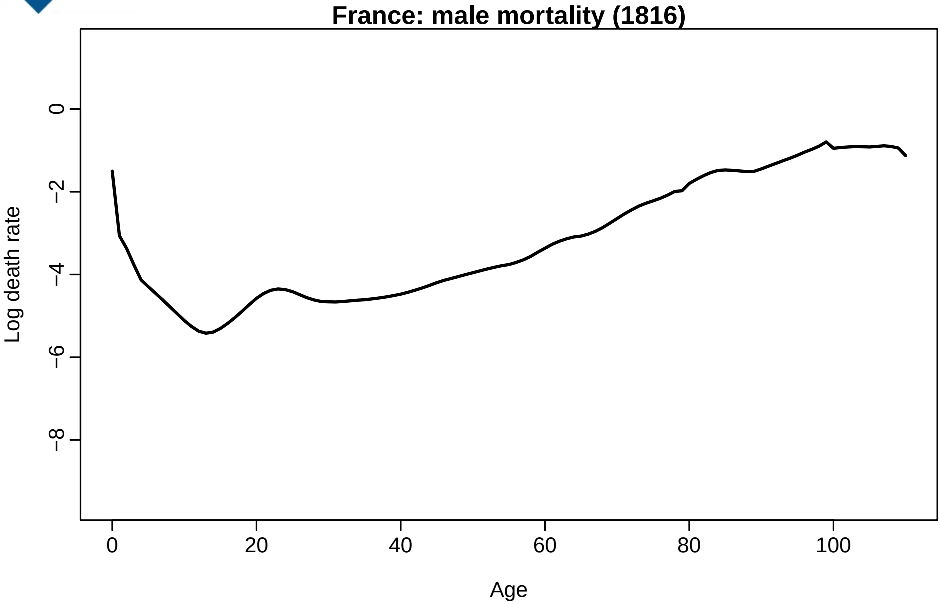
Demographic forecasting using functional data analysis

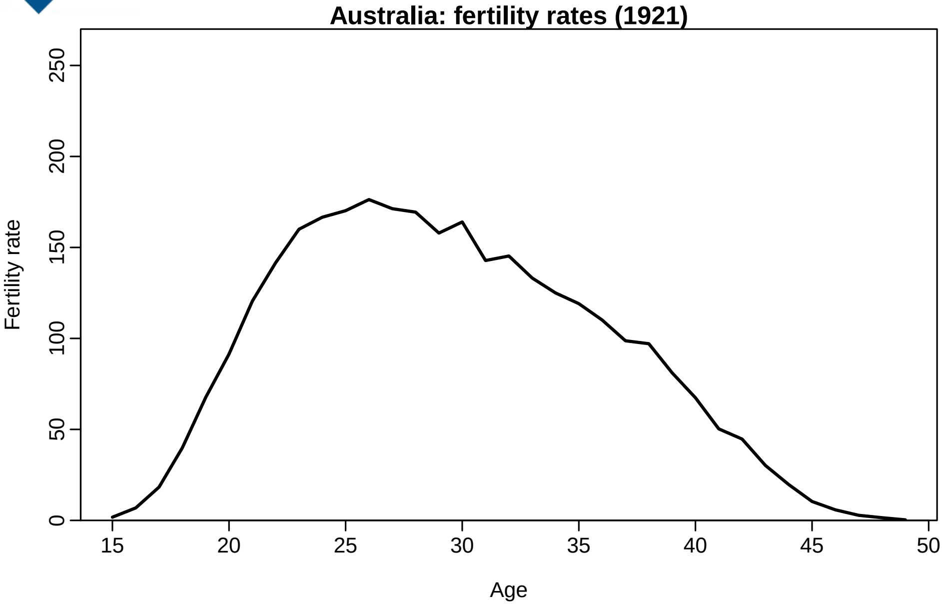
Rob J Hyndman

Joint work with: Heather Booth, Han Lin Shang, Shahid Ullah, Farah Yasmeen.

Mortality rates



Fertility rates



Outline

Outline

Some notation

Let $y_{t,x}$ be the observed (smoothed) data in period t at age x, $t = 1, \dots, n$.

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- Estimate $f_t(x)$ using penalized regression splines.
- Estimate $\mu(x)$ as me(di)an $f_t(x)$ across years.
- Estimate $\beta_{t,k}$ and $\phi_k(x)$ using (robust) functional principal components.
- $\varepsilon_{t,x} \overset{iid}{\sim} N(0,1)$ and $e_t(x) \overset{iid}{\sim} N(0, v(x))$.

Some notation

Let $y_{t,x}$ be the observed (smoothed) data in period t at age x, $t = 1, \dots, n$.

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- Estimate $f_t(x)$ using penalized regression splines.
- Estimate $\mu(x)$ as me(di)an $f_t(x)$ across years.
- Estimate $\beta_{t,k}$ and $\phi_k(x)$ using (robust) functional principal components.
- $\varepsilon_{t,x} \overset{\text{iid}}{\sim} \text{N}(0,1)$ and $e_t(x) \overset{\text{iid}}{\sim} \text{N}(0, v(x))$.

Some notation

Let $y_{t,x}$ be the observed (smoothed) data in period t at age x, $t = 1, \dots, n$.

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- Estimate $f_t(x)$ using penalized regression splines.
- Estimate $\mu(x)$ as me(di)an $f_t(x)$ across years.
- Estimate $\beta_{t,k}$ and $\phi_k(x)$ using (robust) functional principal components.
- $\varepsilon_{t,x} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$ and $e_t(x) \overset{\text{iid}}{\sim} \mathrm{N}(0, v(x))$.

Some notation

Let $y_{t,x}$ be the observed (smoothed) data in period t at age x, $t = 1, \dots, n$.

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- Estimate $f_t(x)$ using penalized regression splines.
- Estimate $\mu(x)$ as me(di)an $f_t(x)$ across years.
- Estimate $\beta_{t,k}$ and $\phi_k(x)$ using (robust) functional principal components.
- $\varepsilon_{t,x} \overset{\text{iid}}{\sim} N(0,1)$ and $e_t(x) \overset{\text{iid}}{\sim} N(0, v(x))$.

Some notation

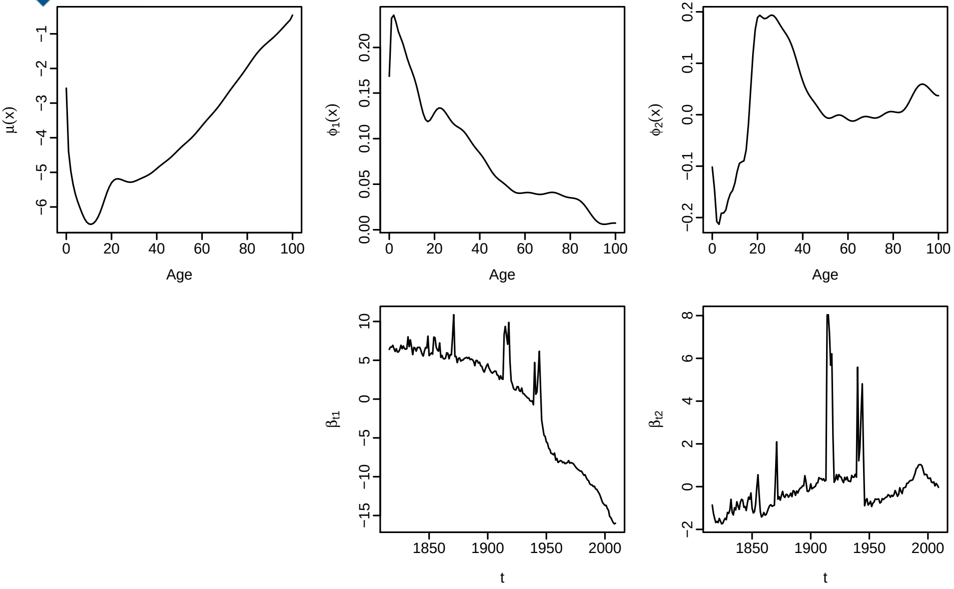
Let $y_{t,x}$ be the observed (smoothed) data in period t at age x, $t = 1, \dots, n$.

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

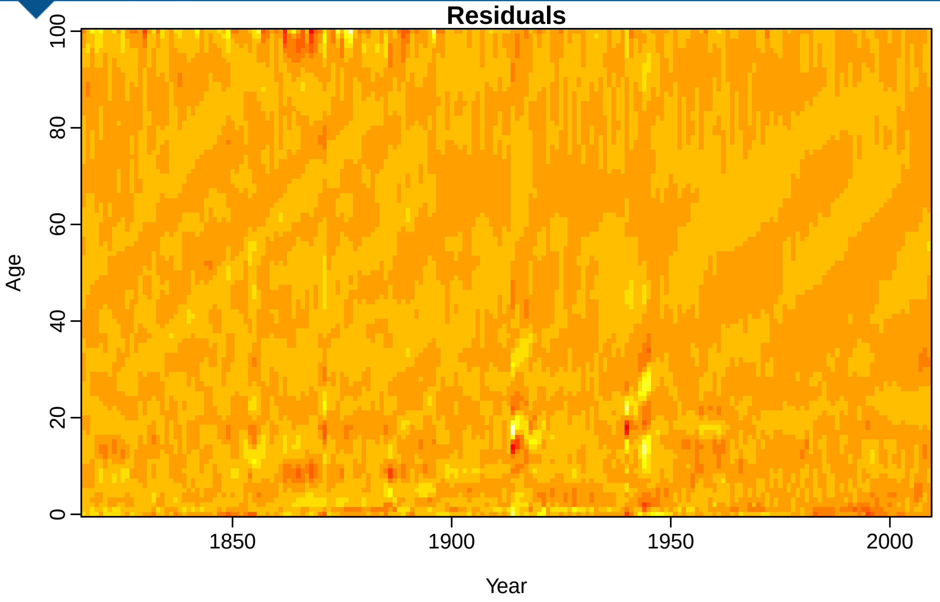
$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- Estimate $f_t(x)$ using penalized regression splines.
- Estimate $\mu(x)$ as me(di)an $f_t(x)$ across years.
- Estimate $\beta_{t,k}$ and $\phi_k(x)$ using (robust) functional principal components.
- $\varepsilon_{t,x} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$ and $e_t(x) \overset{\text{iid}}{\sim} \mathrm{N}(0, v(x))$.

French mortality components

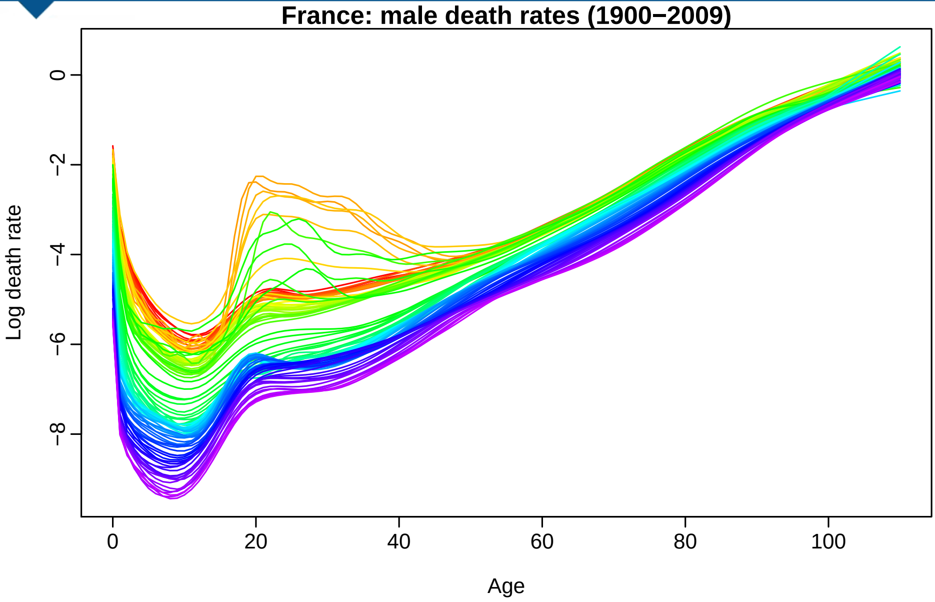


French mortality components

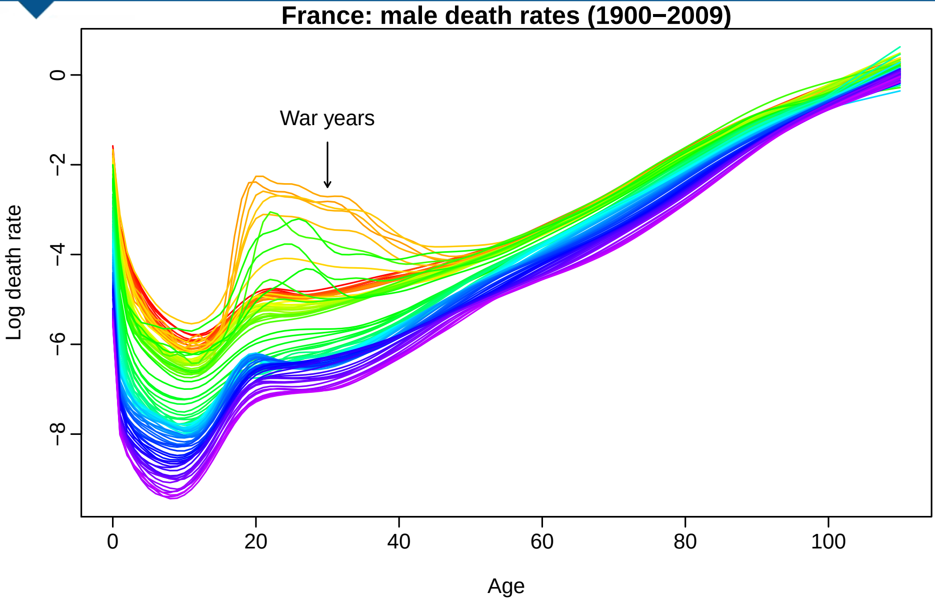


Outline

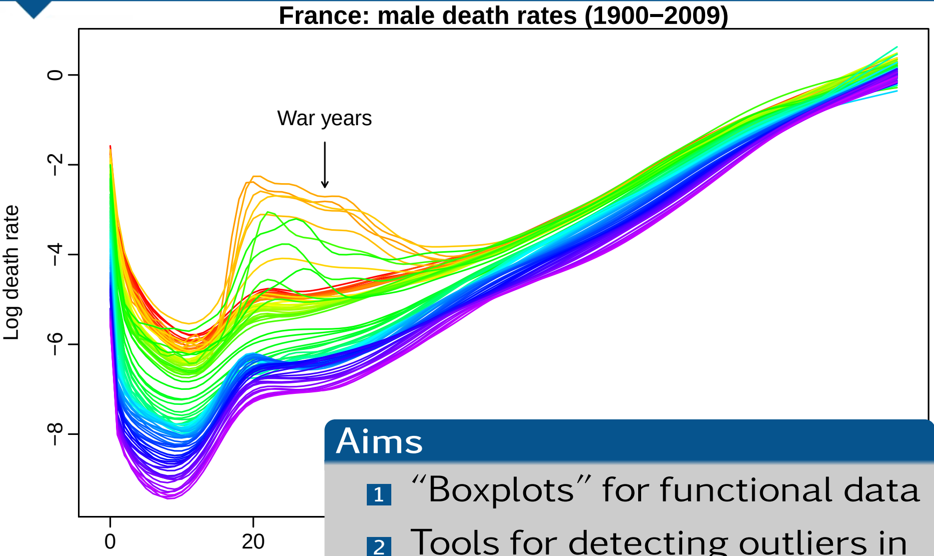
French male mortality rates



French male mortality rates



French male mortality rates



Aims

1. "Boxplots" for functional data
2. Tools for detecting outliers in functional data

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

   - $\mu(x)$ is median curve

2. Plot $\beta_{t,2}$ vs $\beta_{t,1}$

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

   - $\mu(x)$ is median curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

- $\mu(x)$ is median curve
- $\{\phi_k(x)\}$ are principal components
- $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

   - $\mu(x)$ is median curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

- $\mu(x)$ is median curve
- $\{\phi_k(x)\}$ are principal components
- $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

   - $\mu(x)$ is median curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{\beta_{t,k}\}$ are PC scores

2. **Plot $\beta_{i,2}$ vs $\beta_{i,1}$**

   - Each point in scatterplot represents one curve.
   - Outliers show up in bivariate score space

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \ldots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k}\phi_k(x)$$

- $\mu(x)$ is median curve
- $\{\phi_k(x)\}$ are principal components
- $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

➡ Each point in scatterplot represents one curve.

➡ Outliers show up in bivariate score space.

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \ldots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k}\phi_k(x)$$

- $\mu(x)$ is median curve
- $\{\phi_k(x)\}$ are principal components
- $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

➥ Each point in scatterplot represents one curve.

➥ Outliers show up in bivariate score space.

Robust principal components

Let $\{f_t(x)\}$, $t = 1, \dots, n$, be a set of curves.

1. Apply a robust principal component algorithm

$$f_t(x) = \mu(x) + \sum_{k=1}^{n-1} \beta_{t,k} \phi_k(x)$$

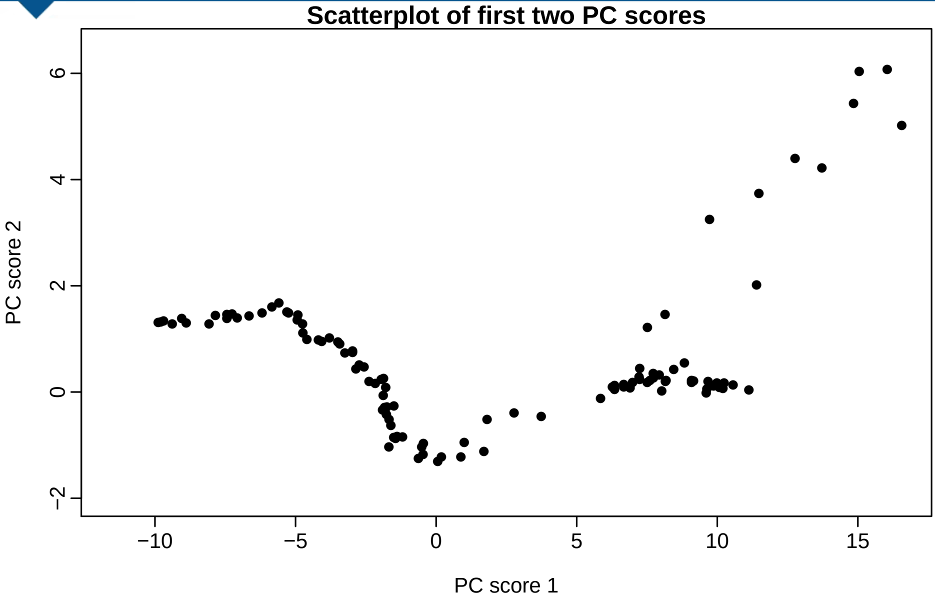
   - $\mu(x)$ is median curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{\beta_{t,k}\}$ are PC scores

2. Plot $\beta_{i,2}$ vs $\beta_{i,1}$

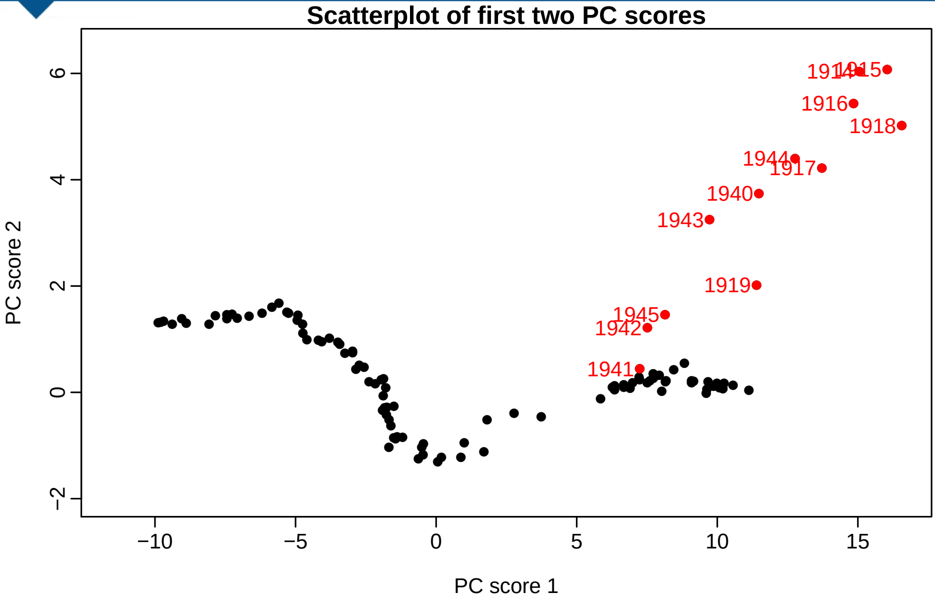
➥ Each point in scatterplot represents one curve.

➥ Outliers show up in bivariate score space.

Robust principal components

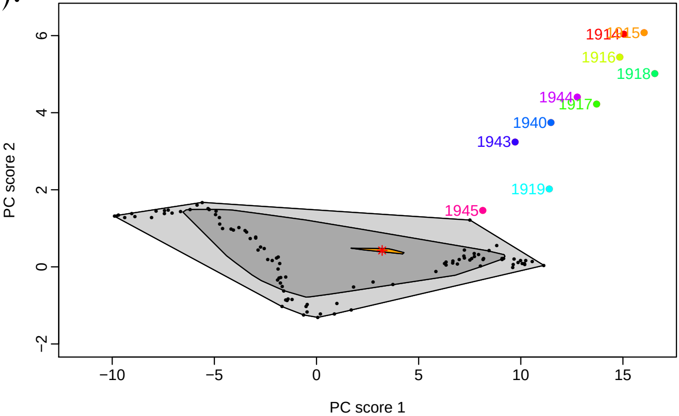


Robust principal components



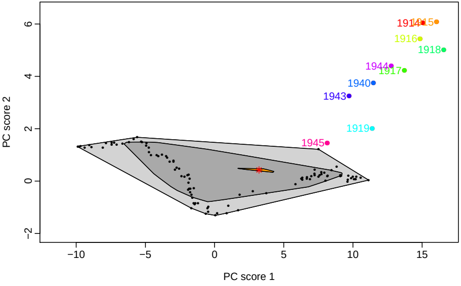
Functional bagplot

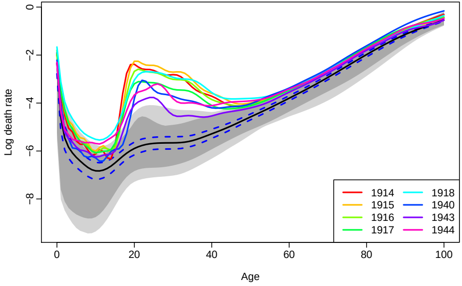
- Bivariate bagplot due to Rousseeuw et al. (1999).
- Rank points by halfspace location depth.
- Display median, 50% convex hull and outer convex hull (with 99% coverage if bivariate normal).



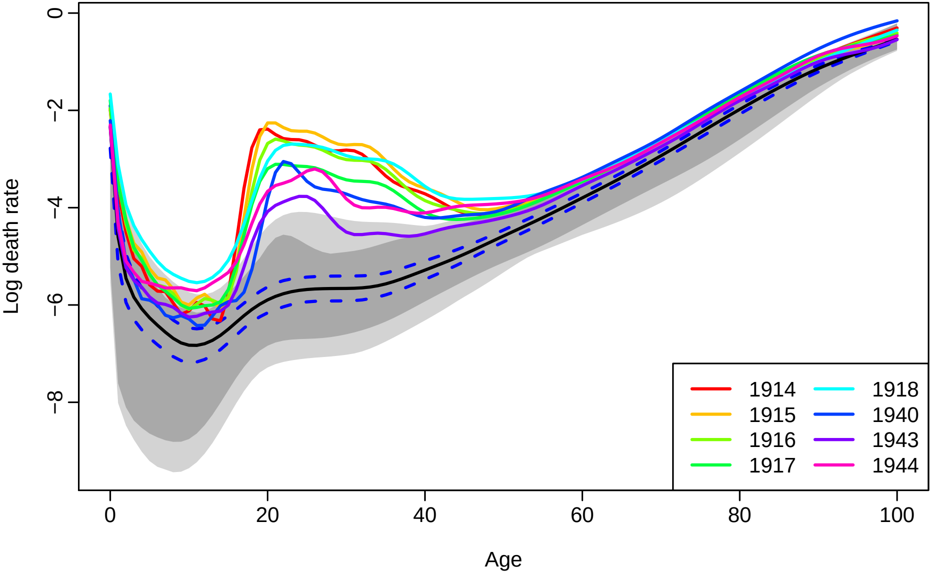
Functional bagplot

- Bivariate bagplot due to Rousseeuw et al. (1999).
- Rank points by halfspace location depth.
- Display median, 50% convex hull and outer convex hull (with 99% coverage if bivariate normal).
- Boundaries contain all curves inside bags.
- 95% CI for median curve also shown.



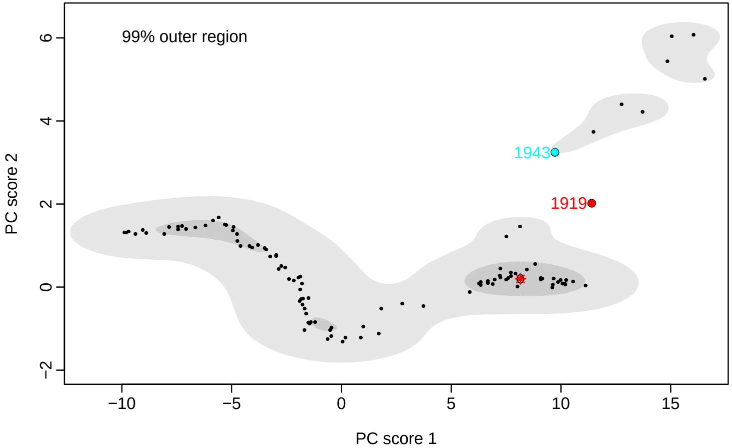


Functional bagplot



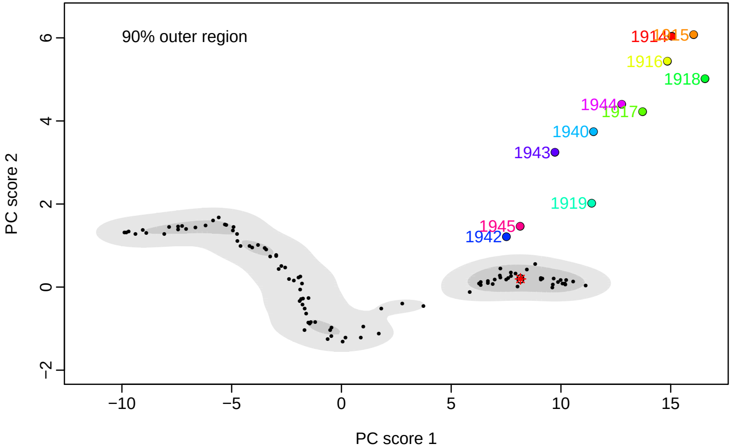
Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.



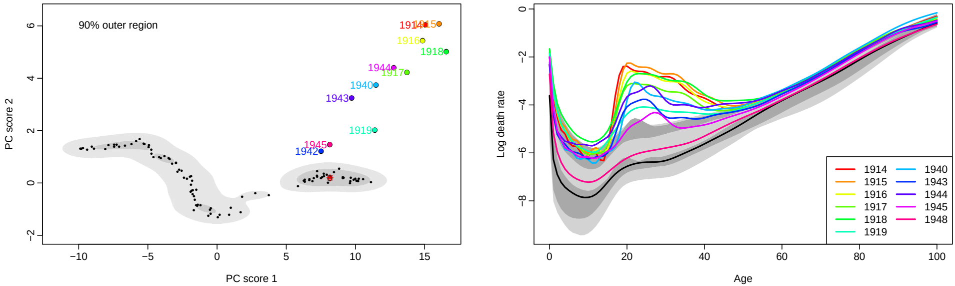
Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.

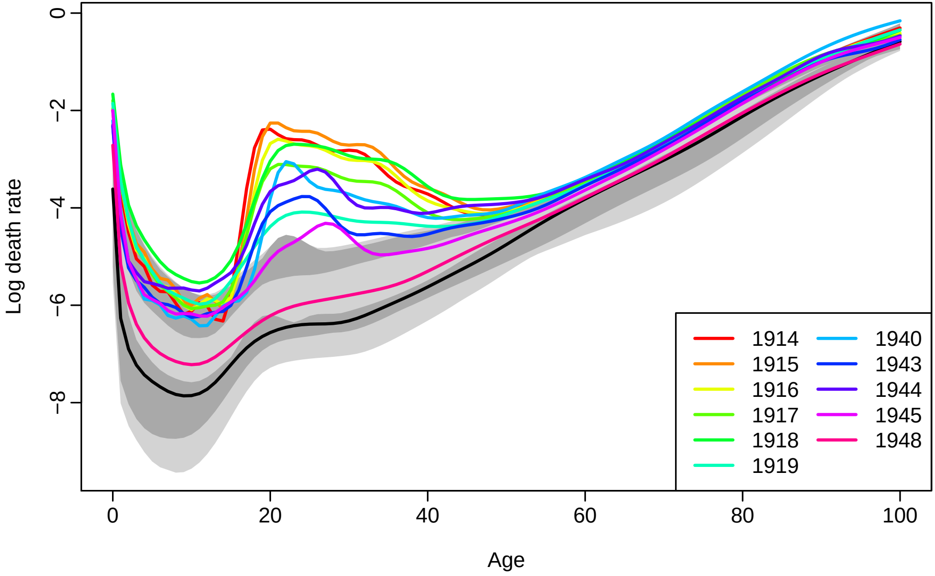


Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.
- Boundaries contain all curves inside HDRs.



Functional HDR boxplot



Outline

Functional time series model

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- The eigenfunctions $\phi_k(x)$ show the main regions of variation.
- The scores $\{\beta_{t,k}\}$ are uncorrelated by construction. So we can forecast each $\beta_{t,k}$ using a univariate time series model.
- Outliers are treated as missing values.
- Univariate ARIMA models are used for forecasting.

Functional time series model

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- The eigenfunctions $\phi_k(x)$ show the main regions of variation.
- The scores $\{\beta_{t,k}\}$ are uncorrelated by construction. So we can forecast each $\beta_{t,k}$ using a univariate time series model.
- Outliers are treated as missing values.
- Univariate ARIMA models are used for forecasting.

Functional time series model

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- The eigenfunctions $\phi_k(x)$ show the main regions of variation.
- The scores $\{\beta_{t,k}\}$ are uncorrelated by construction. So we can forecast each $\beta_{t,k}$ using a univariate time series model.
- Outliers are treated as missing values.
- Univariate ARIMA models are used for forecasting.

Functional time series model

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- The eigenfunctions $\phi_k(x)$ show the main regions of variation.
- The scores $\{\beta_{t,k}\}$ are uncorrelated by construction. So we can forecast each $\beta_{t,k}$ using a univariate time series model.
- Outliers are treated as missing values.
- Univariate ARIMA models are used for forecasting.

Functional time series model

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- The eigenfunctions $\phi_k(x)$ show the main regions of variation.
- The scores $\{\beta_{t,k}\}$ are uncorrelated by construction. So we can forecast each $\beta_{t,k}$ using a univariate time series model.
- Outliers are treated as missing values.
- Univariate ARIMA models are used for forecasting.

Forecasts

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

Forecasts

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

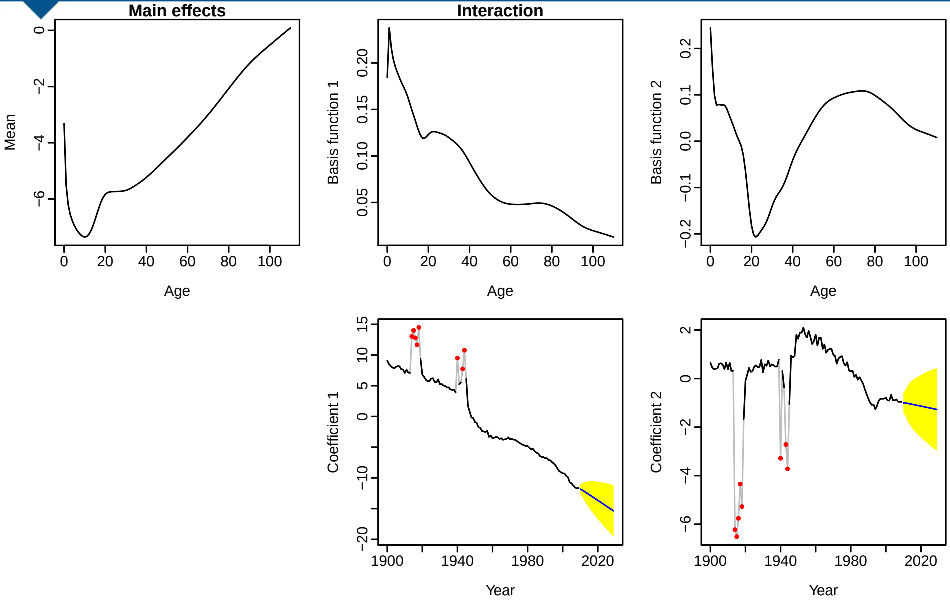
$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

$$\text{E}[y_{n+h,x} \mid \boldsymbol{y}] = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n+h,k}\,\hat{\phi}_k(x)$$

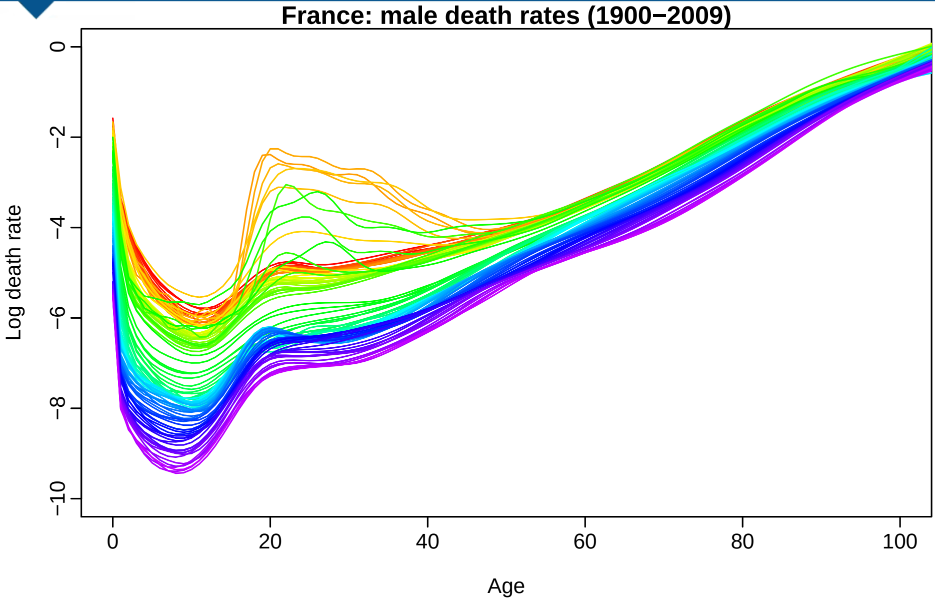
$$\text{Var}[y_{n+h,x} \mid \boldsymbol{y}] = \hat{\sigma}_\mu^2(x) + \sum_{k=1}^{K} v_{n+h,k}\,\hat{\phi}_k^2(x) + \sigma_t^2(x) + v(x)$$

where $v_{n+h,k} = \text{Var}(\beta_{n+h,k} \mid \beta_{1,k}, \dots, \beta_{n,k})$
and $\boldsymbol{y} = [y_{1,x}, \dots, y_{n,x}]$.

Forecasting the PC scores

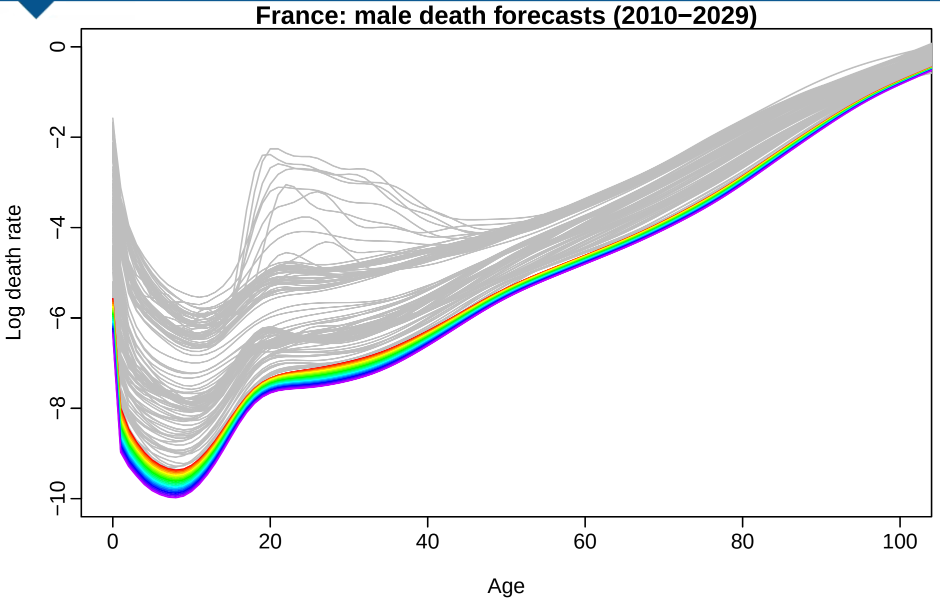


Forecasts of $f_t(x)$

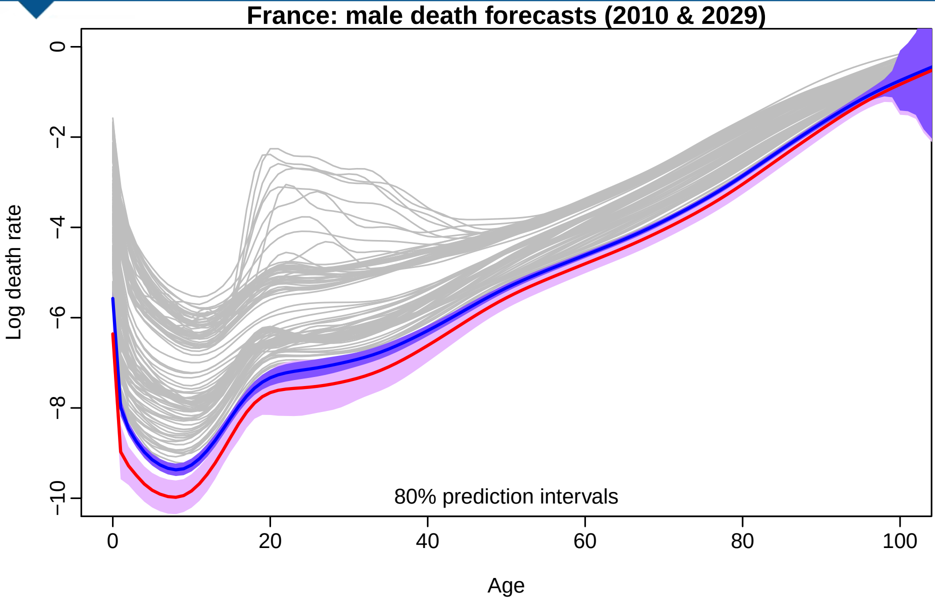


Forecasts of $f_t(x)$

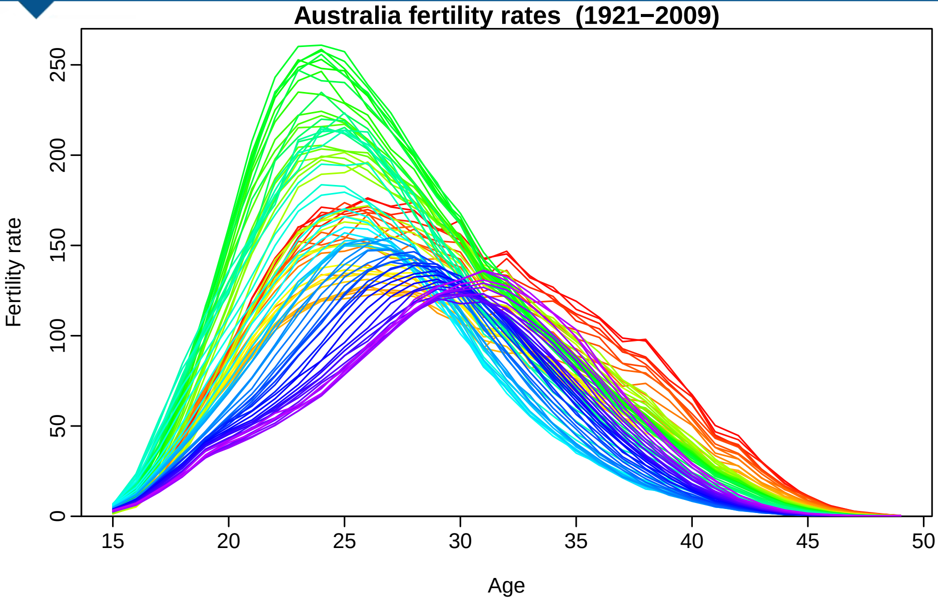
Forecasts of $f_t(x)$



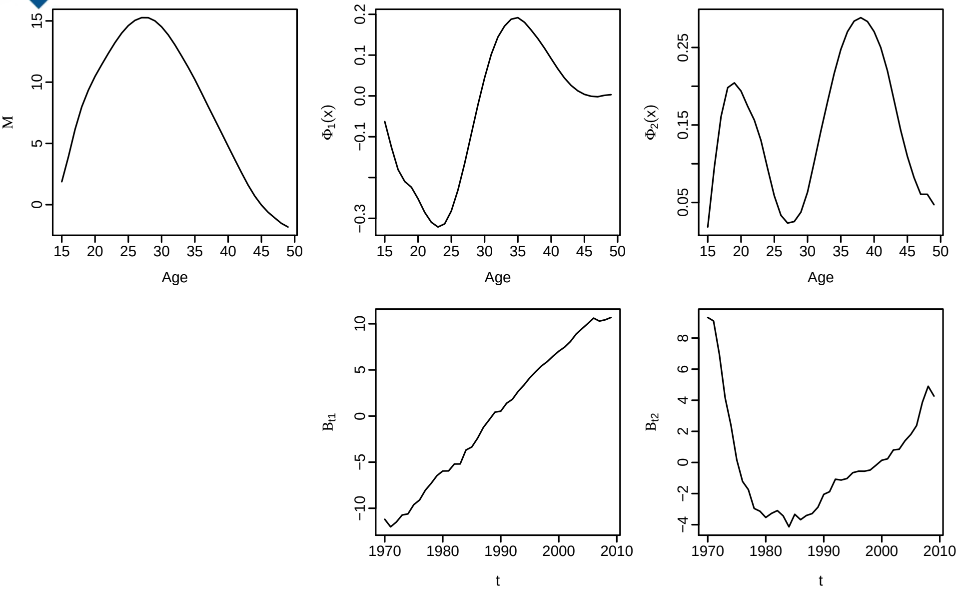
Forecasts of $f_t(x)$



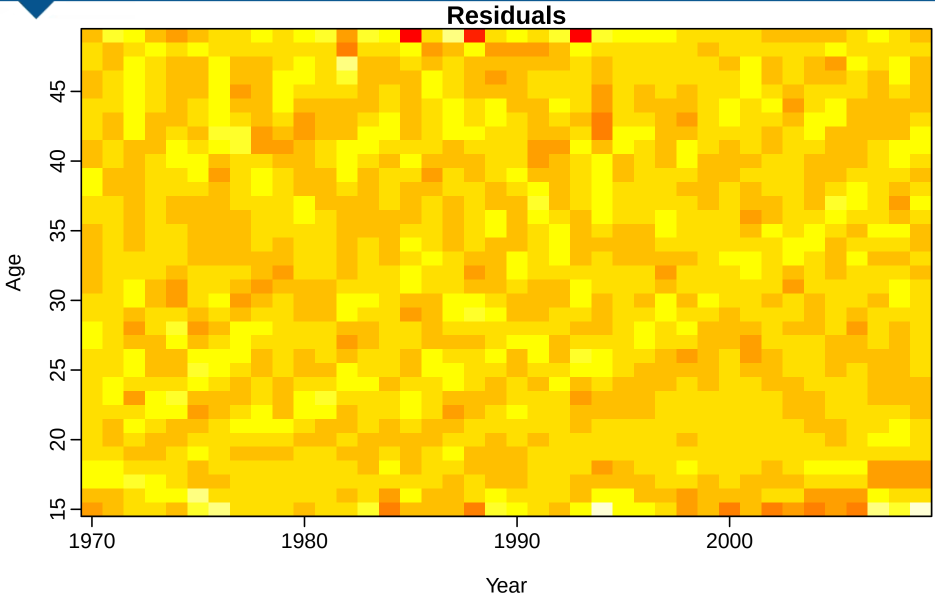
Fertility application



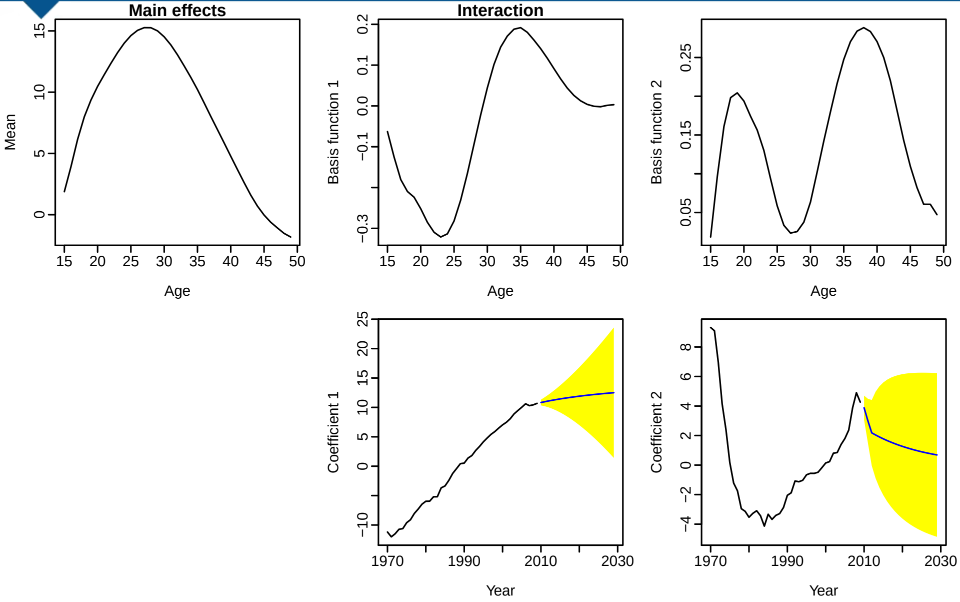
Fertility model



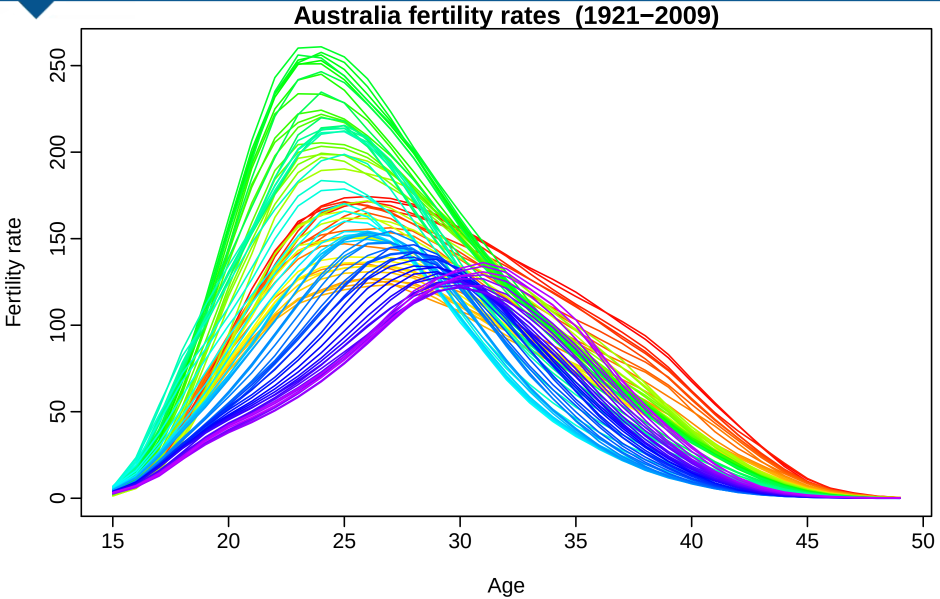
Fertility model



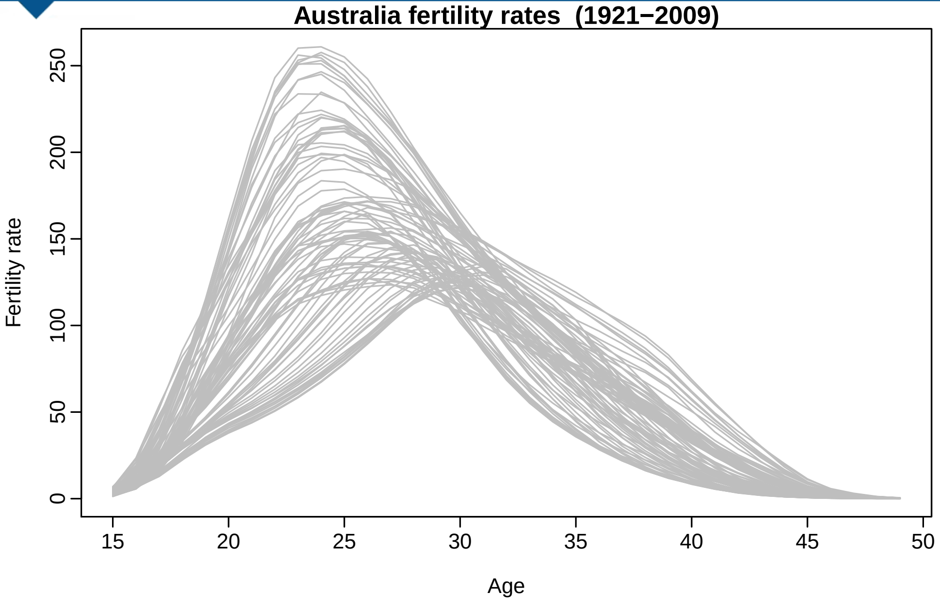
Fertility model



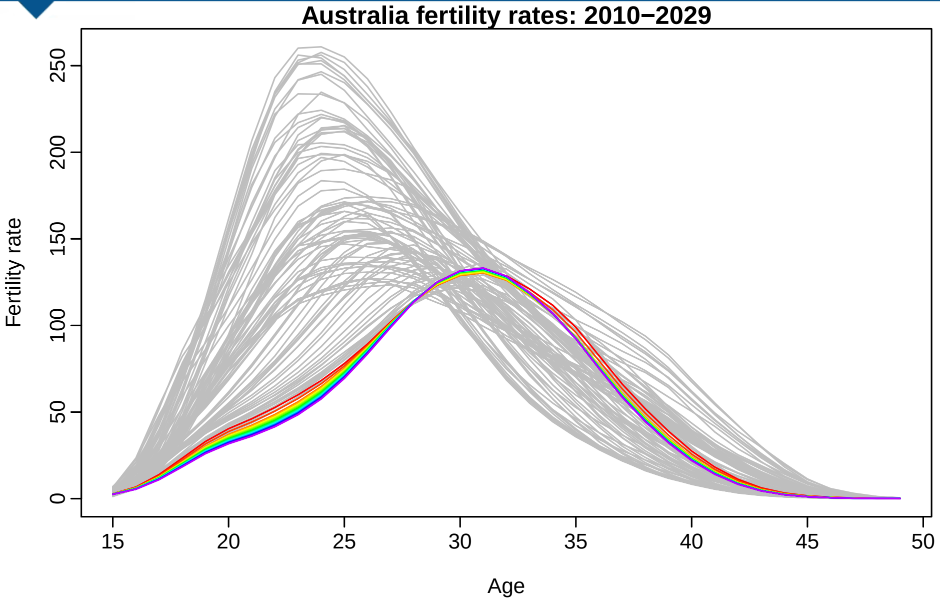
Forecasts of $f_t(x)$



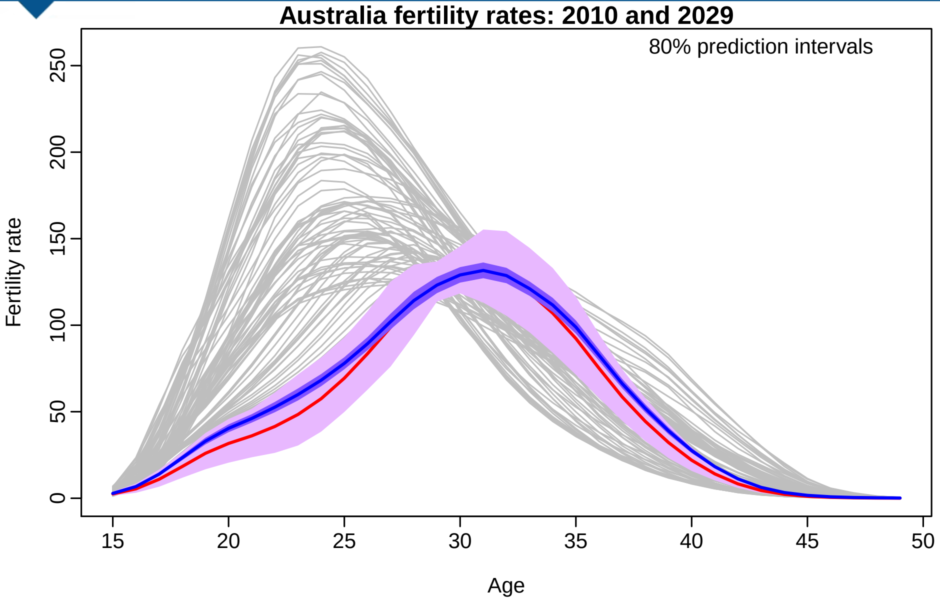
Forecasts of $f_t(x)$



Forecasts of $f_t(x)$



Forecasts of $f_t(x)$



Outline

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- Coherent forecasts do not diverge over time.
- Existing functional models do not impose coherence.

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- Coherent forecasts do not diverge over time.
- Existing functional models do not impose coherence.

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- **Coherent** forecasts do not diverge over time.
- Existing functional models do not impose coherence.

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- **Coherent** forecasts do not diverge over time.
- Existing functional models do not impose coherence.

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- **Coherent** forecasts do not diverge over time.
- Existing functional models do not impose coherence.

The problem

Let $f_{t,j}(x)$ be the smoothed mortality rate for age x in group j in year t.

- Groups may be males and females.
- Groups may be states within a country.
- Expected that groups will behave similarly.
- **Coherent** forecasts do not diverge over time.
- Existing functional models do not impose coherence.

Forecasting the coefficients

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- We use ARIMA models for each coefficient $\{\beta_{1,j,k}, \dots, \beta_{n,j,k}\}$.
- The ARIMA models are non-stationary for the first few coefficients ($k = 1, 2$)
- Non-stationary ARIMA forecasts will diverge. Hence the mortality forecasts are not coherent.

Forecasting the coefficients

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$
$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- We use ARIMA models for each coefficient $\{\beta_{1,j,k}, \dots, \beta_{n,j,k}\}$.
- The ARIMA models are non-stationary for the first few coefficients ($k = 1, 2$)
- Non-stationary ARIMA forecasts will diverge. Hence the mortality forecasts are not coherent.

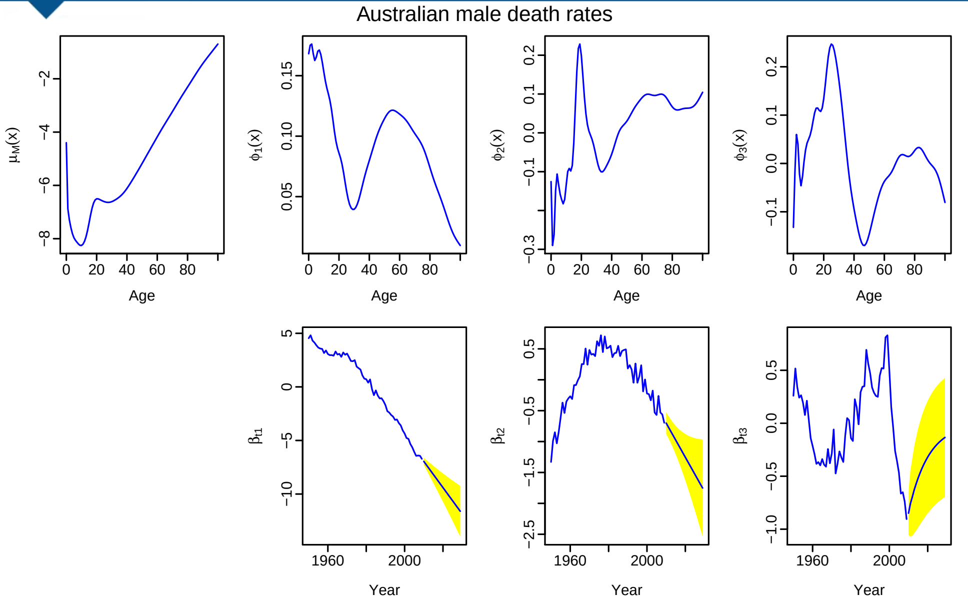
Forecasting the coefficients

$$y_{t,x} = f_t(x) + \sigma_t(x)\varepsilon_{t,x}$$

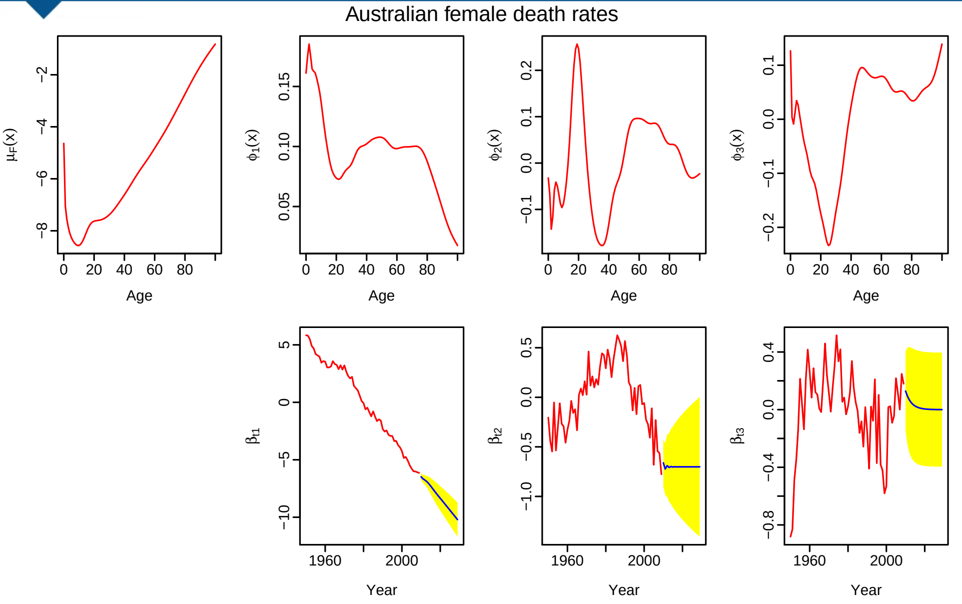
$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x)$$

- We use ARIMA models for each coefficient $\{\beta_{1,j,k}, \dots, \beta_{n,j,k}\}$.
- The ARIMA models are non-stationary for the first few coefficients ($k = 1, 2$)
- Non-stationary ARIMA forecasts will diverge. Hence the mortality forecasts are not coherent.

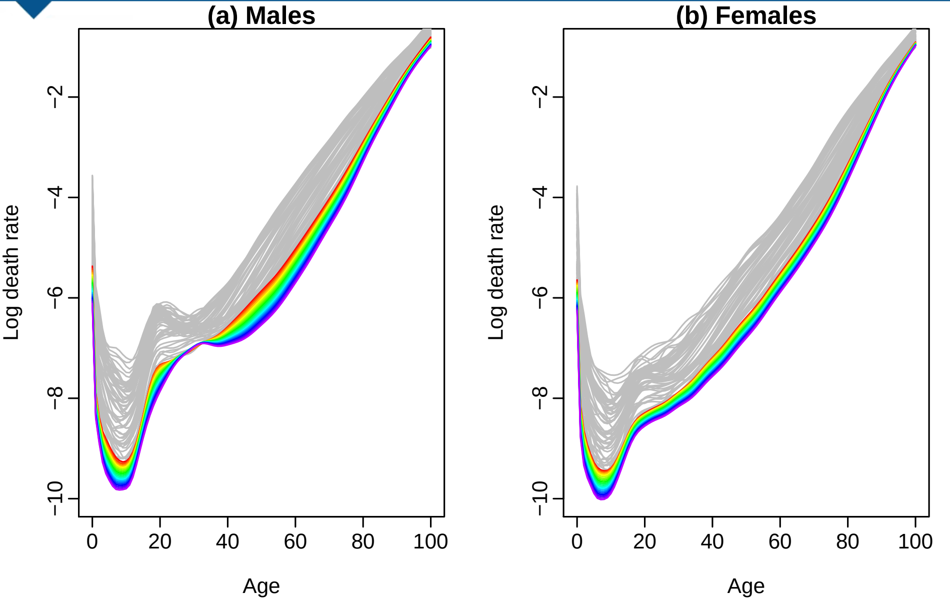
Male fts model



Female fts model



Australian mortality forecasts



Mortality product and ratios

Key idea

Model the geometric mean and the mortality ratio instead of the individual rates for each sex separately.

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

- Product and ratio are approximately independent
- Ratio should be stationary (for coherence) but product can be non-stationary

Mortality product and ratios

Key idea

Model the geometric mean and the mortality ratio instead of the individual rates for each sex separately.

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

- Product and ratio are approximately independent
- Ratio should be stationary (for coherence) but product can be non-stationary.

Mortality product and ratios

Key idea

Model the geometric mean and the mortality ratio instead of the individual rates for each sex separately.

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

- Product and ratio are approximately independent
- Ratio should be stationary (for coherence) but product can be non-stationary.

Mortality product and ratios

Key idea

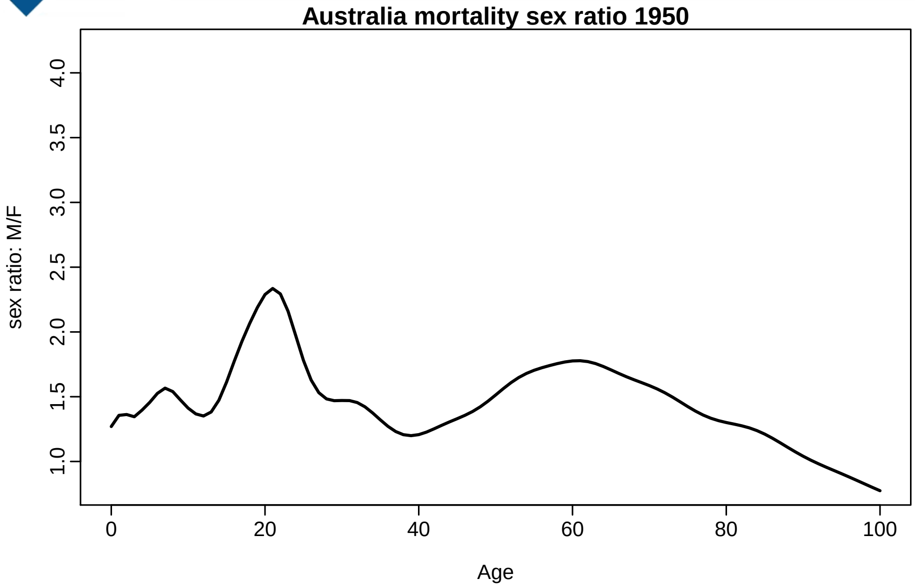
Model the geometric mean and the mortality ratio instead of the individual rates for each sex separately.

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

- Product and ratio are approximately independent
- Ratio should be stationary (for coherence) but product can be non-stationary.

Mortality rates

Mortality rates



Model product and ratios

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_t(x)] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell}\psi_\ell(x) + w_t(x).$$

- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.
- Forecasts: $f_{n+h|n,M}(x) = p_{n+h|n}(x) r_{n+h|n}(x)$

$$f_{n+h|n,F}(x) = p_{n+h|n}(x) \big/ r_{n+h|n}(x).$$

Model product and ratios

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_t(x)] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell}\psi_\ell(x) + w_t(x).$$

- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.
- Forecasts: $f_{n+h|n,M}(x) = p_{n+h|n}(x) r_{n+h|n}(x)$

  $f_{n+h|n,F}(x) = p_{n+h|n}(x) \big/ r_{n+h|n}(x).$

Model product and ratios

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(x) + e_t(x)$$

$$\log[r_t(x)] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell} \psi_\ell(x) + w_t(x).$$

- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p, q) or ARFIMA(p, d, q).
- No restrictions for $\beta_{t,1}, \dots, \beta_{t,K}$.
- Forecasts: $f_{n+h|n,M}(x) = p_{n+h|n}(x) r_{n+h|n}(x)$

  $f_{n+h|n,F}(x) = p_{n+h|n}(x) \big/ r_{n+h|n}(x).$

Model product and ratios

$$p_t(x) = \sqrt{f_{t,M}(x) f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x) \big/ f_{t,F}(x)}.$$

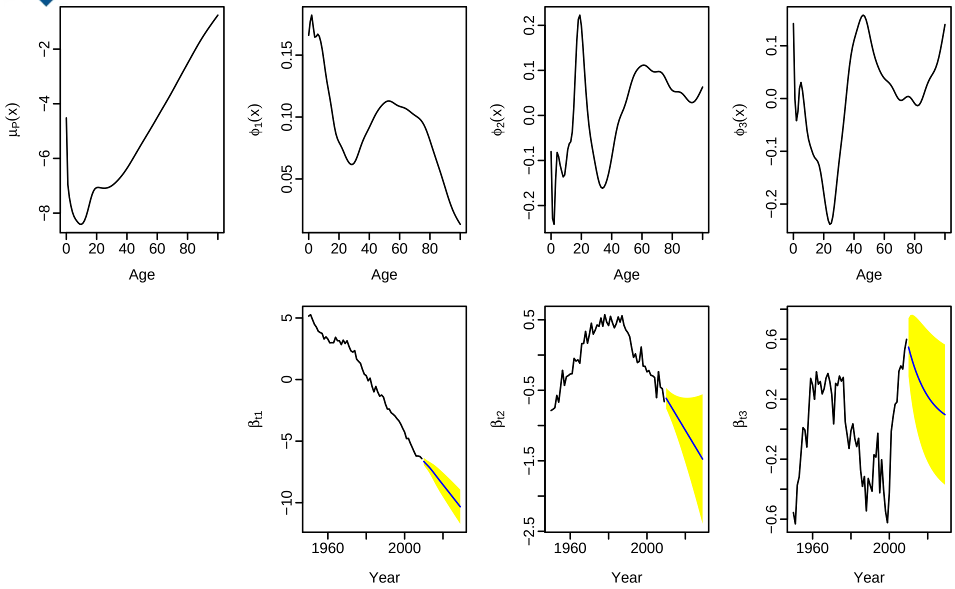
$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_t(x)] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell}\psi_\ell(x) + w_t(x).$$

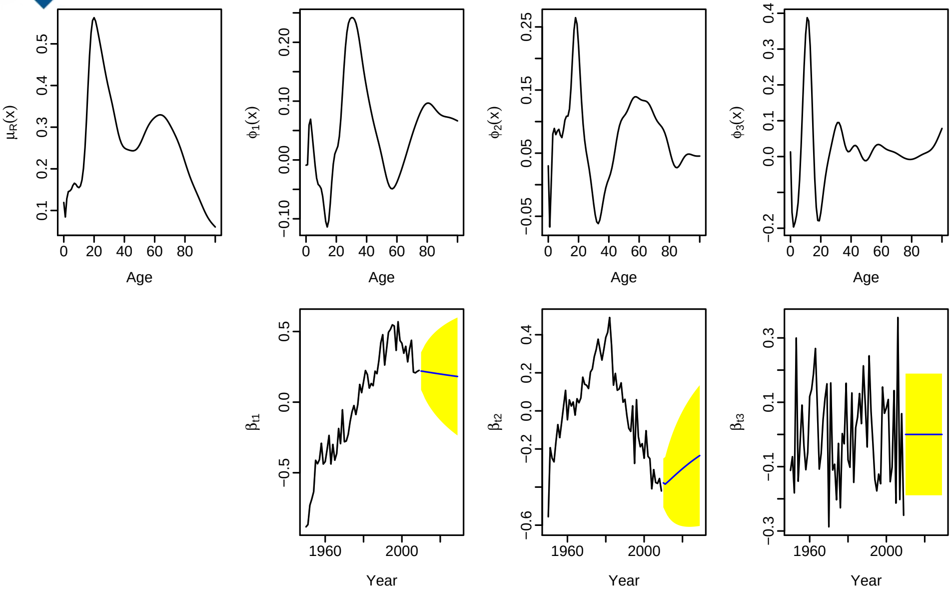
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.
- **Forecasts:** $f_{n+h|n,M}(x) = p_{n+h|n}(x) r_{n+h|n}(x)$

  $f_{n+h|n,F}(x) = p_{n+h|n}(x) \big/ r_{n+h|n}(x).$

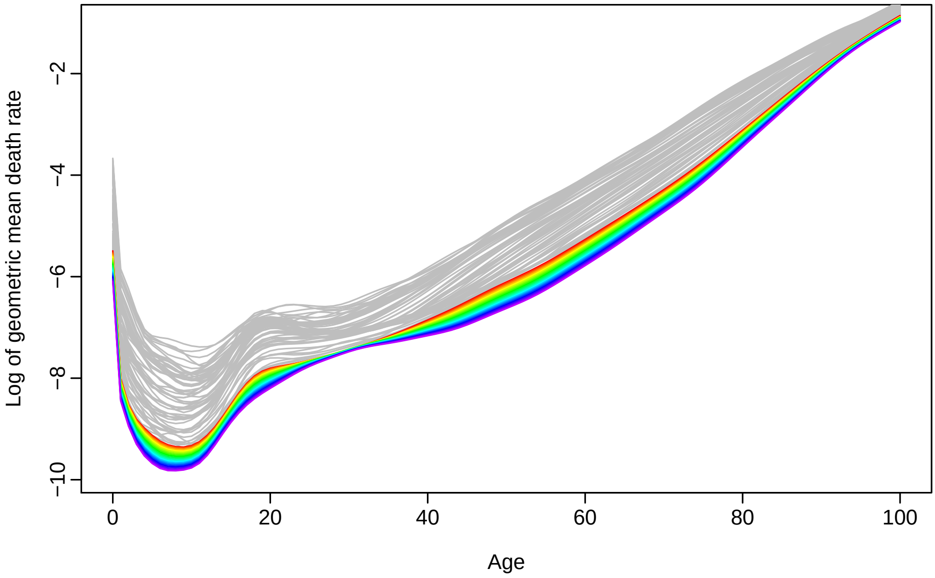
Product model



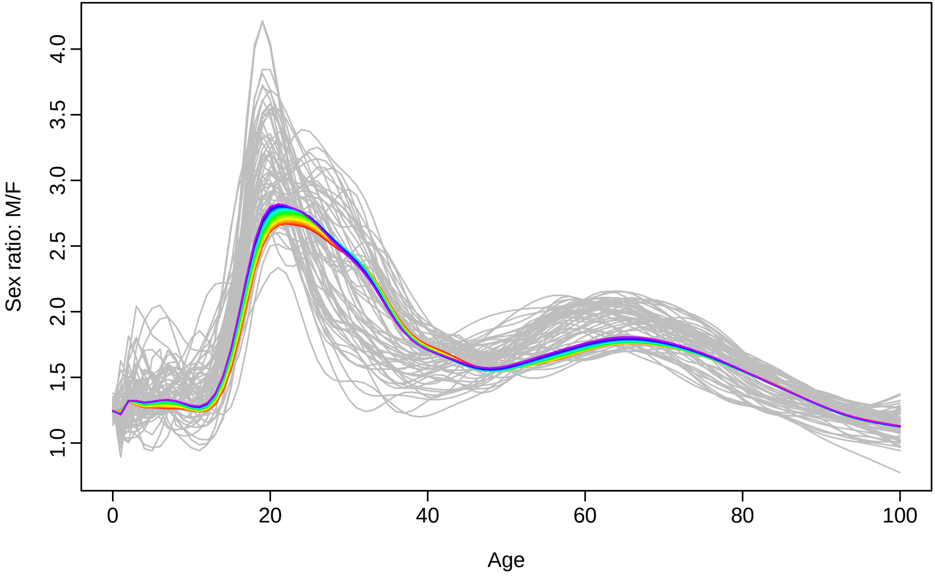
Ratio model



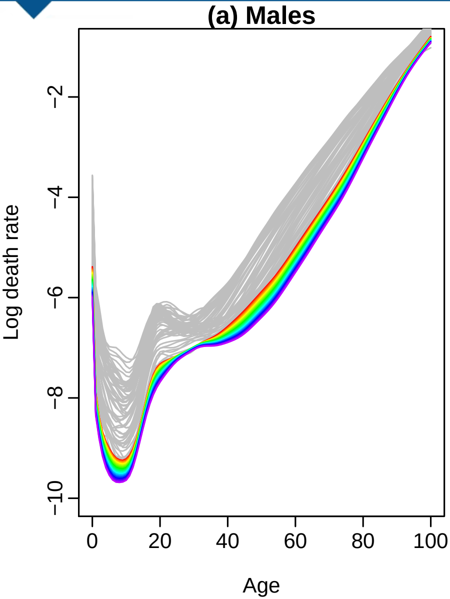
Product forecasts



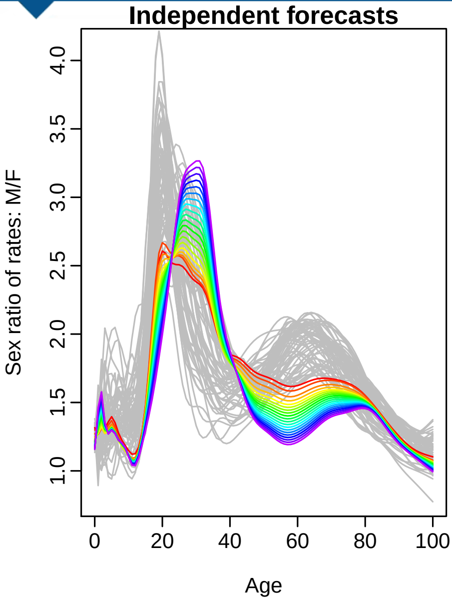
Ratio forecasts

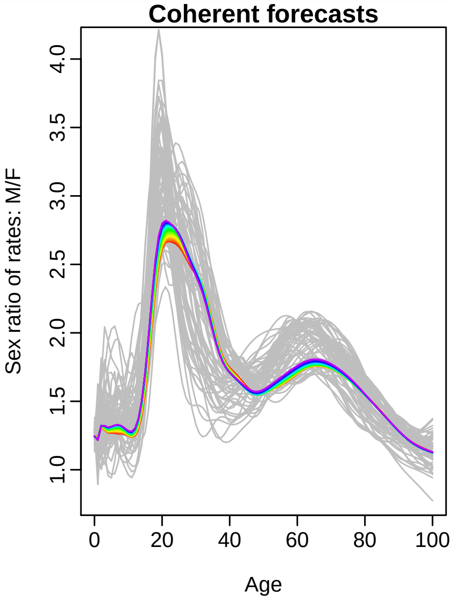


Coherent forecasts

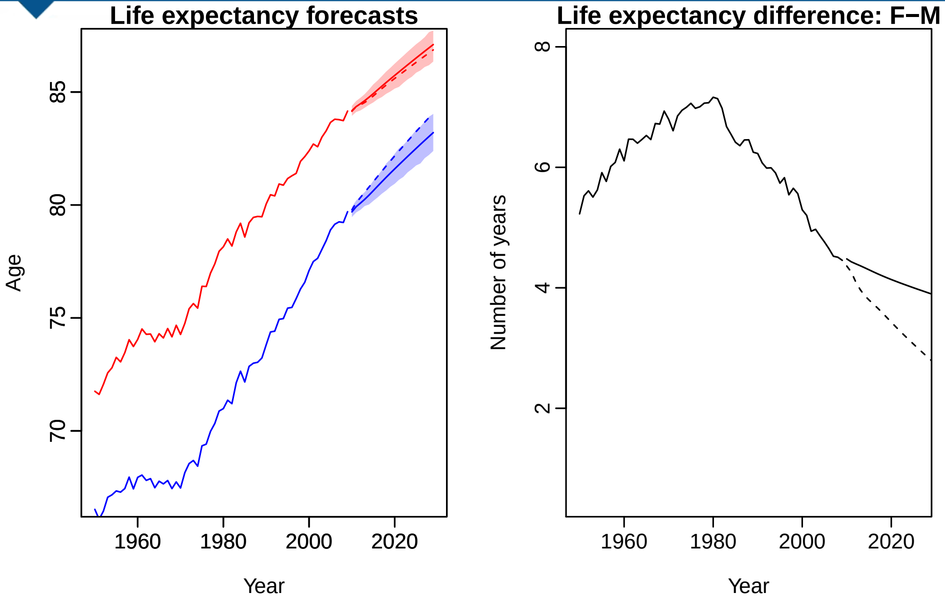


Ratio forecasts





Life expectancy forecasts



Coherent forecasts for J groups

$$p_t(x) = [f_{t,1}(x) f_{t,2}(x) \cdots f_{t,J}(x)]^{1/J}$$

and $$r_{t,j}(x) = f_{t,j}(x) \big/ p_t(x),$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L} \gamma_{t,l,j}\psi_{l,j}(x) + w_{t,j}(x).$$

- $p_t(x)$ and all $r_{t,j}(x)$ are approximately independent.
- Ratios satisfy constraint

Coherent forecasts for J groups

$$p_t(x) = [f_{t,1}(x) f_{t,2}(x) \cdots f_{t,J}(x)]^{1/J}$$

and $$r_{t,j}(x) = f_{t,j}(x) \big/ p_t(x),$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(x) + e_t(x)$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L} \gamma_{t,l,j} \psi_{l,j}(x) + w_{t,j}(x).$$

- $p_t(x)$ and all $r_{t,j}(x)$ are approximately independent.
- Ratios satisfy constraint $r_{t,1}(x) r_{t,2}(x) \cdots r_{t,J}(x) = 1$.
- $\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)]$

Coherent forecasts for J groups

$$p_t(x) = [f_{t,1}(x) f_{t,2}(x) \cdots f_{t,J}(x)]^{1/J}$$

and
$$r_{t,j}(x) = f_{t,j}(x) \big/ p_t(x),$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L} \gamma_{t,l,j}\psi_{l,j}(x) + w_{t,j}(x).$$

- $p_t(x)$ and all $r_{t,j}(x)$ are approximately independent.
- Ratios satisfy constraint $r_{t,1}(x) r_{t,2}(x) \cdots r_{t,J}(x) = 1$.
- $\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)]$

Coherent forecasts for J groups

$$p_t(x) = [f_{t,1}(x) f_{t,2}(x) \cdots f_{t,J}(x)]^{1/J}$$

and $$r_{t,j}(x) = f_{t,j}(x) \big/ p_t(x),$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(x) + e_t(x)$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L} \gamma_{t,l,j} \psi_{l,j}(x) + w_{t,j}(x).$$

- $p_t(x)$ and all $r_{t,j}(x)$ are approximately independent.
- Ratios satisfy constraint $r_{t,1}(x) r_{t,2}(x) \cdots r_{t,J}(x) = 1$.
- $\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)]$

Coherent forecasts for J groups

$$p_t(x) = [f_{t,1}(x)f_{t,2}(x)\cdots f_{t,J}(x)]^{1/J}$$

and $$r_{t,j}(x) = f_{t,j}(x)\big/p_t(x),$$

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K}\beta_{t,k}\phi_k(x) + e_t(x)$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L}\gamma_{t,l,j}\psi_{l,j}(x) + w_{t,j}(x).$$

- $p_t(x)$ and all $r_{t,j}(x)$ are approximately independent.
- Ratios satisfy constraint $r_{t,1}(x)r_{t,2}(x)\cdots r_{t,J}(x) = 1$.
- $\log[f_{t,j}(x)] = \log[p_t(x)r_{t,j}(x)]$

Coherent forecasts for J groups

$$\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)] = \log[p_t(x)] + \log[r_{t,j}]$$
$$= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j}\psi_{\ell,j}(x) + z_{t,j}(x)$$

- $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is group mean
- $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is error term.
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.

Coherent forecasts for J groups

$$\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)] = \log[p_t(x)] + \log[r_{t,j}]$$
$$= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j}\psi_{\ell,j}(x) + z_{t,j}(x)$$

- $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is group mean
- $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is error term.
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.

Coherent forecasts for J groups

$$\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)] = \log[p_t(x)] + \log[r_{t,j}]$$
$$= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j}\psi_{\ell,j}(x) + z_{t,j}(x)$$

- $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is group mean
- $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is error term.
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.

Coherent forecasts for J groups

$$\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)] = \log[p_t(x)] + \log[r_{t,j}]$$
$$= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j}\psi_{\ell,j}(x) + z_{t,j}(x)$$

- $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is group mean
- $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is error term.
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p, q) or ARFIMA(p, d, q).
- No restrictions for $\beta_{t,1}, \ldots, \beta_{t,K}$.

Coherent forecasts for J groups

$$\log[f_{t,j}(x)] = \log[p_t(x) r_{t,j}(x)] = \log[p_t(x)] + \log[r_{t,j}]$$
$$= \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j}\psi_{\ell,j}(x) + z_{t,j}(x)$$

- $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is group mean
- $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is error term.
- $\{\gamma_{t,\ell}\}$ restricted to be stationary processes: either ARMA(p,q) or ARFIMA(p,d,q).
- No restrictions for $\beta_{t,1},\dots,\beta_{t,K}$.

Outline

Demographic growth-balance equation

Demographic growth-balance equation

$$P_{t+1}(x+1) = P_t(x) - D_t(x,x+1) + G_t(x,x+1)$$
$$P_{t+1}(0) = B_t \quad - D_t(B,0) \quad + G_t(B,0)$$

$x = 0,1,2,\ldots.$

$P_t(x) =$ population of age x at 1 January, year t

$B_t =$ births in calendar year t

$D_t(x,x+1) =$ deaths in calendar year t of persons aged x at the beginning of year t

$D_t(B,0) =$ infant deaths in calendar year t

$G_t(x,x+1) =$ net migrants in calendar year t of persons aged x at the beginning of year t

$G_t(B,0) =$ net migrants of infants born in calendar year t

Demographic growth-balance equation

Demographic growth-balance equation

$$P_{t+1}(x+1) = P_t(x) - D_t(x,x+1) + G_t(x,x+1)$$
$$P_{t+1}(0) = B_t \quad - D_t(B,0) \quad + G_t(B,0)$$

$x = 0, 1, 2, \ldots.$

$P_t(x) =$	population of age x at 1 January, year t
$B_t =$	births in calendar year t
$D_t(x,x+1) =$	deaths in calendar year t of persons aged x at the beginning of year t
$D_t(B,0) =$	infant deaths in calendar year t
$G_t(x,x+1) =$	net migrants in calendar year t of persons aged x at the beginning of year t
$G_t(B,0) =$	net migrants of infants born in calendar year t

Key ideas

- Build a **stochastic functional model** for each of mortality, fertility and net migration.
- Treat all observed data as **functional** (i.e., smooth curves of age) rather than discrete values.
- Use the models to **simulate future sample paths** of all components giving the entire age distribution at every year into the future.
- Compute future births, deaths, net migrants. and populations from simulated rates.
- Combine the results to get **age-specific stochastic population forecasts.**

Key ideas

- Build a **stochastic functional model** for each of mortality, fertility and net migration.
- Treat all observed data as **functional** (i.e., smooth curves of age) rather than discrete values.
- Use the models to **simulate future sample paths** of all components giving the entire age distribution at every year into the future.
- Compute future births, deaths, net migrants. and populations from simulated rates.
- Combine the results to get **age-specific stochastic population forecasts**.

Key ideas

- Build a **stochastic functional model** for each of mortality, fertility and net migration.
- Treat all observed data as **functional** (i.e., smooth curves of age) rather than discrete values.
- Use the models to **simulate future sample paths** of all components giving the entire age distribution at every year into the future.
- Compute future births, deaths, net migrants. and populations from simulated rates.
- Combine the results to get **age-specific stochastic population forecasts**.

Key ideas

- Build a **stochastic functional model** for each of mortality, fertility and net migration.
- Treat all observed data as **functional** (i.e., smooth curves of age) rather than discrete values.
- Use the models to **simulate future sample paths** of all components giving the entire age distribution at every year into the future.
- Compute future births, deaths, net migrants. and populations from simulated rates.
- Combine the results to get **age-specific stochastic population forecasts**.

Key ideas

- Build a **stochastic functional model** for each of mortality, fertility and net migration.
- Treat all observed data as **functional** (i.e., smooth curves of age) rather than discrete values.
- Use the models to **simulate future sample paths** of all components giving the entire age distribution at every year into the future.
- Compute future births, deaths, net migrants. and populations from simulated rates.
- Combine the results to get **age-specific stochastic population forecasts.**

The available data

In most countries, the following data are available:

$P_t(x) =$ **population** of age x at 1 January, year t
$E_t(x) =$ **population** of age x at 30 June, year t
$B_t(x) =$ **births** in calendar year t to females of age x
$D_t(x) =$ **deaths** in calendar year t of persons of age x

From these, we can estimate:

- $m_t(x) = D_t(x)/E_t(x) =$ central death rate in calendar year t;
- $f_t(x) = B_t(x)/E_t^F(x) =$ fertility rate for females of age x in calendar year t.

The available data

In most countries, the following data are available:

$P_t(x) =$ **population** of age x at 1 January, year t
$E_t(x) =$ **population** of age x at 30 June, year t
$B_t(x) =$ **births** in calendar year t to females of age x
$D_t(x) =$ **deaths** in calendar year t of persons of age x

From these, we can estimate:

- $m_t(x) = D_t(x)/E_t(x) =$ central death rate in calendar year t;
- $f_t(x) = B_t(x)/E_t^F(x) =$ fertility rate for females of age x in calendar year t.

The available data

In most countries, the following data are available:

$P_t(x) =$ **population** of age x at 1 January, year t
$E_t(x) =$ **population** of age x at 30 June, year t
$B_t(x) =$ **births** in calendar year t to females of age x
$D_t(x) =$ **deaths** in calendar year t of persons of age x

From these, we can estimate:

- $m_t(x) = D_t(x)/E_t(x) =$ central death rate in calendar year t;
- $f_t(x) = B_t(x)/E_t^F(x) =$ fertility rate for females of age x in calendar year t.

Net migration

- We need to *estimate* **migration** data based on difference in population numbers after adjusting for births and deaths.

Demographic growth-balance equation

$$G_t(x,x+1) = P_{t+1}(x+1) - P_t(x) + D_t(x,x+1)$$
$$G_t(B,0) = P_{t+1}(0) \qquad - B_t \qquad + D_t(B,0)$$

$x = 0,1,2,\ldots$

Note: "net migration" numbers also include errors associated with all estimates. i.e., a "residual".

Net migration

- We need to *estimate* migration data based on difference in population numbers after adjusting for births and deaths.

Demographic growth-balance equation

$$G_t(x,x+1) = P_{t+1}(x+1) - P_t(x) + D_t(x,x+1)$$
$$G_t(B,0) = P_{t+1}(0) \quad - B_t \quad + D_t(B,0)$$

$x = 0,1,2,\ldots.$

Note: "net migration" numbers also include errors associated with all estimates. i.e., a "residual".

Net migration

- We need to *estimate* **migration** data based on difference in population numbers after adjusting for births and deaths.

Demographic growth-balance equation

$$G_t(x,x+1) = P_{t+1}(x+1) - P_t(x) + D_t(x,x+1)$$
$$G_t(B,0) = P_{t+1}(0) \qquad - B_t \qquad + D_t(B,0)$$

$x = 0,1,2,\dots.$

Note: “net migration” numbers also include **errors** associated with all estimates. i.e., a “residual”.

Net migration

- We need to *estimate* migration data based on difference in population numbers after adjusting for births and deaths.

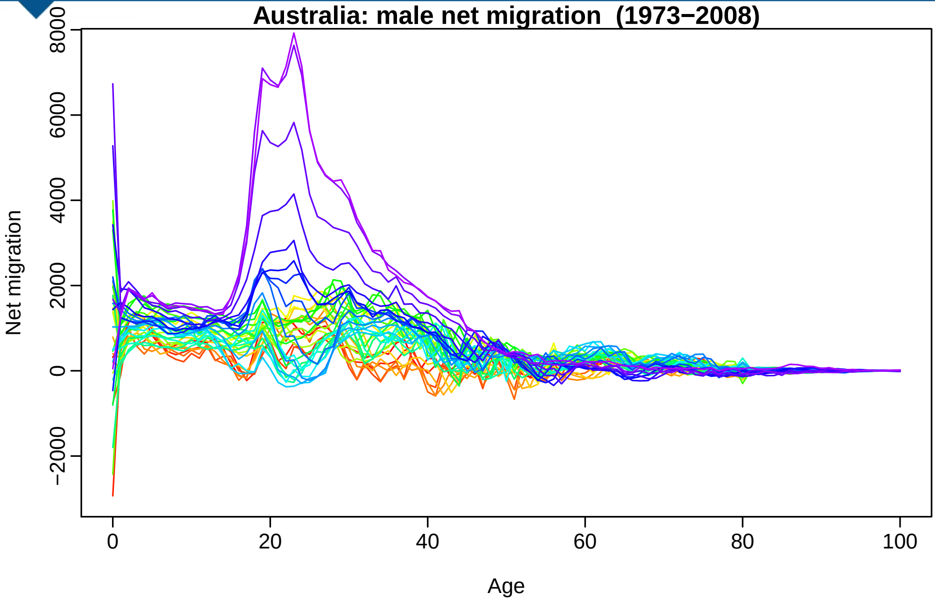
Demographic growth-balance equation

$$G_t(x, x+1) = P_{t+1}(x+1) - P_t(x) + D_t(x, x+1)$$
$$G_t(B, 0) = P_{t+1}(0) \qquad - B_t \qquad + D_t(B, 0)$$

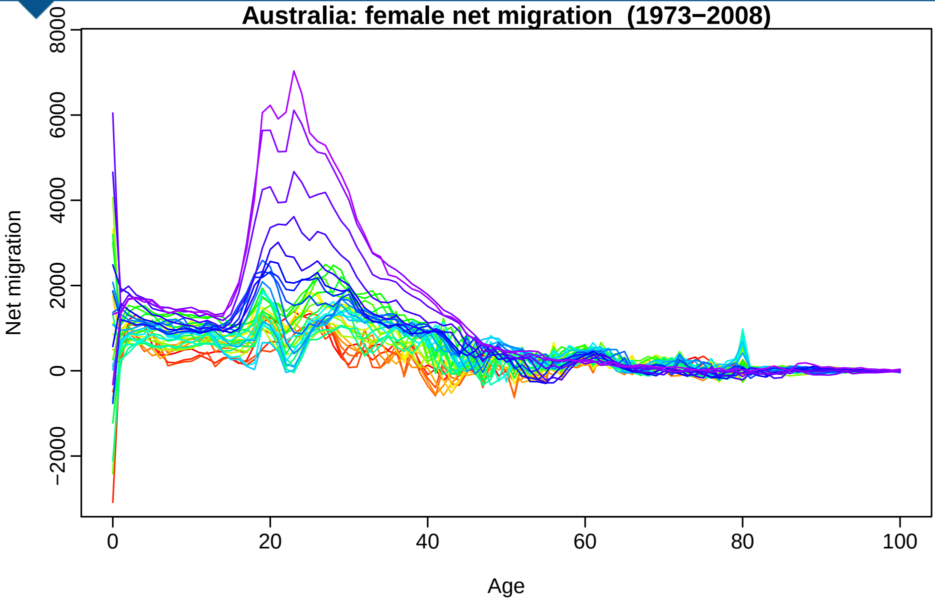
$x = 0, 1, 2, \ldots$

Note: "net migration" numbers also include errors associated with all estimates. i.e., a "residual".

Net migration



Net migration



Stochastic population forecasts

Component models

- Data: age/sex-specific mortality rates, fertility rates and net migration.
- Models: Functional time series models for mortality (M/F), fertility and net migration (M/F) assuming independence between components and coherence between sexes.
- Generate random sample paths of each component conditional on observed data.
- Use simulated rates to generate $B_t(x)$, $D_t^F(x, x+1)$, $D_t^M(x, x+1)$ for $t = n+1, \dots, n+h$, assuming deaths and births are Poisson.

Stochastic population forecasts

Component models

- Data: age/sex-specific mortality rates, fertility rates and net migration.
- Models: Functional time series models for mortality (M/F), fertility and net migration (M/F) assuming independence between components and coherence between sexes.
- Generate random sample paths of each component conditional on observed data.
- Use simulated rates to generate $B_t(x)$, $D_t^F(x,x+1)$, $D_t^M(x,x+1)$ for $t = n+1,\dots,n+h$, assuming deaths and births are Poisson.

Stochastic population forecasts

Component models

- Data: age/sex-specific mortality rates, fertility rates and net migration.
- Models: Functional time series models for mortality (M/F), fertility and net migration (M/F) assuming independence between components and coherence between sexes.
- Generate random sample paths of each component conditional on observed data.
- Use simulated rates to generate $B_t(x)$, $D_t^F(x,x+1)$, $D_t^M(x,x+1)$ for $t = n+1,\dots,n+h$, assuming deaths and births are Poisson.

Stochastic population forecasts

Component models

- Data: age/sex-specific mortality rates, fertility rates and net migration.
- Models: Functional time series models for mortality (M/F), fertility and net migration (M/F) assuming independence between components and coherence between sexes.
- Generate random sample paths of each component conditional on observed data.
- Use simulated rates to generate $B_t(x)$, $D_t^F(x,x+1)$, $D_t^M(x,x+1)$ for $t = n+1,\dots,n+h$, assuming deaths and births are Poisson.

Simulation

Demographic growth-balance equation used to get population sample paths.

Demographic growth-balance equation

$$
\begin{aligned}
P_{t+1}(x+1) &= P_t(x) - D_t(x,x+1) + G_t(x,x+1) \\
P_{t+1}(0) &= B_t \quad - D_t(B,0) \quad + G_t(B,0)
\end{aligned}
$$

$x = 0, 1, 2, \ldots.$

- 10000 sample paths of population $P_t(x)$, deaths $D_t(x)$ and births $B_t(x)$ generated for $t = 2004, \ldots, 2023$ and $x = 0, 1, 2, \ldots,$.
- This allows the computation of the empirical forecast distribution of any demographic quantity that is based on births, deaths and population numbers.

Simulation

Demographic growth-balance equation used to get population sample paths.

Demographic growth-balance equation

$$P_{t+1}(x+1) = P_t(x) - D_t(x,x+1) + G_t(x,x+1)$$
$$P_{t+1}(0) = B_t \quad - D_t(B,0) \quad + G_t(B,0)$$

$x = 0, 1, 2, \ldots.$

- 10000 sample paths of population $P_t(x)$, deaths $D_t(x)$ and births $B_t(x)$ generated for $t = 2004, \ldots, 2023$ and $x = 0, 1, 2, \ldots,$.
- This allows the computation of the empirical forecast distribution of any demographic quantity that is based on births, deaths and population numbers.

Simulation

Demographic growth-balance equation used to get population sample paths.

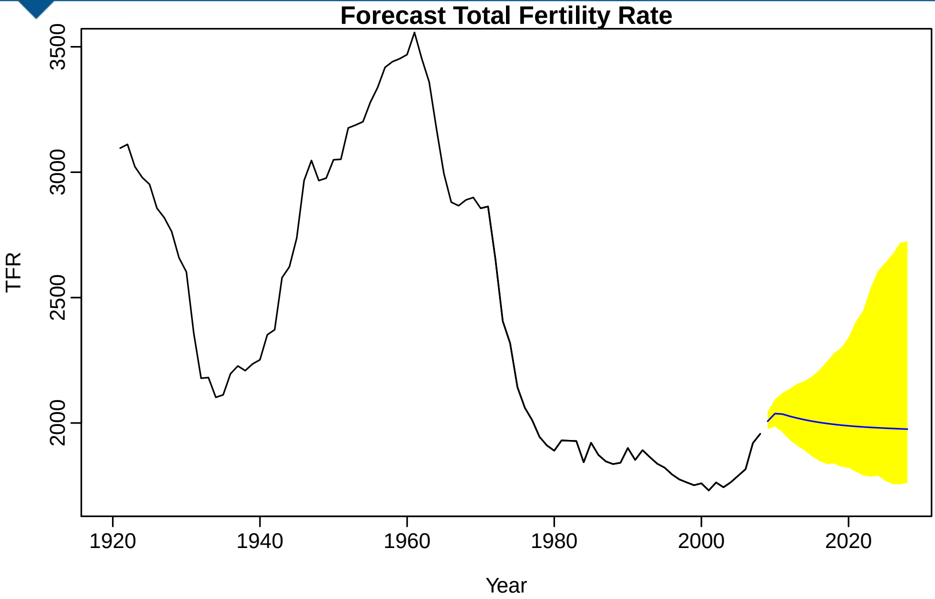
Demographic growth-balance equation

$$P_{t+1}(x+1) = P_t(x) - D_t(x, x+1) + G_t(x, x+1)$$
$$P_{t+1}(0) = B_t \quad - D_t(B, 0) \quad + G_t(B, 0)$$

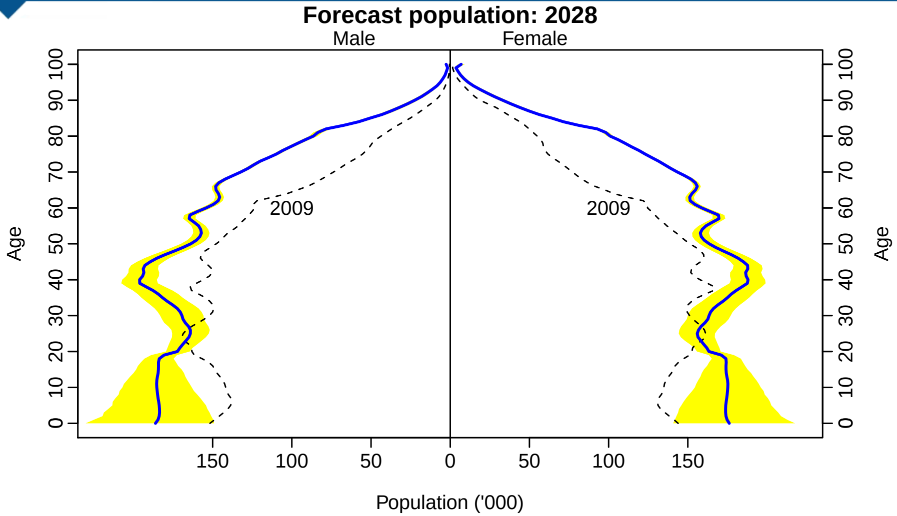
$x = 0, 1, 2, \ldots.$

- 10000 sample paths of population $P_t(x)$, deaths $D_t(x)$ and births $B_t(x)$ generated for $t = 2004, \ldots, 2023$ and $x = 0, 1, 2, \ldots,$.
- This allows the computation of the empirical forecast distribution of any demographic quantity that is based on births, deaths and population numbers.

Forecasts of TFR

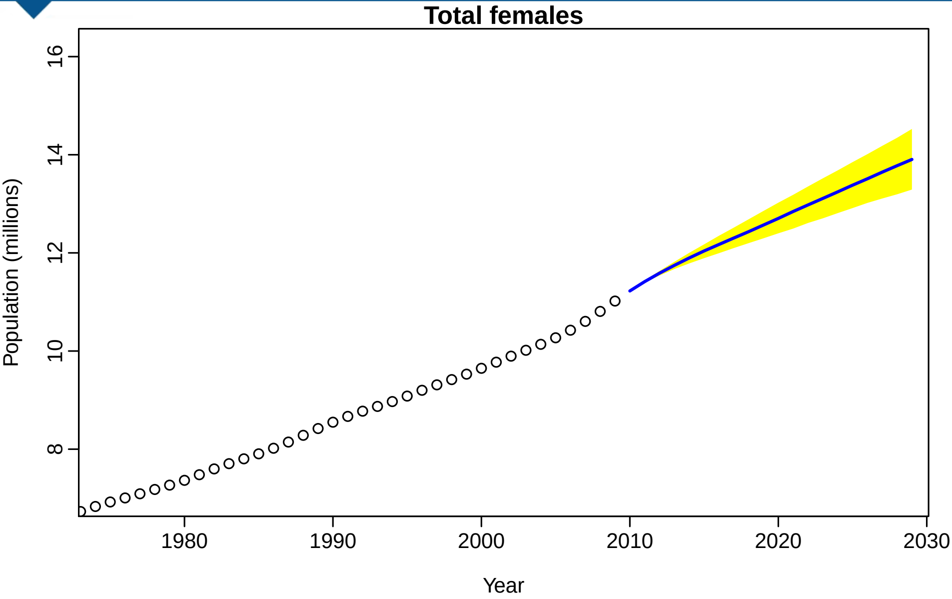


Population forecasts

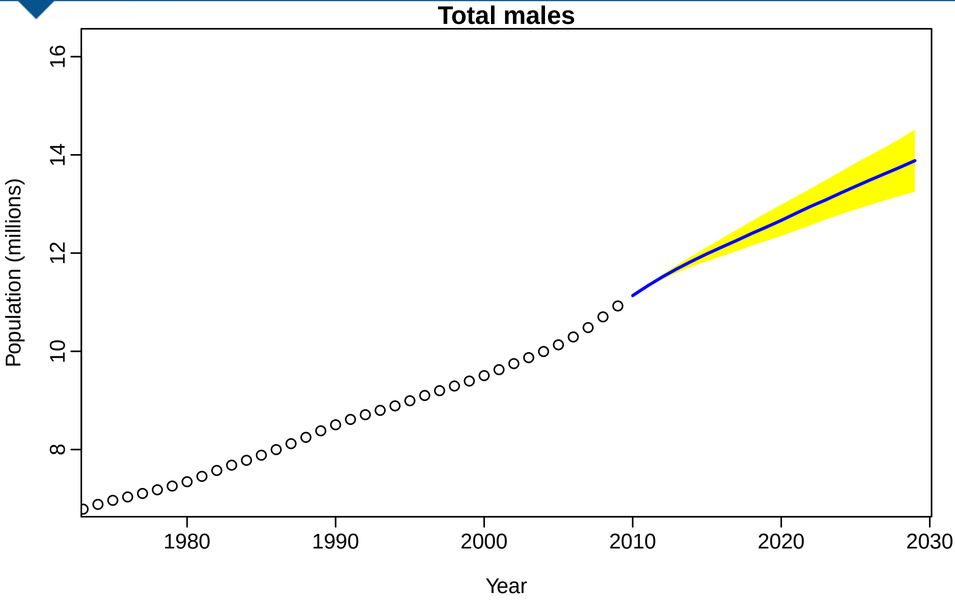


Forecast population pyramid for 2028, along with 80% prediction intervals. Dashed: actual population pyramid for 2009.

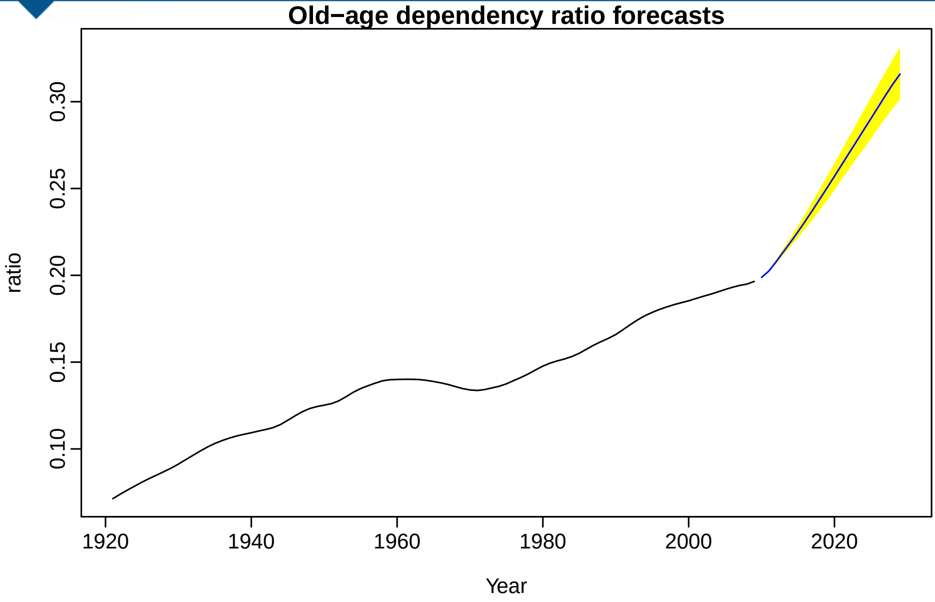
Population forecasts



Population forecasts



Old-age dependency ratio



Advantages of stochastic simulation approach

- Functional data analysis provides a way of forecasting age-specific mortality, fertility and net migration.
- Stochastic age-specific cohort-component simulation provides a way of forecasting many demographic quantities with prediction intervals.
- No need to select combinations of assumed rates.
- True prediction intervals with specified coverage for population and all derived variables (TFR, life expectancy, old-age dependencies, etc.)

Advantages of stochastic simulation approach

- Functional data analysis provides a way of forecasting age-specific mortality, fertility and net migration.
- Stochastic age-specific cohort-component simulation provides a way of forecasting many demographic quantities with prediction intervals.
- No need to select combinations of assumed rates.
- True prediction intervals with specified coverage for population and all derived variables (TFR, life expectancy, old-age dependencies, etc.)

Advantages of stochastic simulation approach

- Functional data analysis provides a way of forecasting age-specific mortality, fertility and net migration.
- Stochastic age-specific cohort-component simulation provides a way of forecasting many demographic quantities with prediction intervals.
- No need to select combinations of assumed rates.
- True prediction intervals with specified coverage for population and all derived variables (TFR, life expectancy, old-age dependencies, etc.)

Advantages of stochastic simulation approach

- Functional data analysis provides a way of forecasting age-specific mortality, fertility and net migration.
- Stochastic age-specific cohort-component simulation provides a way of forecasting many demographic quantities with prediction intervals.
- No need to select combinations of assumed rates.
- True prediction intervals with specified coverage for population and all derived variables (TFR, life expectancy, old-age dependencies, etc.)

Outline

Selected references

- Hyndman, Shang (2010). "Rainbow plots, bagplots and boxplots for functional data". *Journal of Computational and Graphical Statistics* **19**(1), 29–45
- Hyndman, Ullah (2007). "Robust forecasting of mortality and fertility rates: A functional data approach". *Computational Statistics and Data Analysis* **51**(10), 4942–4956
- Hyndman, Shang (2009). "Forecasting functional time series (with discussion)". *Journal of the Korean Statistical Society* **38**(3), 199–221
- Shang, Booth, Hyndman (2011). "Point and interval forecasts of mortality rates and life expectancy : a comparison of ten principal component methods". *Demographic Research* **25**(5), 173–214
- Hyndman, Booth (2008). "Stochastic population forecasts using functional data models for mortality, fertility and migration". *International Journal of Forecasting* **24**(3), 323–342
- Hyndman, Booth, Yasmeen (2012). "Coherent mortality forecasting: the product-ratio method with functional time series models". *Demography*, to appear
- Hyndman (2012). *demography: Forecasting mortality, fertility, migration and population data.* `cran.r-project.org/package=demography`

Selected references

- Hyndman, Shang (2010). "Rainbow plots, bagplots and boxplots for functional data". *Journal of Computational and Graphical Statistics* **19**(1), 29–45
- Hyndman, Ullah (2007). "Robust forecasting of mortality and fertility rates: A functional data approach". *Computational Statistics and Data Analysis* **51**(10), 4942–4956
- Hyndman, Shang (2009). "Forecasting functional time series (with discussion)". *Journal of the Korean Statistical Society* **38**(3), 199–221
- Shang, Booth, Hyndman (2011). "Point and interval forecasts of mortality rates and life expectancy : a comparison of ten principal component methods". *Demographic Research* **25**(5), 173–214
- Hyndman, Booth (2008). "Stochastic population forecasts using

➥ Papers and R code:

robjhyndman.com

➥ Email: **Rob.Hyndman@monash.edu**

`cran.r-project.org/package=demography`