Stochastic Population Analysis of Asia

Lei Fang
Wolfgang K. Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin

http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

Demography

- Welfare policy, insurance and pension industry, children's service planning
- Aging, low fertility, migration, gender unbalance

Total Fertility Rate

- Total Fertility Rate (TFR) ≥ 2.0
- Low TFR: aging problem and pension crisis

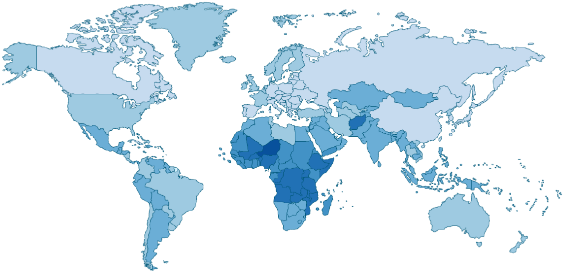

Figure 1: Total fertility rate world map 2012 (source: indexmundi)

Demographic key elements

- Mortality rate: age-specific, male and female, (region-specific)
- Fertility rate: bearing-age specific
- Migration: immigration, emigration
  - Factor in developed countries

Basic Definitions

- Mortality rate is the ratio of number of death and number of exposure, taken as the log transformation.
- Fertility rate is the ratio of number of births per 1000 women at the same age per one calender year.

Note: in following graphs, rates in different years are plotted in rainbow palette so that the earliest years are red and so on.

Demographic Risk

Figure 2: Japan female mortality trend: 1947-2009

Demographic Risk

Figure 3: Japan fertility trend: 1947-2009

Lee-Carter (LC) Method

▸ LC Method

- A benchmark in demographics: Lee and Carter (1992)
- Idea: use SVD to extract a single time-varying index of mortality/fertility rate level
- One component to address demographic rate patterns
  - Take second and higher order PCs
- Take stationarity for granted although structural changes exist
  - Assign higher weights to more recent data

Hyndman-Ullah (HU) Method

- Main ideas of the HU method
  - Nonparametric presmoothing
  - Functional PCA
  - Time series model of factor loading

Objectives

- Employ the LC and HU methods to Asian data sets
- Methods comparison
- Regional trends comparison and discussion

Outline

1. Motivation ✓
2. FDA-based Population Forecasting
3. Empirical Research: Asia
4. Comparisons
5. Discussion
6. References
7. Appendix

Hyndman-Ullah (HU) Method

- Generalization: presmooth, orthogonalize, forecast

- $y_t(x)$ denotes the generic variable: mortality, fertility or migration at age x in year t

$$y_t(x) = s_t(x) + \sigma_t(x)\varepsilon_t \tag{1}$$

$$s_t(x) = \mu_{(}x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x) \tag{2}$$

Constrained and Weighted Smoothing

1. Estimate the smooth functions $s_t(x)$ through the data sets $\{x, y_t(x)\}$ for each t:

$$y_t(x) = s_t(x) + \sigma_t(x)\varepsilon_t$$

- $s_t(x)$ smooth function
- $\sigma_t(x)$ smooth volatility function of $y_t(x)$
- ε_t i.i.d. random error

Constrained and Weighted Smoothing

For each fixed time t,

$$\hat{s}(x) = \underset{s(x)}{\operatorname{argmin}} \sum_{i=1}^{n} \mid y_i - s_i(x) \mid + \lambda \sum_{i=1}^{n-1} \mid s'_{i+1}(x) - s'_i(x) \mid \tag{3}$$

- 1st component denotes the loss part
- 2nd component is the L_1-roughness

Weights

- The residual term $\sigma_t(x)\varepsilon_t$ in (1) determines weight as the inverse standard deviation $\sigma_t^{-1}(x)$ imposed on loss function.
  - Mortality ▸ Binomial Distribution

$$\hat{\sigma}_t^2(x) \approx \{1 - m_t(x)\}N_t^{-1}(x)m_t^{-1}(x) \tag{4}$$

where $m_t(x)$ denotes the mortality rate and $N_t(x)$ denotes the total population of age x in year t.

  - Fertility, by the similar way

$$\hat{\sigma}_t^2(x) \approx \{1000 - f_t(x)\}N_t^{-1}(x)f_t^{-1}(x) \tag{5}$$

where $f_t(x)$ denotes the fertility rate per thousand women of age x in year t.

Constraint

- Constraint for mortality (Wood, 1994)
  Monotonically increasing after some age, like 50.
- Constraint for fertility (He and Ng, 1999)
  Concavity

Functional PCA

2. Use functional principal component analysis (FPCA)

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

- $\mu(x)$ mean of $s_t(x)$ across years
- $\phi_k(x)$ orthogonal basis functional PCs
- $\beta_{t,k}$ uncorrelated PC scores
- $e_t(x)$ is residual function with mean zero

Functional PCA

For a given K, $\{\phi_k(x)\}$ is the solution to minimize the mean integrated squared error

$$MISE = n^{-1} \sum_{t=1}^{n} \int e_t^2(x) dx \tag{6}$$

Estimate the average age term $\mu(x)$ through

$$\hat{\mu}(x) \quad = \quad \underset{\theta(x)}{\operatorname{argmin}} \sum_{t=1}^{n} \| \hat{s}_t(x) - \theta(x) \| \tag{7}$$

where $\|g(x)\| = (\int g^2(x) dx)^{1/2}$ denotes the norm of function g.

Functional PCA

Functional PCA is applied over the $\{\hat{s}_t^*(x)\}$, where $\hat{s}_t^*(x) = \hat{s}_t(x) - \hat{\mu}(x)$ is median-adjusted data.

$$z_{t,k} = w_t \int \phi_k(x)\hat{s}_t^*(x)dx \tag{8}$$

is maximized s.t.

$$\int \phi_k^2(x)dx = 1 \tag{9}$$

$$\int \phi_k(x)\phi_{k-1}(x)dx = 0 \tag{10}$$

Forecasting

3. Due to the way the basis functions $\phi_k(x)$ are chosen, the coefficients $\hat{\beta}_{t,k}$ and $\hat{\beta}_{t,l}$ are uncorrelated for $k \neq l$.

Univariate time series model to forecast the $\beta_{t,k}$:
Optimal ARIMA model

Variants ▸ HUw Method

Demographic Data

- Japan and Taiwan
  Mortality: age-specific (0,110+), male and female
  Extract ages: (0,100)
  Fertility: bearing-age specific (12-,55+)
- Japan
  Fertility: 1947-2009, Mortality: 1947-2012
  Extract years: 1947-2009
- Taiwan
  Fertility: 1976-2010, Mortality: 1970-2010
  Extract years: 1979-2010
- Data Source: Human Mortality Database, Human Fertility Database

Demographic Data

- China
  Mortality: age-specific (0,90+), male and female
  Fertility: bearing-age specific (15-,49+)
- China sample size
  Fertility: 1990-2011 (1992-1994,1997 and 2002 missing)
  Mortality: 1995-2010 (1996, 1997, 2001 and 2006 missing)
- Data Source: China Statistical Year Book
- Missing values are estimated by Moving Average

Japan: mortality

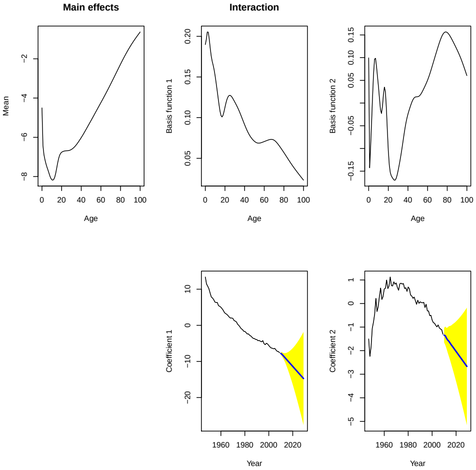

Figure 4: Japan male mortality decomposition

Japan: mortality

Figure 5: Out-of-sample test on Japan's male mortality (1947-1989): forecast rates (black lines) along with 95 % confidence intervals, while actual rates are shown as red circles

Japan: mortality

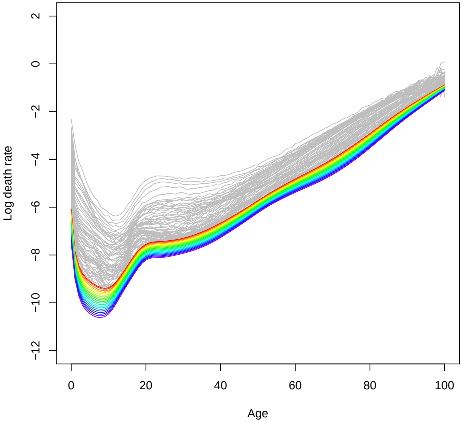

Figure 6: Japan male mortality forecast from 2010 to 2029

Japan: fertility

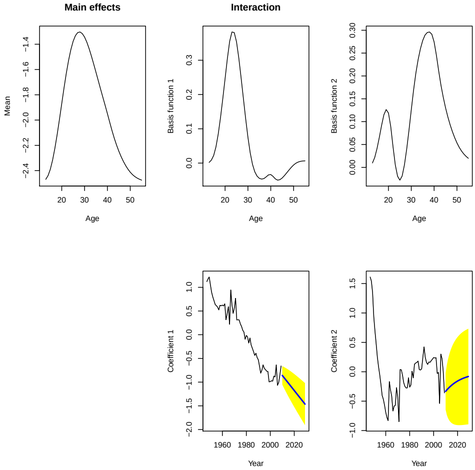

Figure 7: Japan fertility decomposition

Japan: fertility

Figure 8: Out-of-sample test on Japan's fertility (1947-1989): forecast rates (black lines) along with 95 % confidence intervals, while actual rates are shown as red circles

Japan: fertility

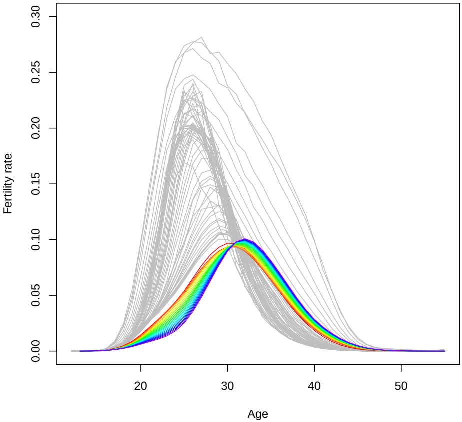

Figure 9: Japan fertility forecast from 2010 to 2029

Methods comparisons

Power of explanation

Female Mortality (%)		
LC model	HU method	Country
96.1	99.9	Japan
86.3	99.0	Taiwan
41.3	98.9	China

Table 1: Explained female mortality variance

Methods comparisons

		1st	2nd	3rd	4th	5th	6th
Japan	*female mortality*	96.5	3.1	0.2	0.1	0.0	0.0
	male mortality	97.0	2.0	0.4	0.3	0.1	0.1
	fertility	58.9	31.0	8.5	1.2	0.2	0.1
Taiwan	*female mortality*	95.1	2.1	0.7	0.5	0.4	0.3
	male mortality	87.6	7.1	2.0	0.8	0.5	0.3
	fertility	90.3	5.5	3.4	0.5	0.1	0.1
China	*female mortality*	84.8	6.1	2.7	2.7	1.5	1.1
	male mortality	78.5	9.3	5.1	2.7	2	0.9
	fertility	47.3	39.1	9.9	2.5	0.5	0.3

Table 2: Explained variance from HU method ($K = 6$)

Methods comparisons

Accuracy of forecast

- Take Japan's female mortality data to compare the accuracy of Lee-Carter model and Hyndman-Ullah Model.
- Divide data set into a fitting period 1947-1989 and forecasting period
- Compare the one-step-ahead forecast and the actual out-of-sample data
- Increase the fitting period by one year until it extends to 2008

Methods comparisons

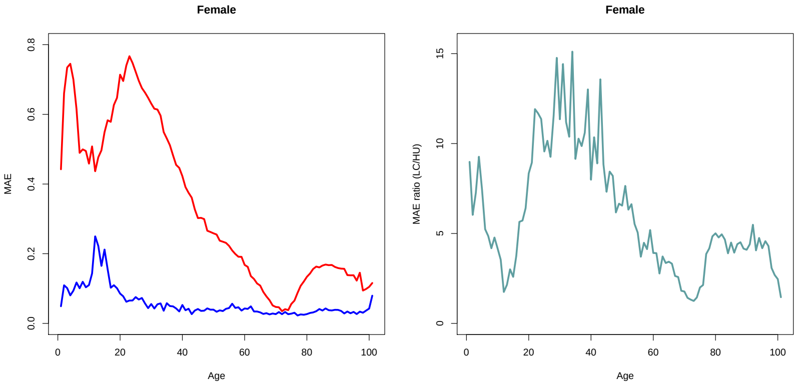

Figure 10: Japan's female mortality Mean Absolute Error for one-step-ahead forecasts averaged over years: LC (red), HU(blue)

Methods comparisons

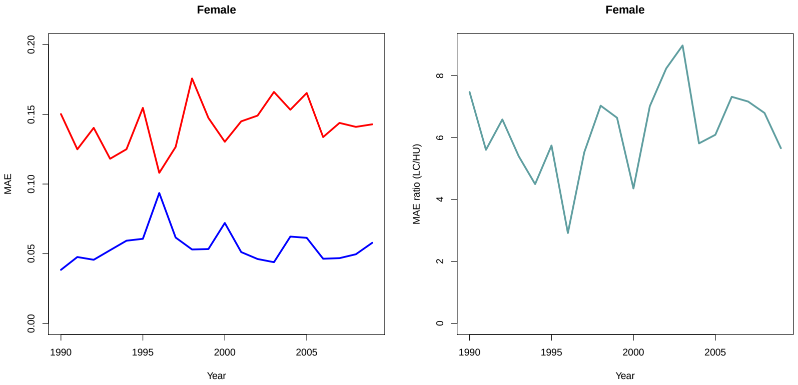

Figure 11: Japan's female mortality Mean Absolute Error for one-step-ahead forecasts averaged over ages: LC (red), HU(blue)

Methods comparisons

Diebold and Mariano (1995) test

- Define the loss differential d_t as $d_t = d_{1t} - d_{2t}$, where $d_{1t} = |\hat{y}_{LC,t} - y_t|$ and $d_{2t} = |\hat{y}_{HU,t} - y_t|$, $t = 1, 2, ..., 20$.
- The null hypothesis is

$$H_0 : E(d_t) = 0, \forall t \tag{11}$$

versus

$$H_1 : E(d_t) > 0. \tag{12}$$

Methods comparisons

- The test statistics is

$$DM = \bar{d}/\sqrt{2\pi\hat{f}_d(0)/T} \tag{13}$$

where $\bar{d} = \sum_{t=1}^{T} d_t$, $\hat{f}_d(0) = \frac{1}{2\pi}\hat{\gamma}_d(0)$,
$\hat{\gamma}_d(0) = \frac{1}{T}\sum_{t=1}^{T}(d_t - \bar{d})^2$ and $T = 20$.

- The p-values obtained from female group and male group are both smaller than 0.01.

- HU method performs better than LC method
- Recent data sets are more fluctuate and difficult to be revealed by decades-ahead data, especially fertility
- Regional similarity in mortality: Japan and China

Regional trends comparisons

- Comparisons on time-varying indices k_t of China and Japan

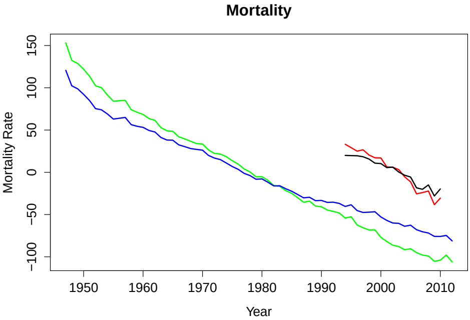

Figure 12: China female mortality (red) vs. Japan female mortality (green) China male mortality (black) vs. Japan male mortality (blue)

Literature

- The Lee-Carter mortality index k_t correlates significantly with macroeconomic fluctuations in some periods, see K. Hanewald (2011).
- Semiparametric comparison of regression curves, see W. Härdle and J.S. Marron (1990).

Stochastic Population Analysis of Asia

Lei Fang
Wolfgang K. Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin

http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

References

K. Hanewald
Explaining mortality dynamics: the role of macroeconomic fluctuations and cause of death trends
North American Actuarial Journal,2011

W. Härdle and J.S. Marron
Semiparametric comparison of regression curves
Annals of Statistics, 1990

R. J. Hyndman and H. Booth
Stochastic Population Forecasts using Functional Data Models for Mortality, Fertility and Migration
International Journal of Forecasting, 2008

References

R. J. Hyndman and Md. S. Ullah
Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach
Computational Statistics and Data Analysis, 2007

R. D. Lee and L. R. Carter
Modeling and Forecasting U.S. Mortality
Journal of the American Statistical Association, 1992

H.L. Shang, H. Booth and R. J. Hyndman
Point and Interval Forecasts of Mortality Rates and Life Expectancy: A Comparision of Ten Principal Component Methods
Demographic Research, 2011

Lee-Carter Method

back

- Take mortality for analysis:

$$\log[m_t(x)] = a_x + b_x k_t + \varepsilon_{x,t}$$

  - $m_t(x)$ observed mortality rate at age x in year t
  - a_x age pattern averaged across years
  - b_x first PC reflecting how fast the mortality changes at each age
  - k_t time-varying index of mortality level
  - $\varepsilon_{x,t}$ residual at age x in year t

Lee-Carter Method

- The LC method is over-parameterized and two constraints are imposed:

$$\sum k_t = 0, \ \sum b_x = 1$$

- Use singular value decomposition (SVD) to derive the parameters k_t and b_x

Lee-Carter Method

- The parameter k_t is forecasted by ARIMA models, and Lee and Carter used a random walk with drift model:

$$k_t = k_{t-1} + d + e_t$$

  - d is the drift parameter reflecting the average annual change
  - e_t is an uncorrelated error

Weighted Hyndman-Ullah (HUw) Method

back

The HUw method takes the same techniques as the HU method, but applies decaying weights in the estimation of $\mu(x)$ and $\phi_k(x)$, and thus realizes higher weights for more recent data

Weighted Hyndman-Ullah Method

1. The weighted function mean $\mu^*(x)$ is estimated by the weighted average

$$\hat{\mu}^*(x) = \sum_{t=1}^{n} w_t f_t(x)$$

- $\{w_t = \lambda(1-\lambda)^{n-t}, t = 1, ..., n\}$ denotes a set of weights, and $0 < \lambda < 1$ denotes a geometrically decaying weight parameter

Weighted Hyndman-Ullah Method

2. By FPCA, the weighted curves is decomposed into orthogonal weighted functional principal components and their uncorrelated scores

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

- $\{\phi_1^*(x), ..., \phi_K^*(x)\}$ denotes a set of weighted functional principal components

Weighted Hyndman-Ullah Method

3. The h-step ahead forecast of $y_{n+h}(x)$ is estimated by the observed data and the set of weighted functional principal components

Binomial Distribution

back

- Take mortality as example:

$$M_t(x) \sim B[N_t(x), m_t(x)]$$

where $M_t(x)$ is the death number of age x in year t.

- $var[m_t(x)] = N_t^{-1}(x)m_t(x)[1 - m_t(x)]$
- The variance of $y_t(x) = log[m_t(x)]$ is obtained by Taylor approximation

$$\hat{\sigma}_t^2(x) \approx [1 - m_t(x)]N_t^{-1}(x)m_t^{-1}(x)$$