Bayesian demographic estimation and inferring counts from (un)reliable data

John Bryant, Bayesian Demography Ltd

United Nations Expert Group Meeting on Methods for the World Population Prospects 2021 and Beyond

8 April 2020

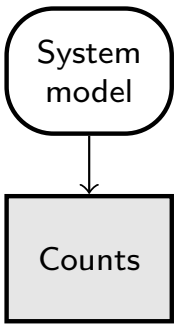
Introduction

Other contributors

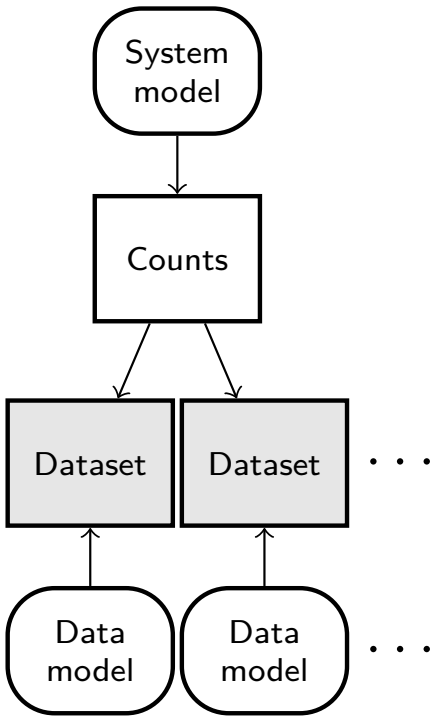
- Junni Zhang, Peking University
- Statistics New Zealand
- Many others!!

Three frameworks

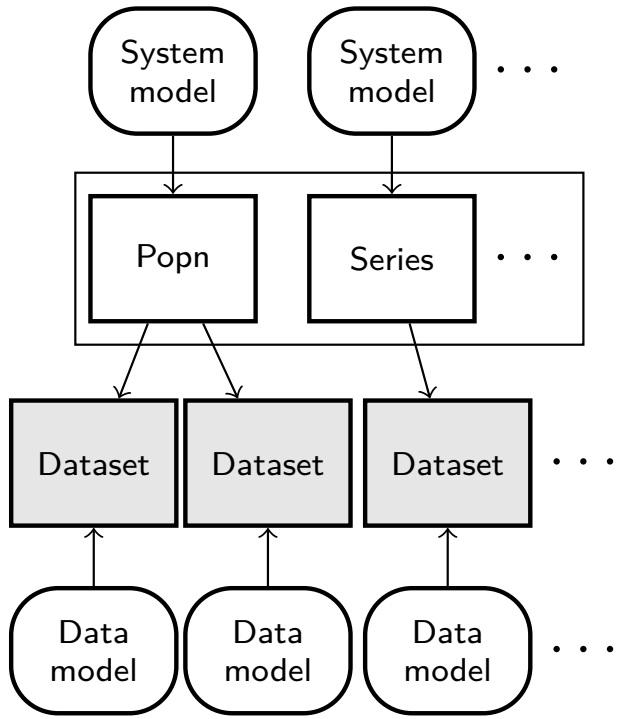
Framework 1
Rates,
reliable data



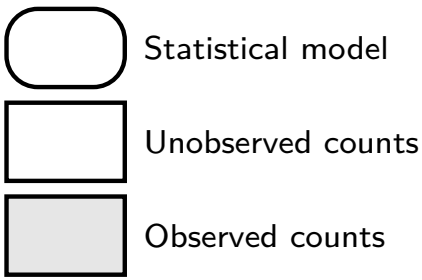
Framework 2
Counts, rates,
unreliable data



Framework 3
Account, rates,
unreliable data



Key

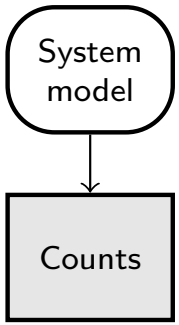


Framework 1: Estimating rates from reliable data

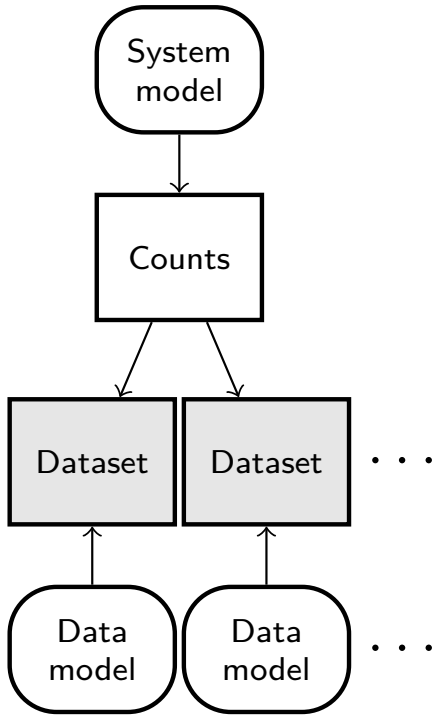
Setting

- Negligible measurement error and complete coverage
  - Eg registered births, deaths in high-income countries
- Methodological challenges for statistical demography
  - Estimation of disaggregated rates
  - Forecasting

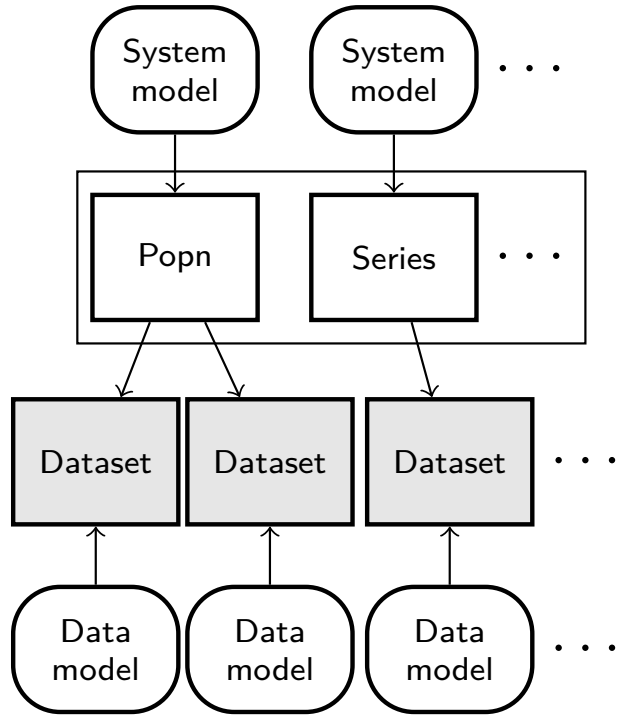
Framework 1
Rates, reliable data



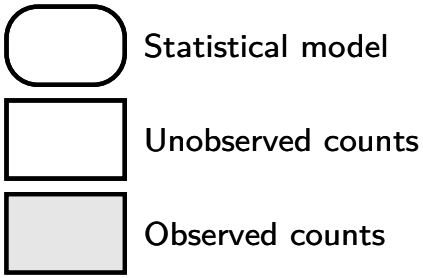
Framework 2
Counts, rates, unreliable data



Framework 3
Account, rates, unreliable data



Key

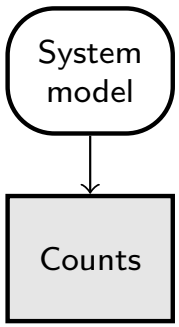


2. Estimating demographic arrays from unreliable data

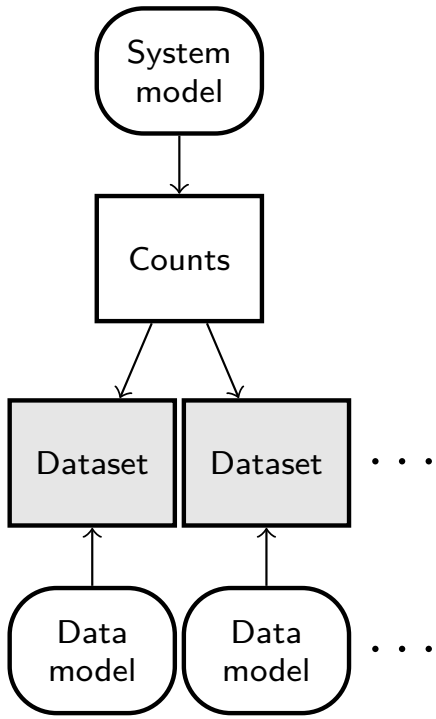
Setting

- Substantial measurement error and incomplete coverage
  - Gaps, missing variables, incorrect values
  - Eg migration in high-income countries, most series elsewhere
- Standard methodological challenges, while also correcting for gaps, measurement errors, coverage errors

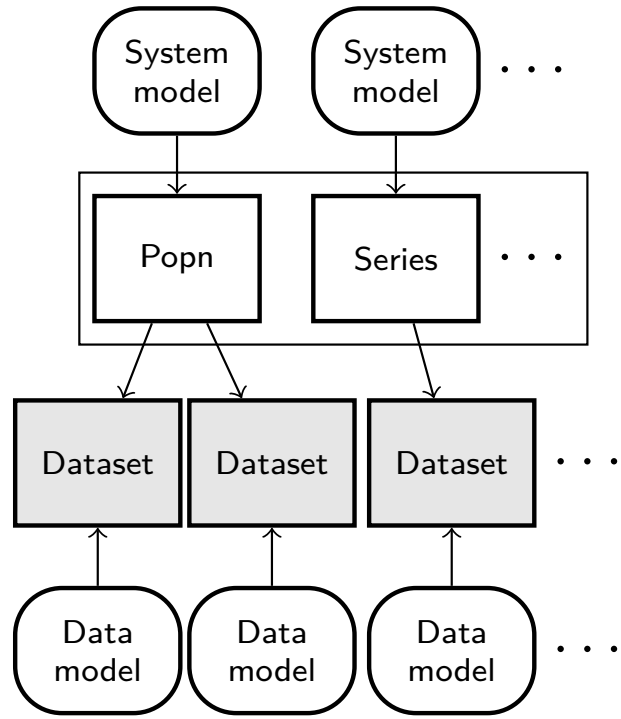
Framework 1
Rates,
reliable data



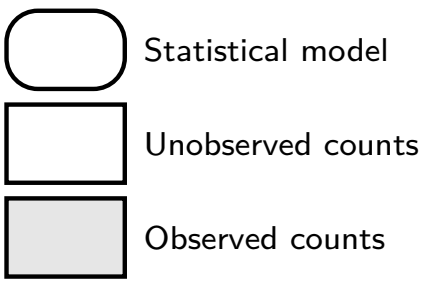
Framework 2
Counts, rates,
unreliable data



Framework 3
Account,
rates,
unreliable data



Key



System models

[Same as for reliable data]

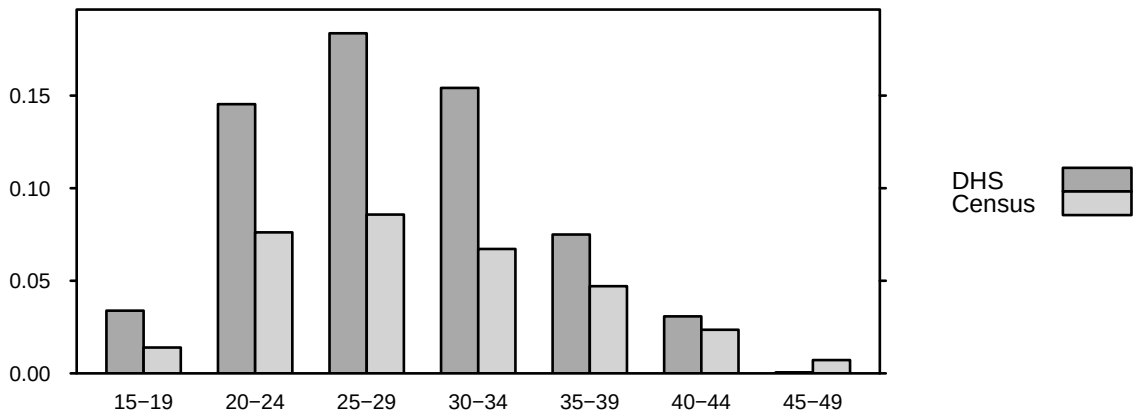
Data models

- Predict data, given true counts
  - Eg predict census counts, given true population counts
  - Eg predict registered births, given true birth counts
- Formal representation of what a demographic analyst knows about the data
  - Magnitude of likely errors
  - How errors vary with age, sex, region, time, ...
- When predicting data from true counts, dealing with gaps in data is easy

Example: Fertility rates in Cambodia

- Estimating age-specific fertility rates for 24 provinces

Data source	Series	Classification	Year	Quality
Census	Births	Age, province	2010	Substantial undercoverage
DHS	Births	Age	2008	Good but small sample
Census	Population	Age, sex, province	2010	OK



Data models

Census:

$$b_{ap}^{\text{cens}} \sim \text{Poisson}(\gamma_{ap} b_{ap}^{\text{true}}) \tag{1}$$

$$\log \gamma_{ap} \sim \text{N}(\beta^0 + \beta_a^{\text{age}} + \beta_p^{\text{prov}}, \sigma^2) \tag{2}$$

$$\beta_p^{\text{age}} \sim \text{RW} \tag{3}$$

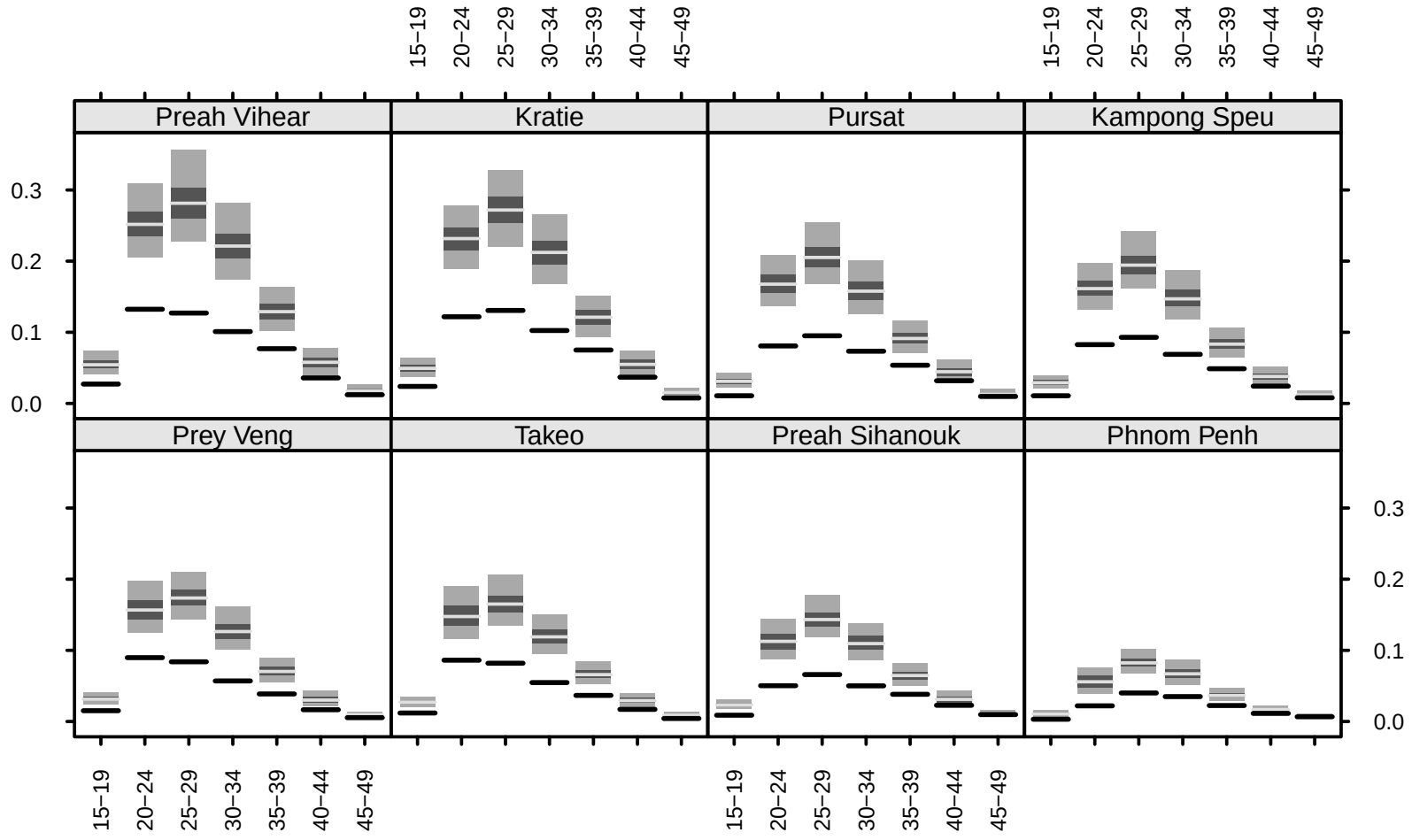
$$\beta_p^{\text{prov}} \sim \text{N}(0, \tau_{\text{prov}}^2) \tag{4}$$

$$\tau_{\text{prov}} \sim t_7^+(0.05^2) \tag{5}$$

DHS:

$$b_a^{\text{dhs}} \sim \text{N}(b_a^{\text{true}}, s_a^2) \tag{6}$$

Results: Age-specific fertility rates for selected provinces



Comparison with other approaches

- Standard separation of system and data models
- We are still scaling up
- Unusual feature 1: latent counts not rates
  - Aggregation easy
  - Maximum flexibility for absorbing data sources
  - But computationally challenging
- Unusual feature 2: Bottom-up
  - Work with disaggregated rates and counts, then aggregate up
  - Natural strategy for general-purpose framework

Working with aggregated data

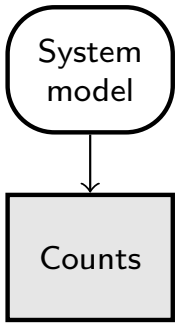
- 1-year estimates from 5-year data
- Potential solution
  - System model, unobserved counts: 1-year
  - Data models, data: 5-year
- Will experiment!

3. Estimating demographic accounts and rates from unreliable data

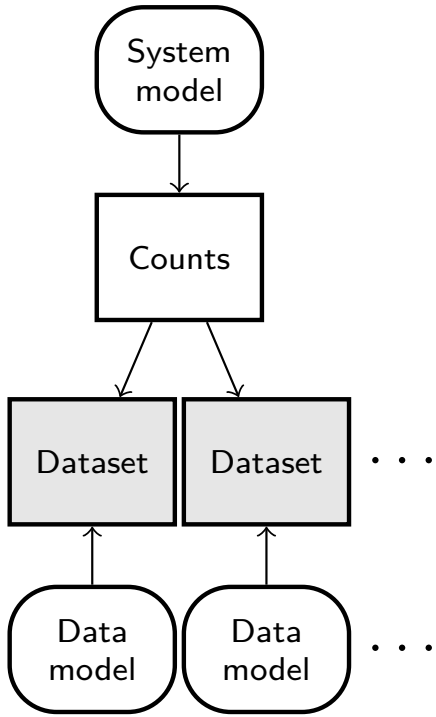
Setting

- Estimating entire demographic system
  - Eg national population with births, deaths, immigration
  - Eg employed, unemployed, not in labor force
  - Eg susceptible, infected, recovered
- Want counts and rates
- Measurement error and incomplete coverage

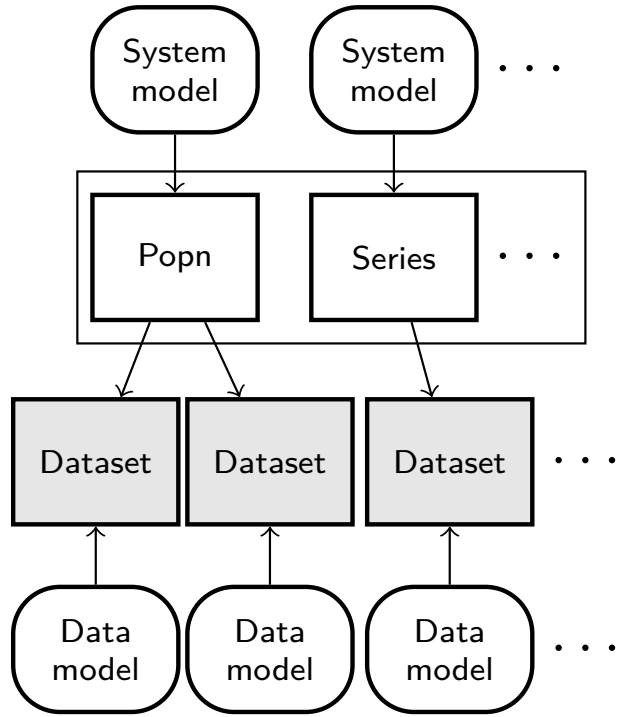
Framework 1
Rates,
reliable data



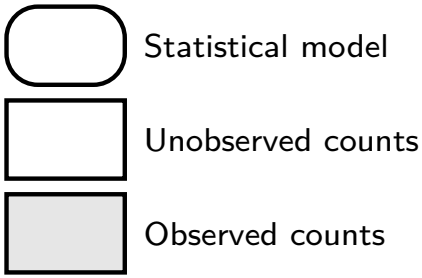
Framework 2
Counts, rates,
unreliable data



Framework 3
Account, rates,
unreliable data



Key



Demographic accounts

- Social equivalent of national accounts
- Much of applied demography?

Population

Region	2010	2020
A	50	40
B	35	20

Deaths

Region	2010-2020
A	10
B	15

Example: Demographic account for New Zealand

Detail required:

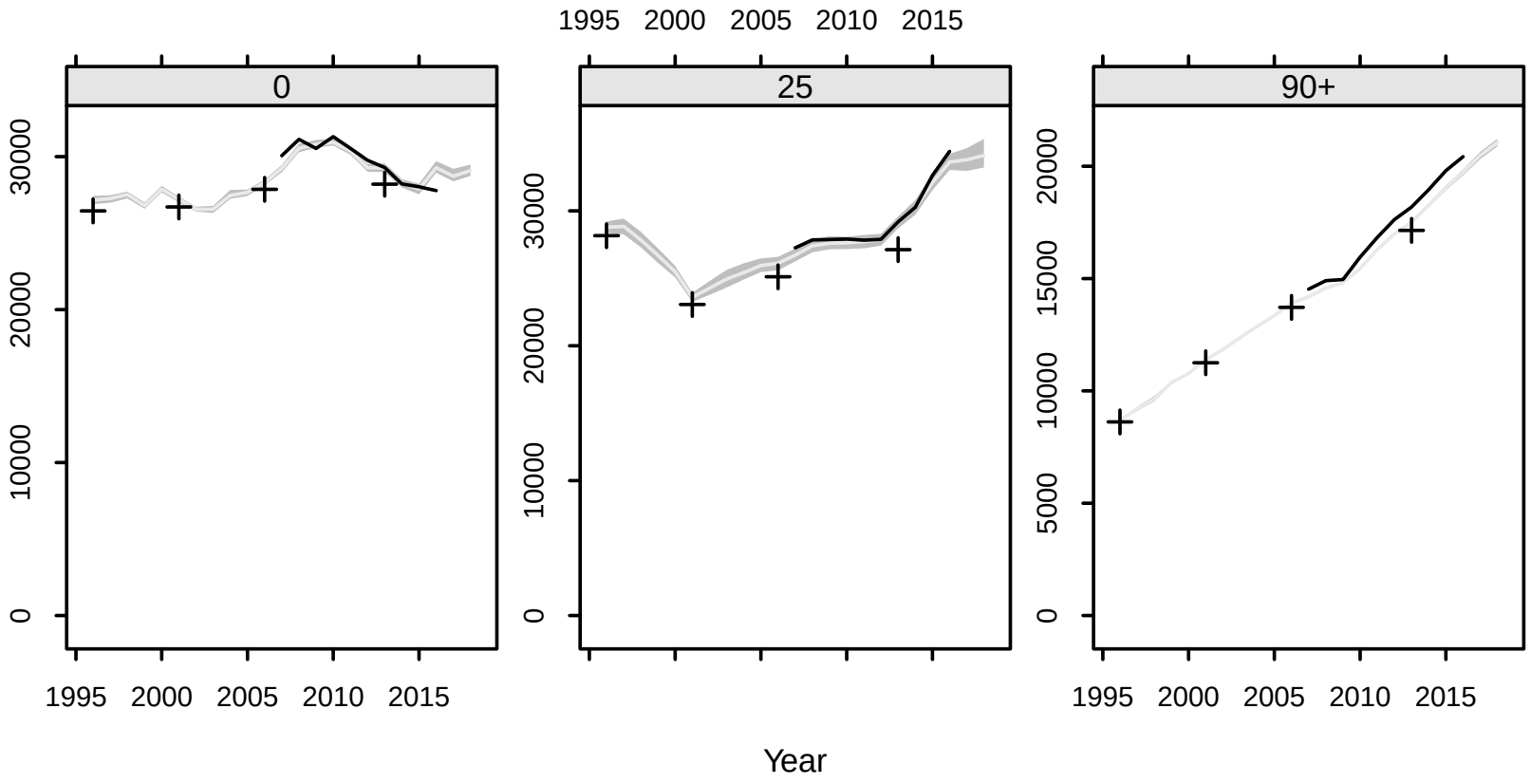
- Population, births, deaths, immigration, emigration
- Age (1-year), sex, Lexis triangle, year
- 1996–2018

Data sources

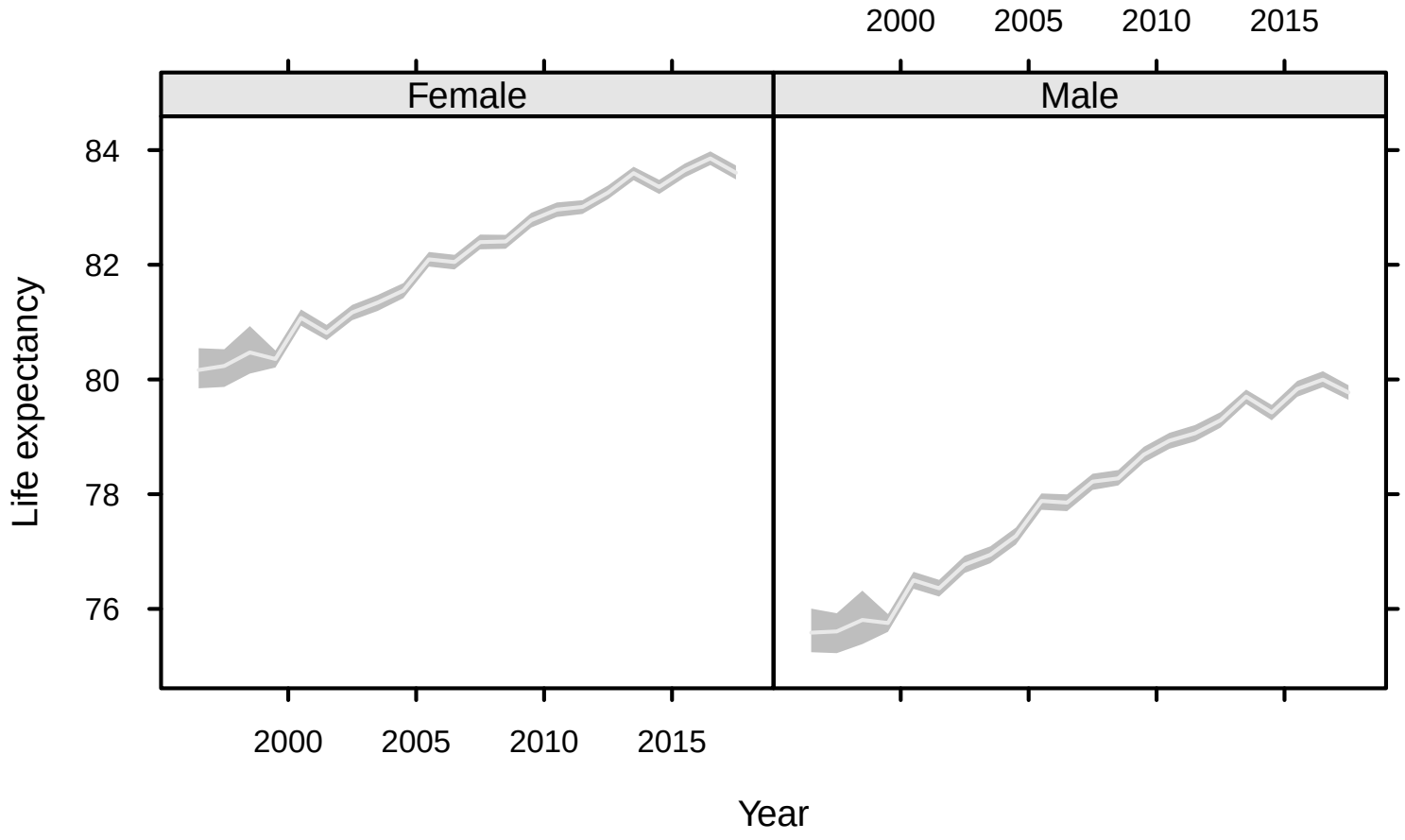
Data source	Series	Classification	Years	Quality
Census	Population	Age(1), sex	1996, 2001, 2006, 2013	Good
PES	[Data model]	Age(15)	1996, 2001, 2006, 2013	Good
Admin	Population	Age(1), sex	2007–2016	Good
Reg births	Births	Age(1)	1996–2018	Excellent
Reg deaths	Deaths	Age(1), sex	2000–2018	Excellent
Arrival cards	Immigration	Age(5), sex	1996–2018	Moderate
12/16 in-migration	Immigration	Age(5), sex	2001–2017	Good
Departure cards	Emigration	Age(5), sex	1996–2018	Moderate
12/16 out-migration	Emigration	Age(5), sex	2001–2017	Good

(All data from Statistics NZ website)

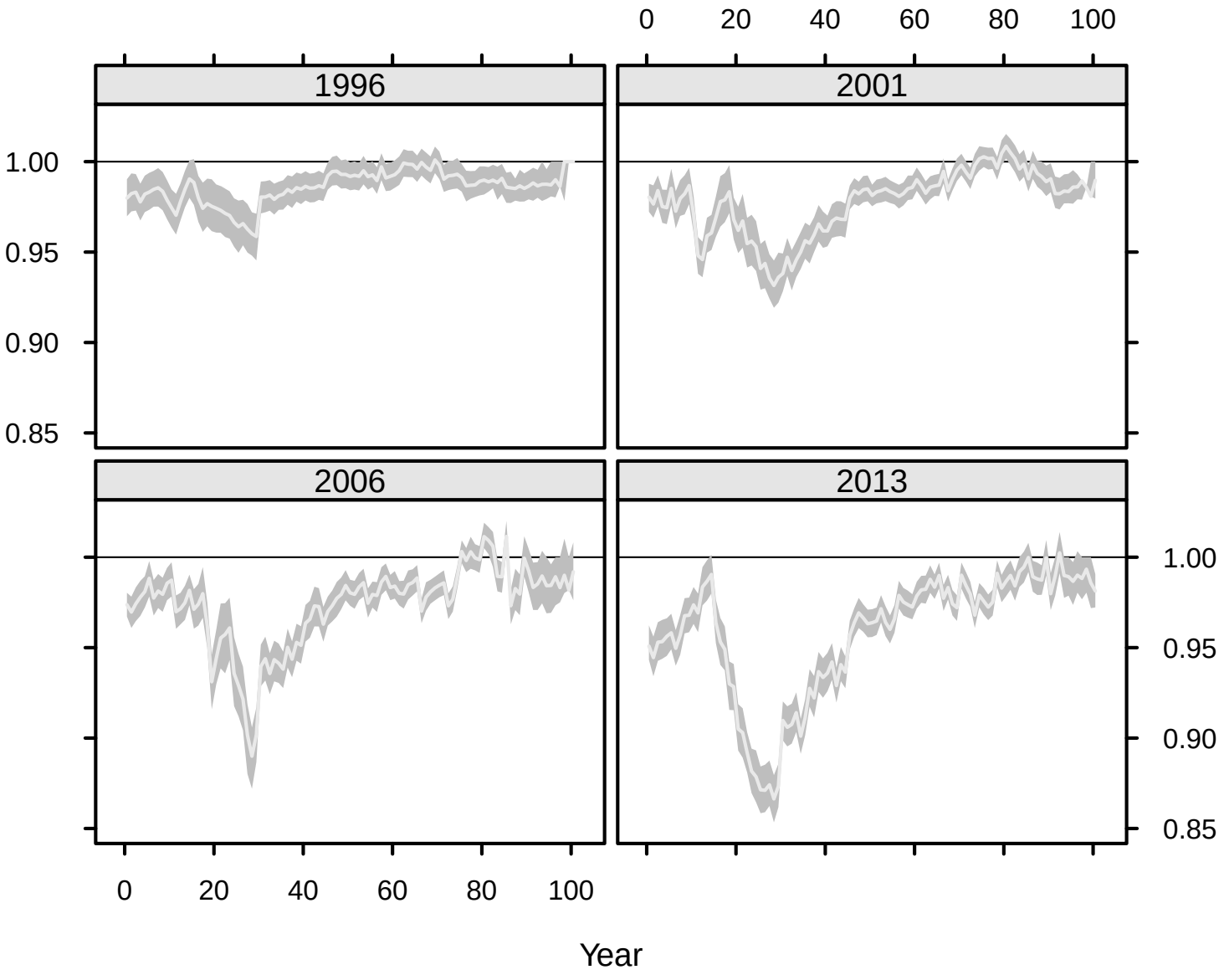
Results: Population



Results: Life expectancy



Results: Census coverage



Comparison with other approaches

- Population reconstruction, integrated population models, ...
- Unusual features
  - Flexibility about components and dimensions
  - Lexis triangles, allowing translation between cohort and period
  - Prior on population series (not just initial population)

Early population estimates for countries with limited data

- Approach of demographic historians
  - Rough estimates of population size, growth
  - Infer possible combinations of birth rates, death rates
- Formalize using Bayesian demographic accounts

Thoughts on demographic inference from unreliable data

Data models have to be simple?

- Our experience
  - System models can be surprisingly complicated
  - Data models must be simple
- But consider the default data model:

$$\text{measured counts} = \text{true counts}$$

Estimates as data

- Bayesianly correct to treat other people's estimates as input data
  - Eg population reconstruction
- Sacrifices some information, but big practical advantages
  - Don't need original data
  - Don't need to repeat analysis
- Mixed approach possible
  - A few standard data models applied to standard data
  - Estimates-as-data for idiosyncratic cases

Informative priors and unreliable data

- Bayesians: Don't be shy about using informative priors
- Data models fit the criteria for when informative priors are helpful
  - We have information beyond what is contained in data itself, eg how the data were generated
  - We can summarize that information quantitatively, eg how much province-to-province variability to expect
- Typically big improvements in convergence, plausibility