Bayesian forecasting of demographic rates for small areas

Junni Zhang

Guanghua School of Management, Peking University
Joint Work with John Bryant, Statistics New Zealand

December 21, 2014

Section 1

Introduction

Demographic Forecasting

- Forecasts of the total number of people are not sufficient.
- Planning for hospitals, bridges, schools, housing, and much else besides requires population forecasts for small areas.
- Forecasts typically extend many years into the future, because planning for infrastructure requires long time horizons.

Our Application

e.g. counts of “permanent and long-term" departures from New Zealand, region=Kaikoura, time=2012

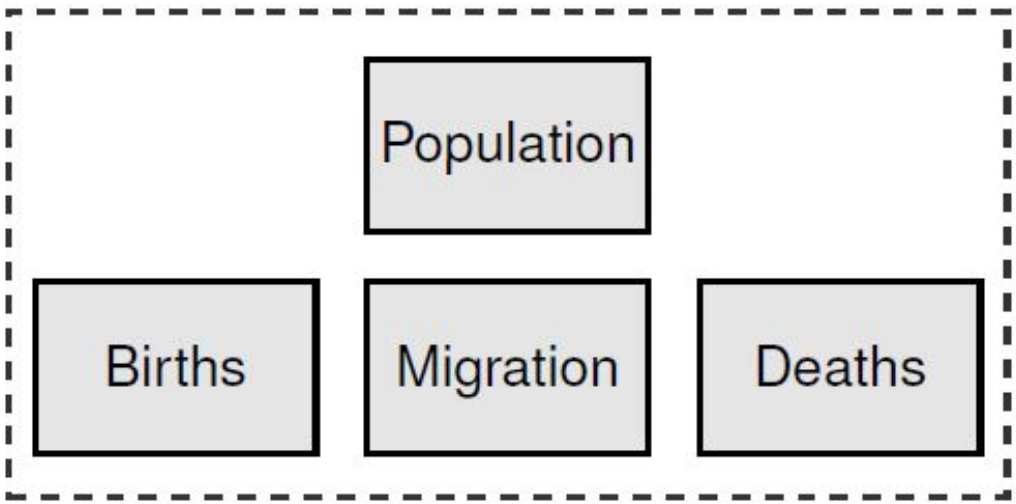
	Sex	
Age	Female	Male
0-4	3	1
5-9	5	3
10-14	3	3
15-19	3	1
20-24	7	5
25-29	3	8
30-34	2	1
35-39	2	4
40-44	5	1
45-49	1	5
50-54	1	0
55-59	2	1
60-64	0	0
65-69	1	0
70-74	0	2
75+	0	0

- Cross-classified counts, rates, probabilities
- Dimensions: age, sex, region, time

Our Application

We produce statistical forecasts for emigration rates for 16 age groups, 2 sexes, and 73 regions over 25 years.

Traditional Demographic Approach



- Forecasts are constructed for birth rates, death rates and migration rates, disaggregated by age and sex.
- The accounting identity is then repeatedly applied, to give forecasted population in each period after the base year.

Traditional Demographic Approach

- Traditionally mathematics, not statistics
- Data evaluation, rich but informal
- Struggling with disaggregation
- Uncertainty: ‘low’, ‘medium’, and ‘high’ variants.
  - When population forecasts are assembled from low, median, and high variants for fertility, mortality and migration, the results are often counter-intuitive.
  - For instance, combinations of variants that lead to large variation in population size may lead to small variation in the ratio of young people to old people.

Probabilistic Approaches

- Researchers have developed probabalistic approaches to forecasting that combine ideas from demography with ideas from the time series literature.
  - Lee and Carter (1992), Booth (2006), Booth and Tickle (2008), Alkema et al., (2011), Raftery et al. (2012), Bijak and Wiśniowski (2010)
- However, almost all the research on population has focused on national-level projections.

Complications with Small Area Forecasts

- Increasing prominence of random variation as the data become more disaggregated.
- Virtually all geographically-disaggregated data contain gaps and breaks due to changes in administrative boundaries.
  - Before 2010 there were 73 territorial authorities in New Zealand.
  - During 2010, seven territorial authorities within greater Auckland were amalgamated into a single unit.

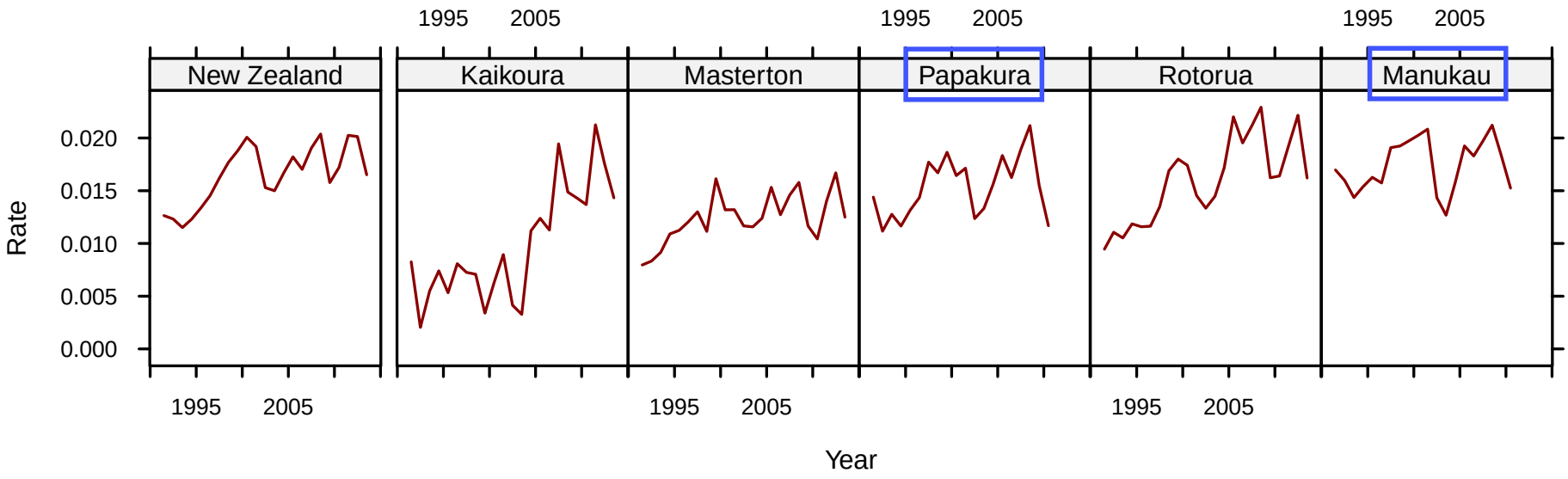
Our Approach

- We draw on ideas from the literature on small area estimation, in addition to demography and time series statistics.
- We develop a Bayesian hierarchical model.

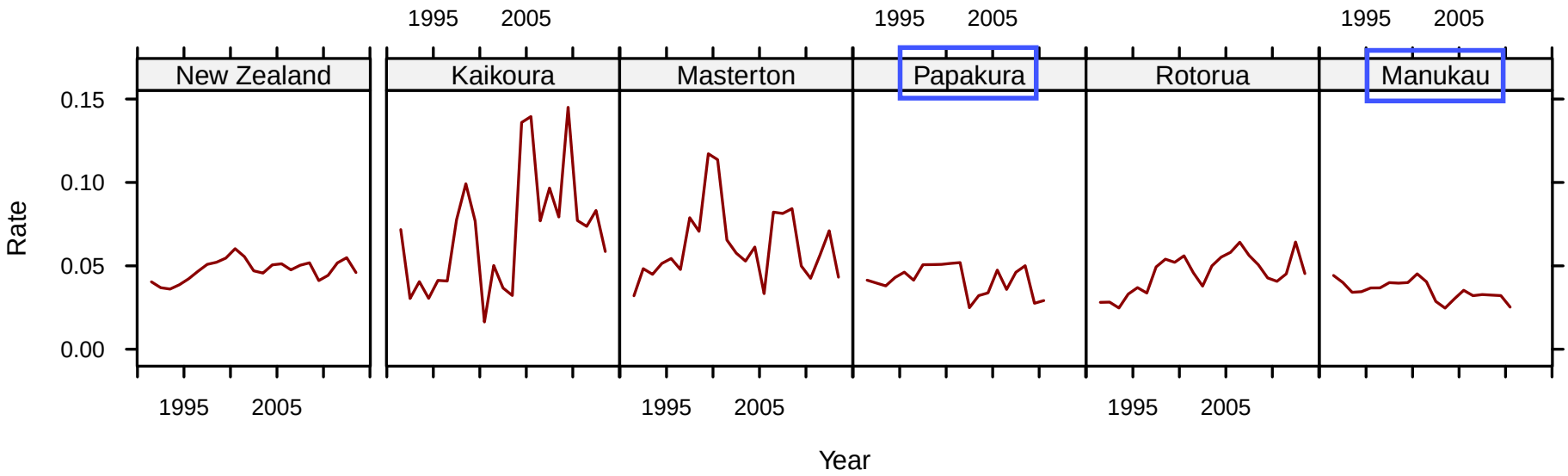
Section 2

Data And Methods

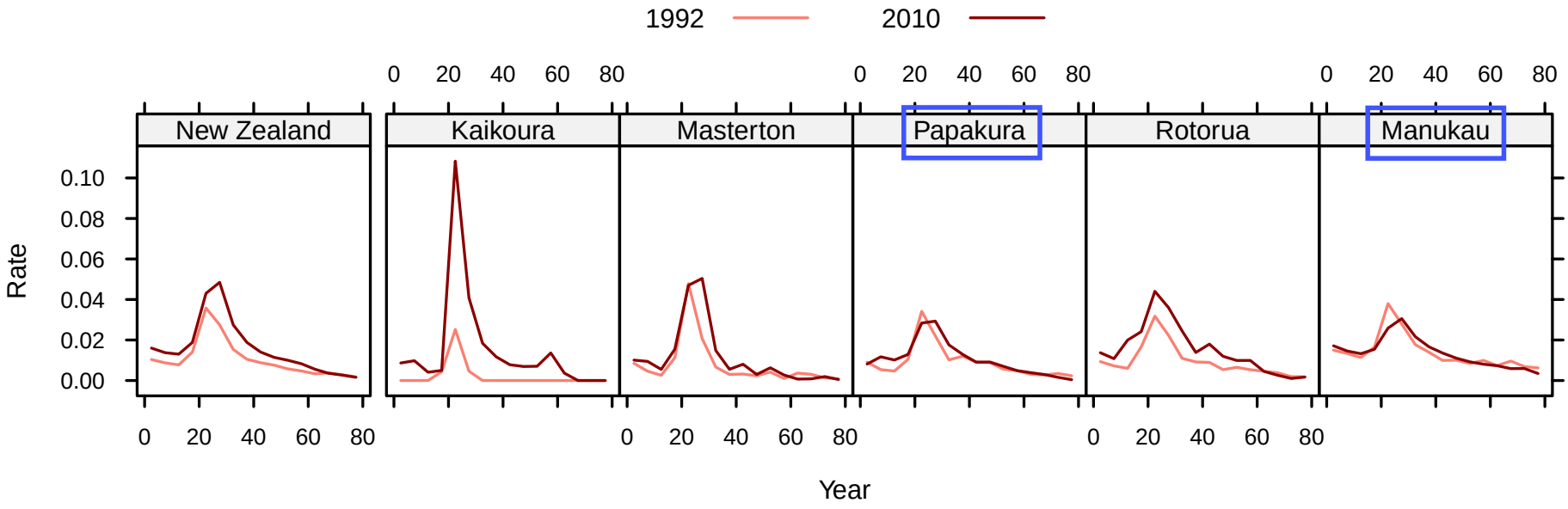
Direct Estimation: Both Sexes, All Ages



Direct Estimation: Female, Age 20-24



Direct Estimation: Both Sexes, Time 1992 and 2010



Basic Model

For age a, sex s, region r and time t, let x_{asrt} denote population size, and let y_{asrt} denote the count of international departures.

$$y_{asrt} \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{asrt} x_{asrt})$$

$$\begin{aligned}\log \lambda_{asrt} = {} & \beta^0 + \beta_a^{\text{age}} + \beta_s^{\text{sex}} + \beta_r^{\text{reg}} + \beta_t^{\text{time}} \\ & + \beta_{as}^{\text{age:sex}} + \beta_{ar}^{\text{age:reg}} + \beta_{sr}^{\text{sex:reg}} + \beta_{asr}^{\text{age:sex:reg}} + \epsilon_{asrt}\end{aligned}$$

$$\epsilon_{asrt} \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon^2)$$

Basic Model

$\beta_t^{\text{time}} \sim$ a non-stationary polynomial trend model with order p

$p = 1$:

$$\begin{aligned} \beta_t^{\text{time}} &= \theta_{t,1} + v_t \\ \theta_{t,1} &= \theta_{t-1,1} + w_{t,1} \end{aligned}$$

$p = 2$:

$$\begin{aligned} \beta_t^{\text{time}} &= \theta_{t,1} + v_t \\ \theta_{t,1} &= \theta_{t-1,1} + \theta_{t-1,2} + w_{t,1} \\ \theta_{t,2} &= \theta_{t-1,2} + w_{t,2} \end{aligned}$$

$\beta_a^{\text{age}} \sim$ a non-stationary polynomial trend model with order q

Basic Model

$$\beta_r^{\text{reg}} = \boldsymbol{\gamma}^\top \boldsymbol{X}_r + u_r, \tag{1}$$

$\boldsymbol{X}_r$ consists of:

- the logarithm of percent of population born overseas for region r.
- the logarithm of percent of population in full-time study for region r.

$u_r \overset{\text{ind}}{\sim} N(0, \sigma_u^2)$

Prior Distributions for Other Parameters

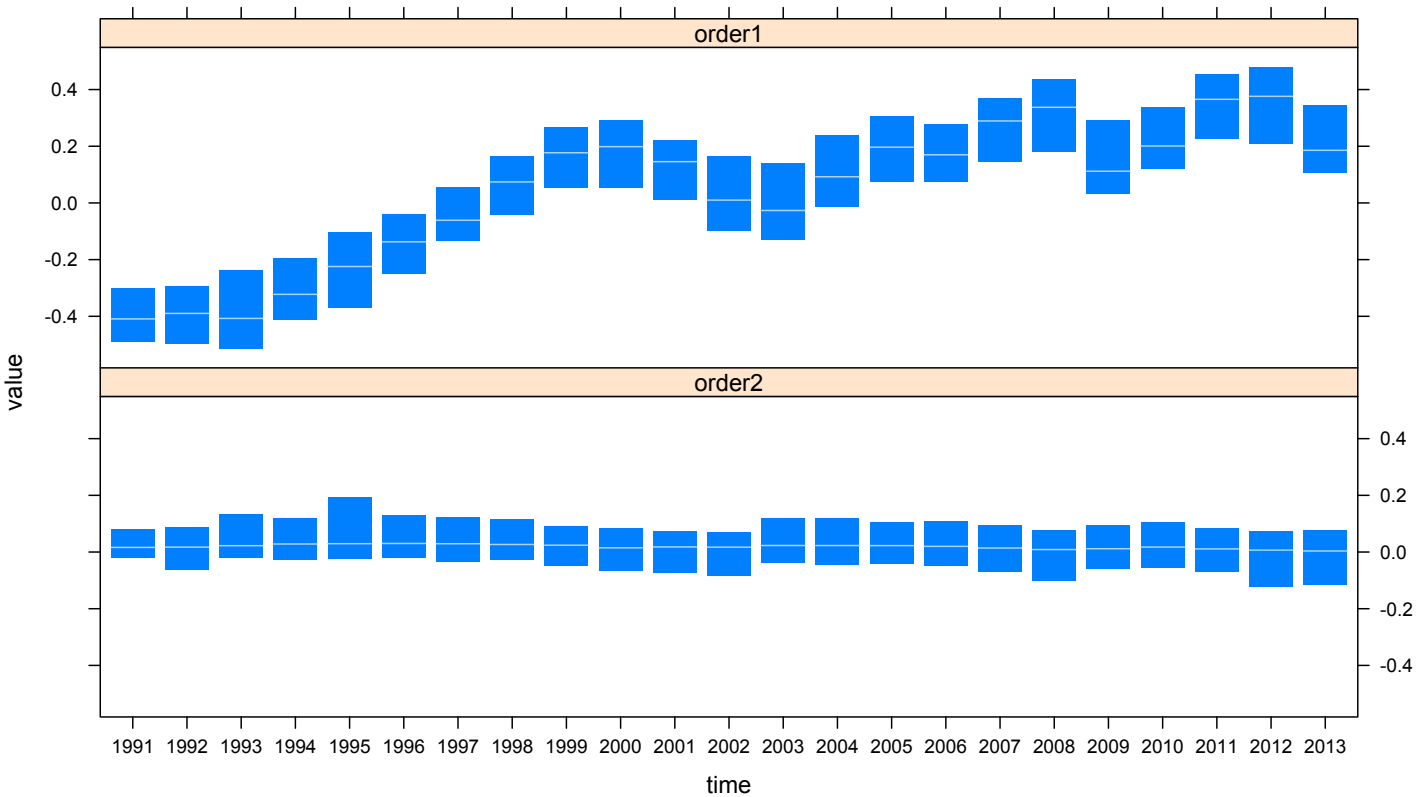
$$
\begin{aligned}
\beta_{as}^{\text{age:sex}} &\overset{\text{ind}}{\sim} N(0, \sigma^2_{\text{age:sex}}),\\
\beta_{ar}^{\text{age:reg}} &\overset{\text{ind}}{\sim} N(0, \sigma^2_{\text{age:reg}}),\\
\beta_{sr}^{\text{sex:reg}} &\overset{\text{ind}}{\sim} N(0, \sigma^2_{\text{sex:reg}}),\\
\beta_{asr}^{\text{age:sex:reg}} &\overset{\text{ind}}{\sim} N(0, \sigma^2_{\text{age:sex:reg}}).
\end{aligned}
$$

The regression coefficients and the standard deviations follow improper uniform prior distributions.

MCMC

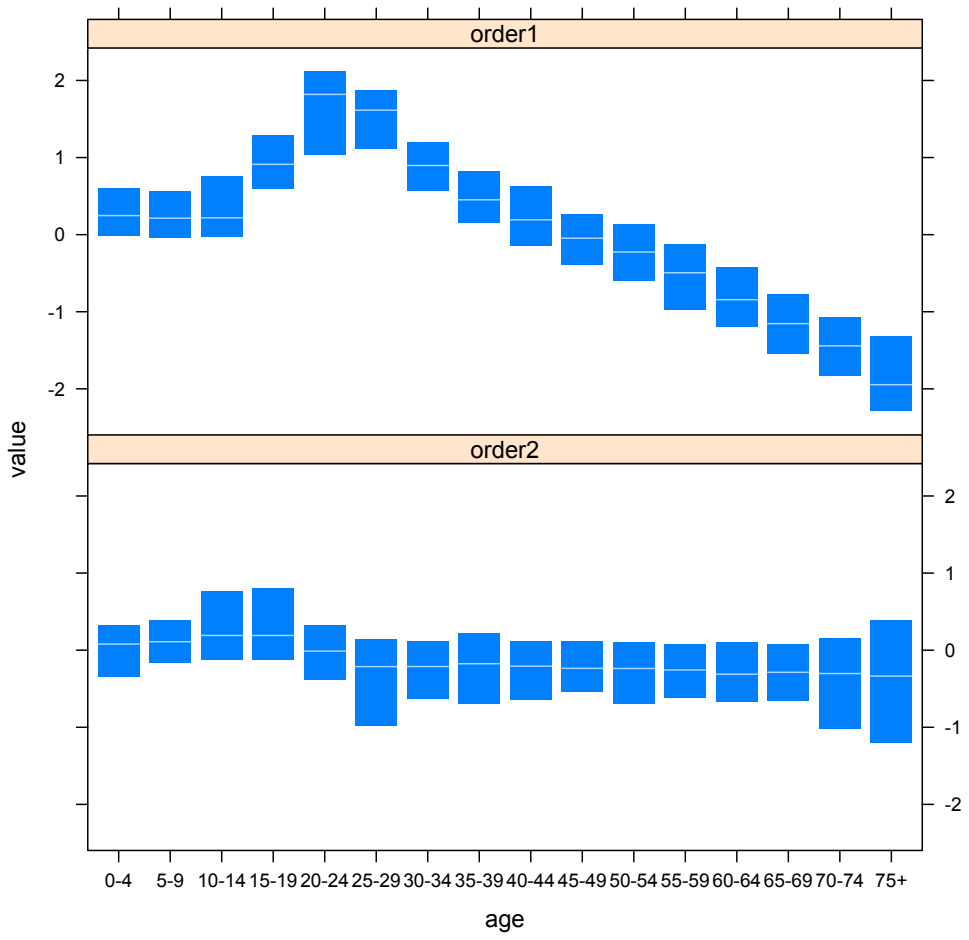
We use a Markov Chain Monte Carlo (MCMC) algorithm to draw the parameters from their posterior distribution.

Time Order Terms



We choose $p = 1$.

Age Order Terms



We choose $q = 2$.

Missing Values for Region: First Problem

7.5% of records have no regional information at all, either because the respondent did not provide it, or because the response could not be coded.

We address this issue through multiple imputation.

Missing Values for Region: First Problem

Statistics New Zealand has information on citizenship that cannot be released publicly.

"... imputation performed by the data collector (e.g. the Census Bureau) has the important advantage of allowing the use of information available to the data collector but not available to an external data analyst. · · · This kind of information, even though inaccessible to the user of a public-use file, can often improve the imputed values." (Rubin 1987)

$$\sum_{r} y_{asrtc}^{mis} = y_{astc}^{obs}$$

Missing Values for Region: Second Problem

From 2010 onwards, if the region is coded as Auckland, there is missing information on which of the seven original territorial 16 authorities within Auckland is associated with the records.

We address this issue through jointly updating these values and the parameters within the MCMC algorithm.

Section 3

Results

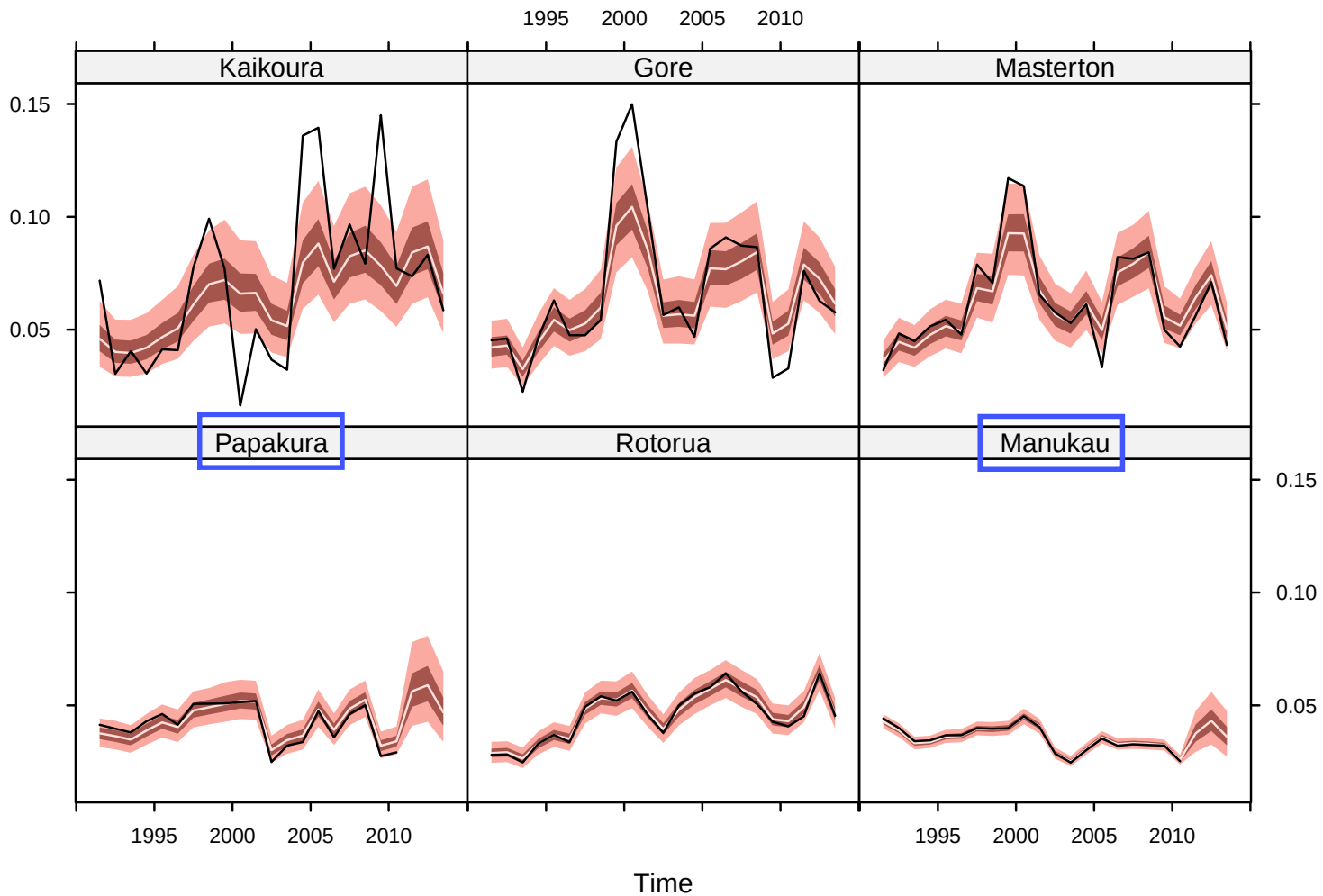
Validation Exercise

- We pretend that the event of merging seven territorial authorities within Auckland happened at 2003.
- The training data include the counts for 73 regions for 1991-2002 and the counts for 67 regions, with the counts for the seven territorial authorities within Auckland merged, for 2003-2005.
- We predict the emigration counts y_{asrt}^{pre} for 2006-2010, and obtain their posterior medians, 50% credible intervals and 90% credible intervals.
- We compare the posterior medians and credible intervals with the observed values of y_{asrt} for 2006-2010.

Results of Validation Exercise

- The median of y_{asrt} to be predicted equals to 9, and the median absolute error of using posterior medians of y_{asrt}^{pre} to predict y_{asrt} is 3.
- The percentage of y_{asrt} lying inside the 50% credible intervals is 57.1%.
- The percentage of y_{asrt} lying inside the 90% credible intervals is 89.6%.

Estimates for Female, Age 20-24



Estimates and Prediction for Female, Age 20-24