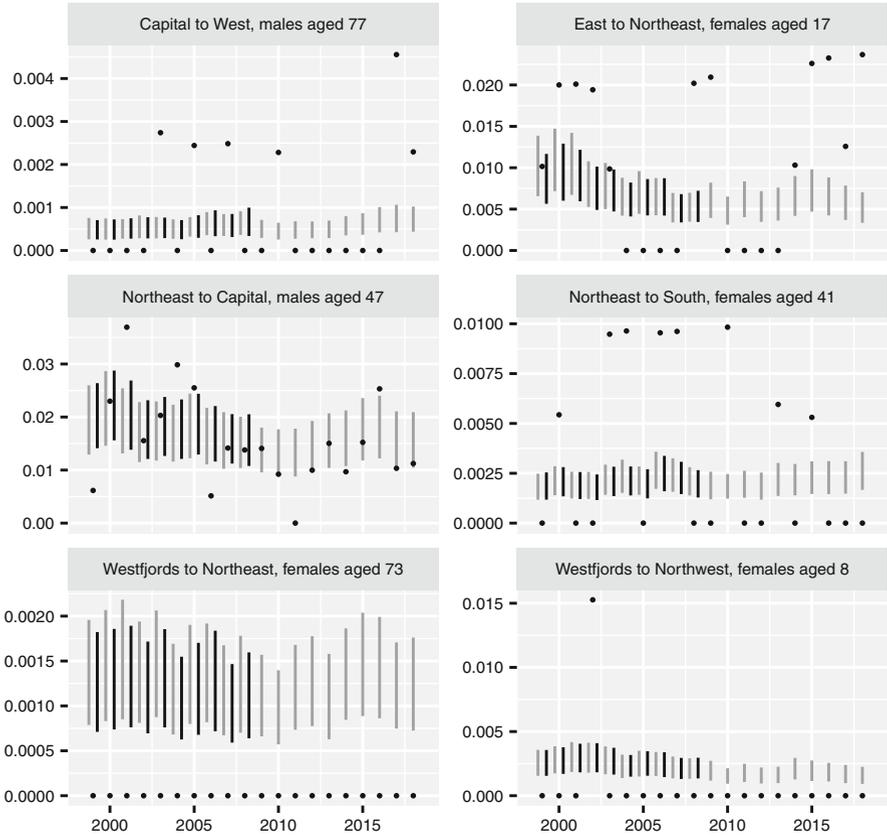


Fig. 10.11 Estimates based on the full dataset vs estimates based only on the training set. The dark grey lines show 95% credible intervals from fitting the revised model to the training set, and the light grey lines show 95% credible intervals from fitting the revised model to the full dataset. The black dots are direct estimates. Each panel shows a randomly-selected combination of origin, destination, age, and sex

rapid rise in distributing computing gives new options for attaining speed through brute force. We suspect that, before long, 80 or 800 regions will be well within reach.

References

Alexander, M., Zaghene, E., & Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, 54(6), 2025–2041.

Alho, J., & Spencer, B. (2005). *Statistical demography and forecasting*. Berlin-Heidelberg: Springer.

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*, 3(1), 133–160.

- Baffour, B., & Raymer, J. (2019). Estimating multiregional survivorship probabilities for sparse data: An application to immigrant populations in Australia, 1981–2011. *Demographic Research*, 40, 463–502.
- Bernard, A., & Bell, M. (2015). Smoothing internal migration age profiles for comparative research. *Demographic Research*, 32, 915–948.
- Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, 70(1), 1–19.
- Bryant, J., & Zhang, J. L. (2018). *Bayesian demographic estimation and forecasting*. Boca Raton: CRC Press.
- Congdon, P. (2008). Models for migration age schedules: A Bayesian perspective with an application to flows between Scotland and England. In J. Raymer & F. Willekens (Eds.), *International migration in Europe: Data, models and estimates*. Chichester: Wiley.
- Gelman, A., Carlin, J., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 4(2), 1360–1383.
- Lynch, S. M. & Brown, J. S. (2010). Obtaining multistate life table distributions for highly refined subpopulations from cross-sectional data: A Bayesian extension of Sullivan's method. *Demography*, 47(4), 1053–1077.
- Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3), 705–767.
- Prado, R., & West, M. (2010). *Time Series: Modeling, Computation, and Inference*. Boca Raton: CRC Press.
- Raymer, J., & Rogers, A. (2007). Using age and spatial flow structures in the indirect estimation of migration streams. *Demography*, 44(2), 199–223.
- Rogers, A., & Castro, L. J. (1981). *Model migration schedules*. Laxenburg: International Institute for Applied Systems Analysis.
- Rogers, A., Little, J., & Raymer, J. (2010). *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. Dordrecht: Springer Science & Business Media.
- Schmertmann, C. P., Cavenaghi, S. M., Assunção, R. M., & Potter, J. E. (2013). Bayes plus Brass: Estimating total fertility for many small areas from sparse census data. *Population Studies*, 67(3), 255–273.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- van Imhoff, E., van der Gaag, N., van Wissen, L., & Rees, P. (1997). The selection of internal migration models for European regions. *International Journal of Population Geography*, 3(2), 137–159.
- Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research*, 23, 191–222.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

