The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.6.677
31(6), 677–692

A comparison and prediction of total fertility rate using parametric, non-parametric, and Bayesian model

Jinho Oh[a,1]

[a]Statistical Research Institute, Statistics Korea

(Received June 28, 2018; Revised August 3, 2018; Accepted October 24, 2018)

Abstract

The total fertility rate of Korea was 1.05 in 2017, showing a return to the 1.08 level in the year 2005. 1.05 is a very low fertility level that is far from replacement level fertility or safety zone 1.5. The number may indicate a low fertility trap. It is therefore important to predict fertility than at any other time. In the meantime, we have predicted the age-specific fertility rate and total fertility rate by various statistical methods. When the data trend is disconnected or fluctuating, it applied a nonparametric method applying the smoothness and weight. In addition, the Bayesian method of using the pre-distribution of fertility rates in advanced countries with reference to the three-stage transition phenomenon have been applied. This paper examines which method is reasonable in terms of precision and feasibility by applying estimation, forecasting, and comparing the results of the recent variability of the Korean fertility rate with parametric, non-parametric and Bayesian methods. The results of the analysis showed that the total fertility rate was in the order of KOSTAT's total fertility rate, Bayesian, parametric and non-parametric method outcomes. Given the level of TFR 1.05 in 2017, the predicted total fertility rate derived from the parametric and nonparametric models is most reasonable. In addition, if a fertility rate data is highly complete and a quality is good, the parametric model approach is superior to other methods in terms of parameter estimation, calculation efficiency and goodness-of-fit.

Keywords: total fertility rate, low fertility trap, parametric model, non-parametric model, Bayesian model

1. 서론

최근 우리나라 합계출산율(total fertility rate; TFR)은 2005년 1.08명 수준으로 회귀(regression)하는 현상을 보이고 있다. 2010년 1.23명, 2015년 1.24명의 증가 추세를 보였으나 2016년에는 1.17명, 2017년은 1.05명 수준이다. 특히 2017년은 총 출생아수가 35만명 수준으로 40만명 아래로 떨어졌고, TFR은 초저출산 덫(low-fertility trap) (Lutz 등, 2006)에 빠진 모습을 보인다. 그리고 이러한 TFR은 OECD국가 중 상당히 낮은 출산율 수준이다.

이런 TFR을 수학적, 통계적 모형으로 가정하고 향후 추이를 예측하는 선행연구 (Alkema 등, 2011; Hyndman과 Ullah, 2007, Hyndman 등, 2013; Jun, 2006; Kim과 Jeon, 2015; Park 등, 2013; Ševčíková 등, 2011)는 크게 세 분류로 나눌 수 있다.

본 논문은 통계청의 공식견해가 아니며 저자의 개인적인 연구결과임을 밝힙니다.

[1]Statistical Analysis Division Statistical Research Institute, 6F, Statistical Center, 713 Hanbatdaero, Seo-gu, Daejeon 35220, Korea. E-mail: comet123@korea.kr

먼저 연령별 출산율(age-specific fertility rate; ASFR) 자료의 완비성(completeness)이 높고 품질(quality)이 좋은 경우 모형을 가정하고 모수를 추정하는 모수적 방법, 다음으로 ASFR 시계열이 단절되거나 변동이 심한 경우 평활과 가중치 기법을 활용하는 비모수적 방법, 끝으로 자료 부족과 품질이 좋지 않은 경우 등으로 선진국의 출산율 3단계 전이현상을 참고하여 이들의 사전분포를 활용하는 베이지안(Bayesian) 방법이다.

국내 연구는 모수적 방법에 초점이 맞춰져, 로그감마모형을 가정하여 출산율 추이를 예측한 연구 (Jun, 2006)와 확률적 출산율 모형을 비교한 연구 (Park 등, 2013)가 대표적이다. Park 등 (2013)은 출산율 예측으로 결정론적(deterministic)과 확률론적(probabilistic) 방법을 소개하고 각각의 장단점을 제시했다. 모수화 모형(parameterized model)으로 감마 (Hoem 등, 1981), Hadwiger (Hadwiger, 1940), 베타 (Hoem 등, 1981), 혼합 Hadwiger (Chandola 등, 1999), PK 1, 2 (Peristera과 Kostaki, 2007) 함수 등에 적용하여 모수를 추정한다. 다음 단계로 이 추정된 모수를 시계열 모형에 적합하여 미래의 출산율을 예측하는데 이러한 접근의 유용성은 과거 출산율 패턴이 미래에도 계속되어야 한다는 가정에 기초하고 있다.

특히 우리나라처럼 최근 변동이 큰 출산율의 경우에는 위의 가정이 적합하지 않을 수 있으므로 Park 등 (2013)은 평균출산연령을 중심으로 서로 다른 두 개의 분포가 합쳐진 형태인 혼합정규함수(mixture of normal functions)을 제안하였다. 그리고 출산에 영향을 미치는 혼인관계를 반영한 연령별 출산 예측모형 (Eom과 Kim, 2013)과 출산은 혼인을 전제로 하는 경향이 강하므로 출산율 예측에 관한 연구 이전에 초혼, 재혼 이혼을 고려한 혼인모형을 개발한 연구도 있다 (Kim과 Jeon, 2015).

국외는 모수적 방법 외에 Ramsay와 Silverman (2005)과 Hyndman과 Ullah (2007)은 함수적 자료 분석 패러다임을 사용하여 출산율을 모델링하고 예측하기 위한 비모수적 방법인 함수적 데이터 모형(functional data model; FDM)이 있다. 그들은 관측치에 존재하는 측정오차(measurement error)와 질병이나 전쟁 등으로 인구동태 자료에서 나타나는 불규칙적인 패턴을 교정하기 위해 함수적 자료 분석을 이용하여 출산율 모형을 구축하였는데 측정오차 또는 불규칙 패턴을 교정하기 위해 비모수 평활기법(non-parametric smoothing)을 이용한다 (Kim과 Oh, 2017; Kim 등, 2018).

그리고 출산율 자료 부족으로 완비성이 낮고 품질까지 좋지 못한 경우 선진국의 출산율 3단계 전이현상을 참고하여 이들의 사전분포를 활용하는 베이지안 방법이 있다 (Alkema 등, 2011; Raftery 등, 2012, 2014; Sevcikova 등, 2011). 지면 관계상 베이지안 방법론의 자세한 설명은 본론에 소개한다.

본 논문은 다음과 같은 점에서 선행연구들과 차별된다.

첫째, 우리나라 ASFR에 대해서 모수적 방법에 기초한 추정, 예측결과가 주류이며 상대적으로 비모수와 베이지안 방법을 적용한 연구결과는 미약한 수준이다. 따라서 이들 방법론의 차이점을 모색하고 우리나라 ASFR과 TFR에 어떠한 모형이 적합한지 논의하고자 한다.

둘째, 우리나라 출산율에 3가지 통계적 방법의 결과를 통해 시사점을 도출하고자 한다. 그리고 최근 우리나라 TFR은 12년 만에 2005년 1.08명 초저출산 수준으로 회귀하고 있고 변동 또한 작지 않으므로 모수, 비모수, 베이지안 등 여러모형을 적용한 다각도적인 심층 논의가 이루어져야 한다고 본다. 이러한 차별화된 연구를 수행하기 위해 본 논문은 총 3개의 장으로 구성한다.

제 2장은 출산관련 인구학적 변수들과 1970년부터 2017년까지의 TFR 추이를 살펴본다. 그리고 모수, 비모수, 그리고 베이지안 방법을 소개한다. 제 3장에서는 3가지 통계적 방법 적용과 도출된 출산율 예측을 알아보고, 결과를 비교하여 시사점을 제시한다. 마지막으로 제 4장에서는 결론과 제언을 정리하였다.

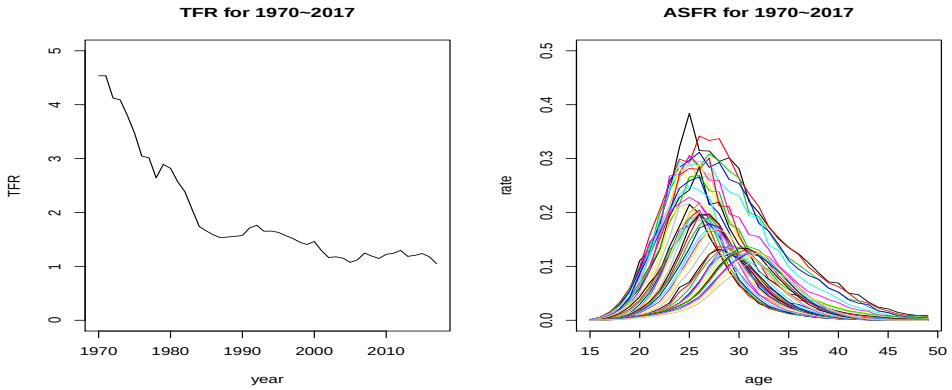


Figure 2.1. Trend of TFR and ASFR for 1970–2017. TFR = total fertility rate; ASFR = age-specific fertility rate.

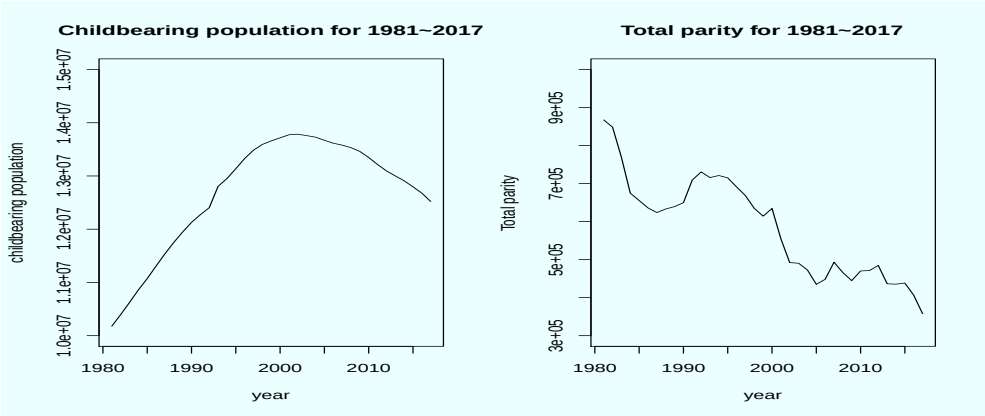


Figure 2.2. Trend of childbearing population and total parity.

2. 합계출산율 추이와 예측방법론

2.1. 출산관련 인구학적 변수 추이

Figure 2.1과 Figure 2.2는 1970–2017년 TFR과 ASFR, 1981–2015년 가임연령인구(15–49세)와 출생아수이다.

먼저 우리나라 TFR은 2005년 1.08명 최저점을 찍고 완만한 상승세를 보였으나 2016년 이후 이전과 다른 하락세를 보여 2017년은 1.05명 수준으로 떨어졌다. 이는 마치 2005년 1.08명 수준으로 회귀하는 양상을 보인다. 그리고 ASFR의 변곡점이 20대 중, 후반에서 30대 초반으로 이동되는 것을 확인할 수 있는데 이는 만혼과 출산지연을 여실히 보여주는 단면이다.

가임여성인구는 2000년 초반까지 증가세를 보이다가 하락세로 돌아서서 점진적 감소추세를 보이고 있고, 이와 연관된 출생아수는 자연히 감소되는 현상을 나타낸다. 특히, 2005년 1.08명 때의 가임연령인구와 출생아수는 1367만명, 43만명이었지만 1.07명인 2017년은 1252만명, 40만명이 깨진 36만명을 보이고 있다.

종합해보면 가임여성과 출생아수의 감소추세와 ASFR의 변곡점이 30대 초, 중반으로 이동되는 양상으로 인해 향후 TFR의 증가는 낙관적이지 못하다. 특히, 1980년대 이후 선진국 출산율은 예전과 다른 상수적인 패턴특성을 보이고 사망률과 국제이동률과는 다르게 쉽게 변화하지 않는 변수 형태가 아닌 일정한 수준을 유지하는 상수적인 패턴, 한번 상승이나 감소기조로 흘러가면 다시 되돌리기 쉽지 않은 비가역성(irreversibility)을 나타낸다. 여기서 선진국은 일본, 프랑스, 독일이다. 특히 프랑스와 독일은 출산장려정책을 성공리에 마쳐 출산 제고를 이룬 나라들로, 프랑스는 1993년, 독일은 1994년, 일본과 한국은 2005년이 출산율 최저점이다. 그런데 한국을 제외한 3개국은 출산율 제고를 보이면서 1.5명 이상의 수준을 보이고 있다.

하지만 우리나라는 출산율 상수패턴과 비가역성에서 벗어나는 출산율 추이를 보이고 있고 변동이 작지 않고 초저출산 덫에 빠진 형국이다. 또한 10여년 동안 이들 국가들의 출산율 추이와는 사뭇 다른 경향을 보이므로 향후 TFR예측은 어느 때 보다도 중요하다 할 수 있다.

2.2. 출산력 예측의 모수, 비모수, 베이지안 방법

그동안 ASFR의 패턴을 설명하기 위해 많은 모수화 모형이 소개되어 왔다. 이들 모수화 모형은 Hoem 등 (1981)가 제안한 베타함수를 제외하고는 모두 지수족(exponential family) 형태를 띠고 있다. 이는 ASFR의 대략적인 분포형태가 평균출산연령을 중심으로 좌우 대칭을 띠고 있기 때문이다.

통계청은 ASFR 적합, 예측을 위해 일반화 로그감마(generalized log gamma; GLG) 모형을 활용한다 (KOSIS, 2011, 2016). 식 (2.1)은 GLG으로 모수 4개 (C_i, μ, b, λ)가 포함됨을 알 수 있다.

$$f_i(x) = \frac{C_i|\lambda|}{b\Gamma(1/\lambda^2)} \left(\frac{1}{\lambda^2}\right)^{\lambda^{-2}} \exp\left[\frac{1}{\lambda}\left(\frac{x-\mu}{b}\right) - \frac{1}{\lambda^2}\exp\left(\lambda\left(\frac{x-\mu}{b}\right)\right)\right], \tag{2.1}$$

여기서 $f_i(x)$는 연령 x세의 출산율, C_i는 특정의 출생코호트가 가임연령동안 출산순위 i번째 자녀의 출산을 경험할 확률, μ와 b는 출산연령의 평균과 표준편차, λ는 분포형태를 나타내는 모수이다. 따라서 GLG 모형은 결혼의 출생순서(birth-order)에 대한 연령패턴의 규칙성(regularity)을 수학적으로 표현한 것 (Kaneko, 2003)으로 이해할 수 있다. 최종적으로 로그감마모형으로 모수를 추정한 후 이 추정된 모수를 시계열 모형에 적합하여 미래의 출산율을 예측한다.

다음으로 비모수 모형의 출산율 적합과 예측이다. 본 연구에서 제안된 FDM 모형의 구조는 식 (2.2)와 (2.3)이다.

$$f_t(x) = \begin{cases} \frac{1}{\lambda}\left(f_t^*(x)^\lambda - 1\right), & 0 < \lambda < 1, \\ \ln\left(f_t^*(x)\right), & \lambda = 0. \end{cases} \tag{2.2}$$

식 (2.2)에서 $f_t^*(x)$는 시간 t와 연령 x에서 출산율을 의미한다. $f_t^*(x)$의 Box-Cox (Box와 Cox, 1964)변형은 $f_t^*(x)$의 값에 따라 증가하는 변동을 줄여주거나 정규화과정으로 λ는 Box-Cox 변형에서 강도를 뜻한다.

$$\begin{aligned} f_t(x) &= s_t(x) + \sigma_t(x)\epsilon_{t,x}, \\ s_t(x) &= \mu(x) + \sum_{i=1}^{J} \beta_{t,j}\phi_j(x) + e_t(x), \end{aligned} \tag{2.3}$$

여기서 $\mu(x)$는 $\sum_{t=1}^{n} s_t(x)/n$에 의해 추정된 평균함수로 평활된 연령에 따른 로그출산율평균이고, $(\beta_{t,j}\phi_j(x);\ t = 1, \ldots, n,\ j = 1, \ldots, J)$는 함수적 주성분분석(functional principal components analysis; FPCA) (Ramsay와 Silverman, 2005; Kang과 Ahn, 2006; Hyndman과 Ullah, 2007)을 사용

하여 추정되어지며 $J < n$는 사용된 주성분 수이다. $\Phi = \{\phi_1(x), \ldots, \phi_J(x)\}$는 J개의 함수적 주성분의 집합으로 직교 기저함수(orthogonal basis function)이고 $B = \{\beta_{t,1}, \ldots, \beta_{t,J}\}$는 비상관 주성분 점수(uncorrelated principal component scores)들의 집합으로 시계열 계수를 의미한다. 식 (2.3)에서 $f_t(x)$는 시간 t의 연령 x에 대한 관찰된 로그출산율 $\ln f_t(x)$이고, $s_t(x)$는 평활함수(smooth function), $\epsilon_{t,x}$는 독립적이고 동일하게 분포된 표준정규 확률변수이고, $\sigma_t(x)$는 시간 t의 연령 x에 따라 변하는 잡음의 양이다. 즉, $\sigma_t(x)\epsilon_{t,x}$는 관측된 로그출산율과 평활된 곡선의 차이인 관측오류를 의미한다. 식 (2.3)의 두 번째 식은 시간에 따라 변화하는 $s_t(x)$의 변화를 설명하는 부분으로 하나 이상의 주성분을 사용하고 FPCA을 사용하여 평활된 곡선 $s_t(x)$를 직교함수 주성분과 비상관 주성분 점수로 분해한 것이다.

일반적으로 출산율은 연도에 따른 이산형 자료이지만, 시간적 추이에 따른 출산율의 점진적 변화는 임의의 곡선으로 묘사 가능하므로 일종의 함수(function)라고 할 수 있으며 이러한 데이터를 분석하는 통계적 분야는 함수적 데이터 분석(functional data analysis; FDA)이다. 이런 이산형 자료를 가지고 모함수 형태를 근사시키기 위한 방법으로 보간법, 평활법, 기저의 선형결합이 있는데 본 연구는 평활법과 기저 함수의 선형결합을 활용하고, FDA에서 자료의 변동성을 알아보기 위해 FPCA를 사용하였다. 보다 자세한 설명은 Ramsay와 Silverman (2005)나 Kang과 Ahn (2006)을 참고하기 바란다.

FDM은 첫 번째 주성분에 직교하는 고차원 주성분에 대해서는 다른 시계열 모형들의 주성분 점수가 도출된다. 모든 성분에 FDM 방법은 최적 시계열 모형을 Akaike information criterion (AIC) 등과 같은 모형 판별 기준에 의거하여 선택한다. 이 모형에 대한 자세한 설명은 Hyndman과 Ullah (2007), Hyndman과 Booth (2008), Hyndman 등 (2013), Kim과 Oh (2017), Kim 등 (2018)을 참조하면 된다. 다음으로 실측치 $Z = \{f_1(x), \ldots, f_n(x)\}$와 $\Phi = \{\phi_1(x), \ldots, \phi_J(x)\}$의 조건부로 $f_{n+h}(x)$의 h단계 예측치를 구할 경우 식 (2.4)에 의해서 도출 가능하다.

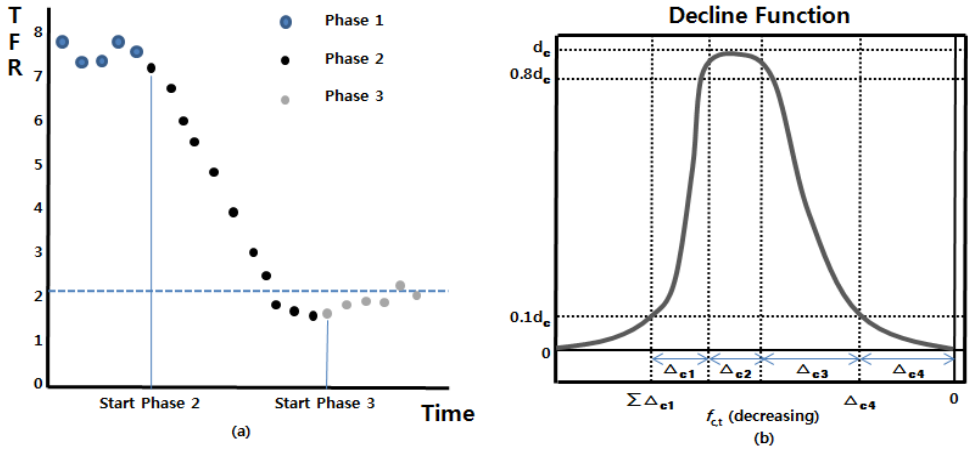
$$\hat{f}_{n+h|n}(x) = E[f_{n+h}(x)|Z, \Phi] = \hat{\mu}(x) + \sum_{j=1}^{J} \hat{\beta}_{n+h|n,j}\phi_j(x), \tag{2.4}$$

여기서 $\hat{\beta}_{n+h|n,j}$은 Hyndman과 Booth (2008)에 의한 지수 평활 또는 자기회귀누적이동평균과정(autoregressive intergrated moving average model; ARIMA) 모형과 같은 일변량 시계열모형을 활용하여 도출한 $\beta_{n+h,j}$의 h 단계 예측을 의미한다. 그리고 모든 주성분들은 직교성을 보이므로 예측분산을 다음과 같이 도출 가능하다.

$$\hat{v}_{n+h|n}(x) = \mathrm{Var}[f_{n+h}(x)|Z, \Phi] = \sigma_\mu^2(x) + \sum_{j=1}^{J} u_{n+h|n,j}\phi_j^2(x) + v(x) + \sigma_t^2(x), \tag{2.5}$$

여기서 $\sigma_\mu^2(x)$는 $\hat{\mu}(x)$의 분산, $u_{n+h|n,j}$은 $\beta_{n+h,j}|\beta_{1,j}, \ldots, \beta_{n,j}$의 분산이다. 그리고 $v(x)$, $\sigma_t^2(x)$는 각각 $e_t(x)$와 $\sigma_t(x)$의 분산이다.

끝으로 베이지안 방법이다. 최근 유엔인구처(UN Population Division; UNPD)는 TFR 추정과 예측을 위해 TFR 전이단계(TFR 추이를 3단계로 나누어 정의하는 것으로 1단계는 출산율 5–7명의 높은 출산율 수준, 2단계는 높은 출산율 수준에서 인구대체선 수준이나 미만으로 감소하는 구간, 3단계는 낮은(최저) 출산율에서 서서히 증가해 인구대체선까지 도달하는 단계, Figure 2.3(a)참조) 중 2단계에 5년 전후 TFR의 차이(gap, $f_t - f_{t+5}$)를 관찰하여 감소함수(decline function)로 정의하고, 식 (2.6)의 더블로지스틱(double logistic) 모형으로 적합한다 (Alkema 등, 2011; Raftery 등, 2012, 2014; Sevcikova 등, 2011, 2018).



자료출처: Alkema 등 (2011), Raftery 등 (2014)

Figure 2.3. The three phases for TFR and trajectory on decline function.

그리고 일반적인 TFR 추이는 전이 1단계에서 감소하다가 최소점을 지나 점진적인 증가추세를 보이므로 5년 전후의 TFR 차이를 그려보면 음의 값으로 증가하다가 최고점을 찍고 다시 하락하는 추이를 도출할 수 있다. 이들 값들을 0을 중심으로 부호(sign)을 반대로 부여해 그리면 Figure 2.3(b)가 되고 Alkema 등 (2011)과 Raftery 등 (2012, 2014)은 이를 감소함수라고 명명하였다.

$$
\begin{aligned}
f_{c,t+5} &= f_{c,t} - r(f_{c,t}|\delta^c) + a_{c,t}, \quad \tau_c \le t < \lambda_c, \\
\tau_c &= \begin{cases} \max\{t : (M_c - L_{c,t}) < 0.5\}, & L_{c,t} > 5.5, \\ < 1950 \sim 1955, & \text{otherwise}, \end{cases} \\
\lambda_c &= \min\{t : f_{c,t} > f_{c,t-1}, f_{c,t+1} > f_{c,t} \quad \text{and} \quad f_{c,p} < 2, p = t-1, t, t+1\} \qquad (2.6) \\
r(f_{c,t}|\delta^c) &= \frac{-d_c}{1 + \exp\left[-2\ln(9)(f_{c,t} - \sum_{i=2}^{4} \triangle_{ci} + 0.5\triangle_{c1})/\triangle_{c1}\right]} \\
&\quad + \frac{d_c}{1 + \exp\left[-2\ln(9)(f_{c,t} - \triangle_{c4} - 0.5\triangle_{c3})/\triangle_{c3}\right]}
\end{aligned}
$$

여기서, M_c는 국가별 TFR 최대값, $L_{c,t}$는 부분 최대치(local maxima)이고, 국가별 전이 3단계의 시작점을 λ_c라고 했을 때, 이는 TFR 2 이하 부분에서 인접 두시점에서 증가로 판정된다면 이는 전이 3단계로 접어들었다고 간주하는데 대체적으로 λ_c는 2005–2010년 이후에 발생한다. 그리고 $\delta^c = (\triangle_{c1}, \triangle_{c2}, \triangle_{c3}, \triangle_{c4}, d_c)$는 국가별 모수벡터, $a_{c,t} \overset{\text{ind}}{\sim} N(0, \sigma(t, f_{c,t})^2)$는 출산수준과 시간변동에 따른 표준편차를 의미한다.

이때 $\triangle_{c1}, \triangle_{c2}, \triangle_{c3}, \triangle_{c4}, d_c$은 Figure 2.3(b)의 곡선 궤적 단계를 표현하는 값으로 이해하면 된다. 출산율이 점진적으로 감소하는 구간은 $\triangle_{c1}, \triangle_{c2}$에 해당하고 최고점은 d_c, 그리고 최저점에서 점진적으로 증가하는 구간은 $\triangle_{c3}, \triangle_{c4}$에 속한다. 특히 $\triangle_{c4}$는 예측구간에 해당되는 경우가 많은데 UN (2017)에서는 상한을 인구대체율인 2.1명으로 하한은 안전선인 1.5명으로 정하고 있다.

그리고 국가별 모수 δ^c는 널리 퍼진 사전분포(diffuse prior distribution)를 가지는 21개국 선진국 출산율 분포로부터 도출되는 것을 가정한다. Raftery 등 (2014)은 이런 일련의 분석과정을 마코브 체인

몬테 카를로(Markov chain Monte Carlo; MCMC) 방법을 활용하여 출산 전이 모형에 대한 각 모수들의 사후분포 표본(posterior distribution sample)을 도출한다. MCMC 자세한 알고리즘은 깁스샘플링(Gibbs sampling), 메트로폴리스-해스팅(Metropolis-Hastings) 그리고 슬라이스 샘플링 단계(slice sampling steps)의 조합 (Neal, 2003)이라고 밝히고 있다.

다음으로 3단계는 국가별(country-specific) 식 (2.7)과 같은 자기상관회귀(국가별 출산율 평균(μ_c)에 근사하는 1차 자기회귀 시계열 모형 AR(1)) 시계열 모형으로 적합하여 출산율을 추정한다. Alkema 등 (2011)이 출산율 전이 3단계를 도출할 때, 21개국 선진국(19개 유럽국가, 미국, 싱가포르) 출산율을 참고하여 전이 3단계의 출산율이 2.1명(μ)로 회귀한다는 강한(strong) 가정과 AR(1) 모형으로 추정하였다. 자세한 모형은 아래와 같다.

$$f_{c,t+5} - \mu = \rho(f_{c,t} - \mu) + b_{c,t}, \quad b_{c,t} \sim N\left(0, \sigma_b^2\right) \tag{2.7}$$

그리고 모수 ρ와 σ_b는 출산율 전이 3단계에 접어든 21개국의 1950–2014년 출산율로부터 최대우도추정량으로 추정($\hat{\rho} = 0.89$, $\hat{\sigma}_b = 0.10$)하였다.

그러나 위의 가정들을 토대로 도출된 결과에 대해서 Basten 등 (2012)은 동아시아(한국, 일본, 홍콩, 싱카포르, 대만) 출산율 환경과는 괴리감이 있다고 지적하였다. 즉, 도출된 TFR은 너무 높기 때문에 Reftery 등 (2014)은 국가별(μ_c)로 AR(1) 모형을 식 (2.8)과 같이 변경해서 제안하였다.

$$f_{c,t+5} - \mu_c = \rho(f_{c,t} - \mu_c) + b_{c,t}, \quad b_{c,t} \sim N\left(0, \sigma_b^2\right), \ \mu_c \sim TN_{[0,\infty)}\left(\bar{\mu}, \sigma_\mu^2\right), \ \rho \sim TN_{[0,1]}\left(\bar{\rho}, \sigma_\rho^2\right) \tag{2.8}$$

여기서, $TN_{[a,b]}(\mu, \sigma^2)$은 a와 b사이로 절단(truncated)되어진 평균이 μ이고 표준편가가 σ인 절단 정규분포(truncated normal distribution)를 의미한다. 그리고 Raftery 등 (2014)는 이들 모수에 대한 사전분포를 다음과 같이 제시했다.

$$\bar{\mu} \sim U[0, 2.1]; \quad \sigma_\mu \sim U[0, 0.318]; \quad \rho_c \sim U[0, 1]; \quad \sigma_p \sim U[0, 0.289]; \quad \sigma_\epsilon \sim U[0, 0.5]. \tag{2.9}$$

지금까지 설명한 일련의 분석과정을 Sevcikova 등 (2011)은 베이지안 계층적 출산율 모형으로 구현하고 R 프로그램 bayesTFR 패키지로 소개해 제공하여 세계 200여개국의 출산율을 예측할 수 있게 하였다.

그리고 현재 출산율 전이 2단계, 3단계에 해당되는 국가별 향후 출산율 예측은 다음과 같다.

출산전이 2단계(시점 t, 국가 c의 $f_{c,t+5}$ 예측은 일종의 평가(evaluated), 검증(cross-validation) 방법과 유사하다. $f_{c,t+5}^{(i)}$는 $f_{c,t+5}^{(i)} = f_{c,t}^{(i)} - d_{c,t}^{(i)} + \epsilon_{c,t}^{(i)}$로 도출되는 i번째 표본군 예측으로, 여기서 $f_{c,t}^{(i)}$는 시점 t에서 도출된 TFR 결과값 표본의 i번째를 의미하고, $d_{c,t}$는 $f_{c,t}^{(i)}$와 모수 벡터 $\delta_c^{(i)}$에서 평가된 감소함수에 의해 도출되는 기대 감소(expected decrement)를 뜻한다. 출산전이 3단계의 시작점은 두 조건($\min_t\{f_{c,t}^{(i)}\} \leq \triangle_{c4}^{(i)}$과 $f_{c,t}^{(i)} > f_{c,t-1}^{(i)}$)을 만족하는 가장 최근의 시점 t로 정한다. t시점 이후의 출산전이 3단계로 진입한 것으로 간주하여 이 시점부터는 식 (2.8)의 AR(1) 모형을 사용하여 예측한다. 3단계에 접어든 선진국이나 개발도상국 일부 국가들의 TFR의 패턴은 약간의 상승이나 일정한 수준을 유지하는 상수적인 패턴을 보인다. 따라서 AR(1) 모형을 적용해도 무리가 없을 것으로 판단된다.

지금까지 서술한 베이지안 방법을 기초로, 본 연구는 우리나라 출산율을 추정하기 위해 bayesTFR 패키지를 활용한 R 프로그램을 부록에 제시하였다. 관심있는 독자는 참고하길 바란다.

3. 합계출산율 예측과 비교

3.1. 합계출산율 예측

본 연구는 통계 R 프로그램을 활용하여 모수적 모형인 GLG모형을 구현하고, 4개 모수 미래 시계열 예

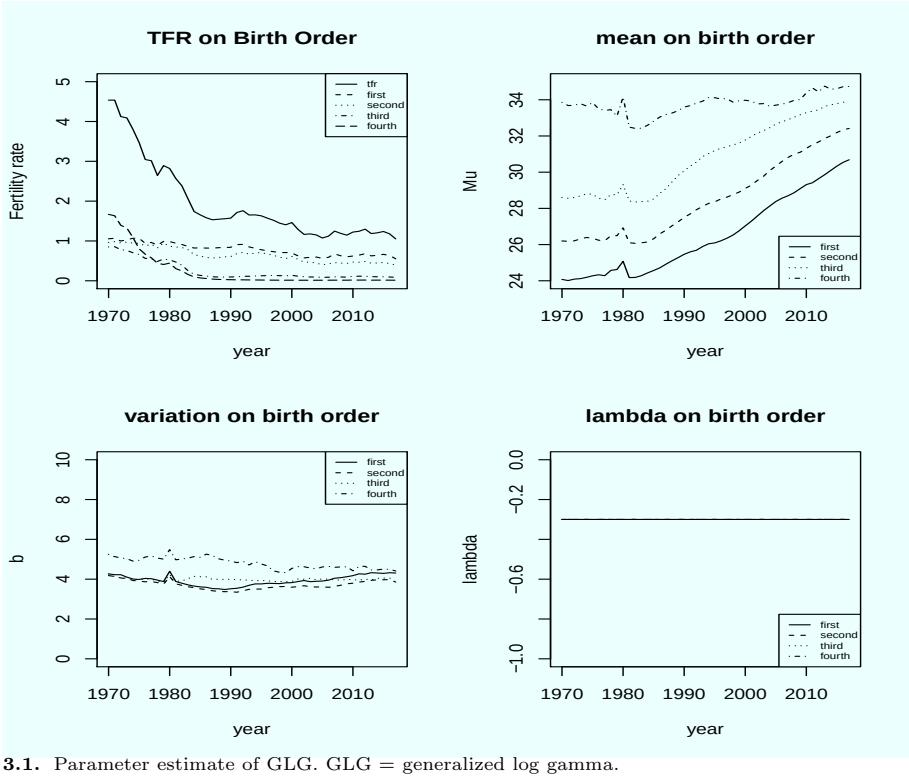


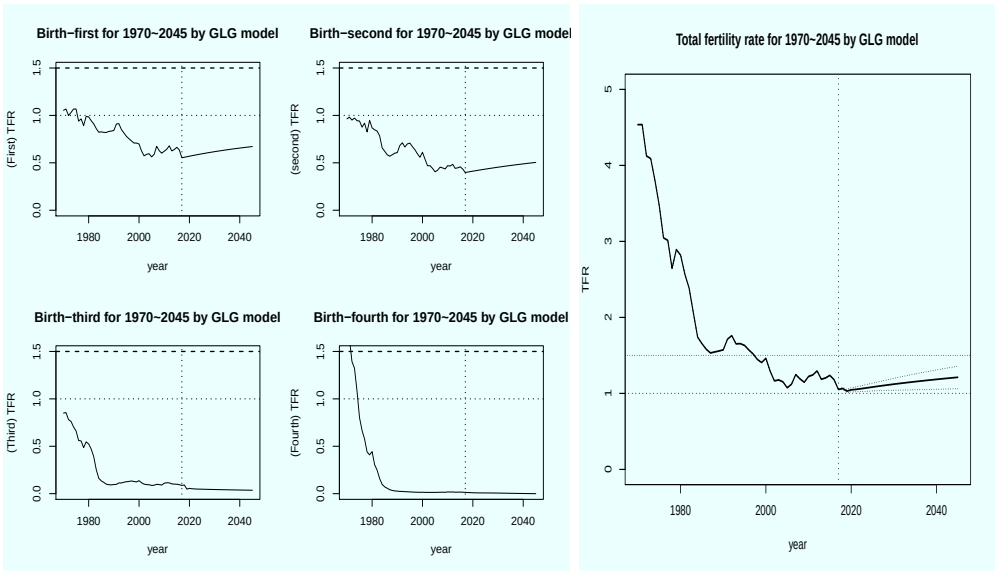
Figure 3.1. Parameter estimate of GLG. GLG = generalized log gamma.

측에 ARIMA모형을 적용하였다. 그리고 비모수와 베이지안 분석을 위해 오픈 소스로 제공하는 R 프로그램 Demography 패키기 (Hyndman 등, 2013)와 bayesTFR 패키지 (Sevcikova 등, 2018)를 활용하였다. 분석을 위해 통계청 공식통계인 1970–2017년의 ASFR 자료(1970–2017년 통계청 'KOSIS-인구,가구-인구동향조사-인구동태건수 및 동태율 추이'에서 제공)와 BayesTFR에서 제공하는 WPP 2017(WPP2017에 사용된 자료는 UNSD(유엔 통계부)가 매년 여러 국가 통계청에서 제공하는 인구동태 통계를 통합한 것으로, 이 자료를 토대로 UNPD는 세계인구를 추정)를 활용하였으며, 2045년 이후 동일한 수준을 보이는 통계청의 TFR과 비교하기 위해 2045년까지 예측하기로 한다.

먼저 GLG 모형 적합과 예측 결과이다. 1970–2017년 출산 순위별 ASFR 데이터(2017년은 ASFR은 최근 3년간 출산 감소율을 반영해 추정)에 대해서 GLG 모형으로 적합한 후, 출산 순위별 4개 모수의 적합값을 산출한다 (Figure 3.1 참조). 이들 추정값을 토대로 시계열 방법을 적용해 미래 모수값들의 추이를 알아보았다 (Figure 3.2 참조).

그 결과 2020년 1.04명, 2025년 1.07명, 2030년 1.11명, 2035년 1.15명, 그리고 2045년은 1.21명으로 산출되었다. 1.21명은 통계청의 장래인구추계 2045년 TFR의 중위수준보다 저위수준에 가깝고, 2045년까지의 TFR과는 상당한 차이를 보인다. 그 이유는 2017년까지의 최신 ASFR 자료 반영 유무와 4개 모수 예측 시계열 모형의 차이 때문이다. GLG모형 4개 모수의 통계청 시계열 적합 모형은 알 수 없지만 본 연구에서는 ARIMA를 적용하고 그 결과를 Table 3.4에 제시하였다.

그리고 Figure 3.1과 같이 λ값은 일정한 상수형태를 보이고, μ는 증가, b는 감소, 그리고 C_i는 감소하다가 2005년 최저점 이후로 증가 한 후에 2015년 이후에 또 다시 하락하는 패턴을 보였다. GLG 모형



(a) (b)

Figure 3.2. Prediction for birth-order and TFR by GLG. TFR = total fertility rate; GLG = generalized log gamma.

으로 도출되는 ASFR은 C_i와 μ는 정의 관계, b와 λ는 역의 관계다. 미래의 μ는 출산기피와 만혼 등으로 점진적인 상승을 보일 것이고, C_1와 C_2값은 2020년 이전까지는 감소하다가 2020년 이후로는 완만한 상승세로 예측되었다. 이를 종합해 보면 GLG 모형에 의한 미래 TFR값은 Figure 3.2처럼 완만한 상승으로 이어진다. 이러한 결과는 출산율 미래 추이의 복원성 가정(under a stationary fertility policy)에 기초하고 있다. 만약 이런 가정과 달리 출산율 전이 2, 3단계를 비정상 시계열로 간주하여 예측할 때 본 연구 결과와는 다를 수 있음을 밝힌다.

다음으로 비모수모형 적합과 예측결과이다. 우선 FDM의 주성분 결정은 mean absolute error (MAE, $\sum |f_x - \hat{f}_x|/n$), mean absolute percentile error (MAPE, $\sum |(f_x - \hat{f}_x)/\hat{f}_x n| * 100\%$)를 기준치로 선택하였다. 또한 식 (2.4)의 $\hat{\beta}_{n+h|n,j}$을 추정하기 위해 지수형 평활 상태공간 모형(exponential smoothing state space model; ets), ARIMA, 확률보행(random walk; RW), 절편이 있는 확률보형(RW drift) 등과 같은 시나리오를 고려하였다. 이들 특징은 미래 추이가 RW drift는 시간이 흐름에 따라 수준이 점차로 증가 또는 감소 경향을, RW는 확률보행 경향, ARIMA는 비정상 시계열을 정상이 되도록 한 후 정상시계열 분석에 쓰이고 증가 혹은 감소를 병행, 그리고 ets는 지수형 평활방법이므로 앞 시차의 추세를 반영하는 경향을 나타낸다.

분석결과 Figure 3.3과 같이 주성분 2 이상은 4개 옵션 모두 거의 동일한 궤적을 보인다. 그리고 'ets'와 'RW drift'는 큰 차이를 보이고 ARIMA와 RW는 동일한 추세를 나타낸다. 하지만 2045년 RW drift의 출산율 예측 결과인 0.5명은 합리적인 결과치라고 보기에는 어렵다.

이처럼 다양한 결과들 중에서 본 연구는 이전 출산 코호트의 행동은 간접적으로 후의 코호트들에게 영향을 미친다 (Ryder, 1990; Evans, 1986)는 선행연구 결과와 최근 우리나라 출산율 추이 환경에 유사한 ets 예측결과를 채택하였다. ets는 2020년 1.05명, 2025년 1.06명, 2030년 1.07명, 2035년 1.10명, 2040년 1.14명, 그리고 2045년은 1.19명으로 GLG 모형 예측 결과와 유사한 증가추세를 보였다. 참고

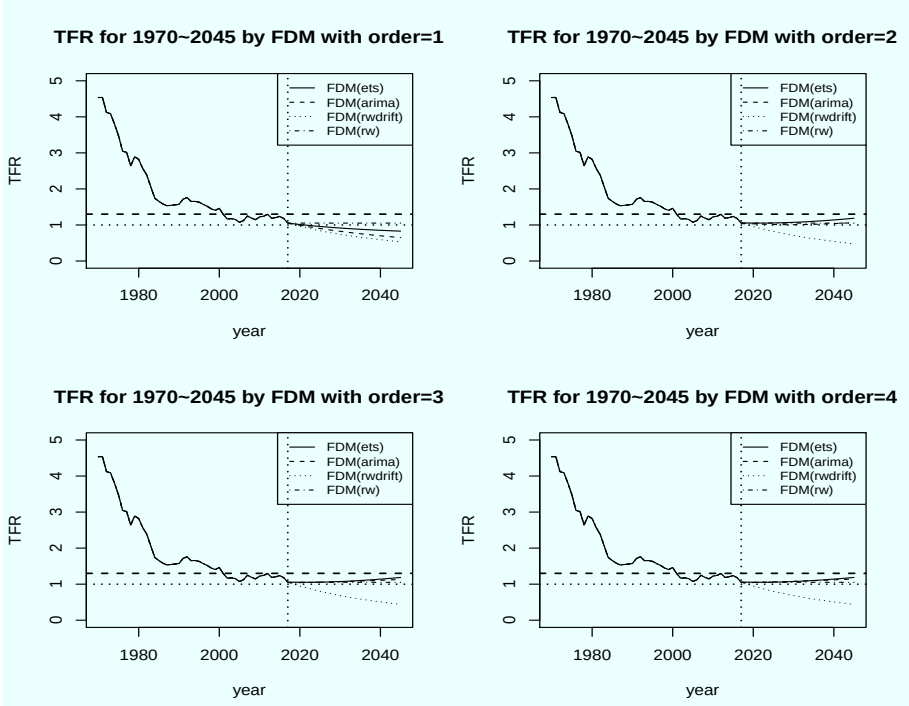


Figure 3.3. Estimation and prediction for TFR by FDM. TFR = total fertility rate; FDM = functional data model.

Table 3.1. TFR for Korea of WPP 2017

연도	1950–1955	1955–1960	1960–1965	1965–1970	1970–1975	1975–1980	1980–1985
TFR	5.653	6.332	5.600	4.650	4.002	2.919	2.234
연도	1985–1990	1990–1995	1995–2000	2000–2005	2005–2010	2010–2015	-
TFR	1.567	1.676	1.501	1.214	1.173	1.233	-

TFR = total fertility rate.

적으로 ARIMA와 RW는 2020년 1.03명, 2025년 1.01명, 2030년 1.01명, 2035년 1.02명, 2040년 1.04명, 그리고 2045년은 1.07명으로 도출되었다. 그리고 주성분에 따른 MAE, MAPE와 주성분 수에 따른 직교 기저함수의 설명력은 Table 3.3에 제시하였다. 주성분 수가 2개 이상이면 설명력은 거의 동일함을 알 수 있다.

대체적으로 2025년 이전까지는 모수적, 비모수적 모형 적용의 예측결과가 유사한 수준이지만, 그 이후로는 모수적 모형 산출치 보다는 낮은 수준을 나타났다. TFR 변동이 큰 경우 비모수적 접근이 모수적 접근보다 경험적으로 보수적인 예측값으로 산출된다.

끝으로 베이지안 모형적합과 예측결과이다. 분석전 WPP 2017의 우리나라 TFR 추이를 살펴보았으며 (Table 3.1), 우리나라 TFR 추세와 유사한 패턴으로 보여진다.

bayesTFR은 국가별 출산율 전이 3단계 추정과 예측을 위해 다양한 가정 설정을 요구한다. 기본적으로 2005–2010년 이후에 출산 전이 2단계에서 3단계로 접어들기 때문에 Figure 2.3에서 $\triangle_{c_4}$의 상한과 하한, 식 (2.8)의 μ_c에 대한 목표출산율 설정이 중요하다. bayesTFR은 출산율의 글로벌 패턴 분석을 위

Table 3.2. Results on prediction of scenarios for bayesTFR

Various options	2020	2025	2030	2035	2040	2045
$\triangle_{c_4}$ 하한(1.0), 상한(2.5), $\mu = 2.1$, mintfr = 0.5	1.32	1.37	1.45	1.52	1.58	1.64
$\triangle_{c_4}$ 하한(0.9), 상한(1.2), $\mu = 1.5$, mintfr = 0.9	1.17	1.12	1.12	1.12	1.14	1.15
$\triangle_{c_4}$ 하한(0.9), 상한(1.3), $\mu = 1.5$, mintfr = 0.9	1.18	1.15	1.12	1.16	1.19	1.22
$\triangle_{c_4}$ 하한(0.9), 상한(1.4), $\mu = 1.5$, mintfr = 0.9	1.24	1.25	1.28	1.28	1.33	1.36

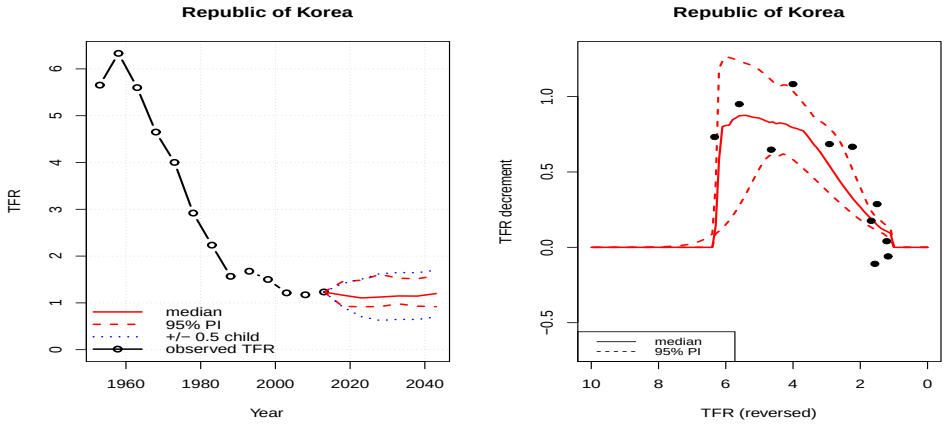


Figure 3.4. Prediction by Bayesian and pattern on declined function for Korea. TFR = total fertility rate.

해 $\triangle_{c_4}$의 상한을 2.5명, 하한을 1명, μ_c의 목표출산율을 인구대체선인 2.1명, 최소 TFR을 1.5명으로 설정(default)해 제공하고 있다. 국가별로 이러한 설정이 상이하므로 조정이 필요하다. 특히 우리나라 출산율 추이를 참고했을 때 위의 설정은 적합하지 않은 가정으로 판단된다.

따라서 본 연구는 2015년 이후 하락세를 보이는 우리나라 2016–2017년 TFR 패턴을 반영할 수 있도록 bayesTFR의 run.tfr.mcmc 옵션설정에서 TFR 하한과 상한을 다양하게 설정해 보았다. 동시에 향후 TFR 예측을 위해 tfr.predict의 μ는 2를 1.5로 min.tfr은 1.5를 0.9로 수정하였음을 밝힌다.

이러한 옵션 변경 이유는 첫째, $\triangle_{c_4}$는 시기상으로 출산 전이 2단계의 마지막과 3단계 초기 진입에 해당된다. 이 시기의 우리나라는 2005–2010년 이후이고, TFR이 1.3명을 넘어선 적이 없다. 둘째, 최근 우리나라 TFR은 1.1명 이하로 떨어졌고 가임연령인구와 출생아수 감소를 감안했을 때 향후 인구대체율 2.1명 수준으로의 회귀는 비현실적이다.

Table 3.2는 다양한 가설을 적용한 우리나라 TFR 시나리오 예측 결과이다. GLG와 FDM 예측결과와 2000년 이후의 출산율 추이를 참고할 때 4결과 중 기본 설정(default)과 $\triangle_{c_4}$의 하한(1.0), 상한(1.4)조건보다 나머지 두 결과가 합리적으로 판단된다. 본 연구는 $\triangle_{c_4}$의 하한(0.9), 상한(1.3)과 mu = 1.5, mintfr = 0.9 가정 결과를 채택해 타 모형과 비교한다. 이 조건의 결과는 2020년 1.18명, 2025년 1.15명, 2030년 1.12명, 2035년 1.16명, 2040년 1.19명, 그리고 2045년 1.22명 수준으로 예측되었다. Figure 3.4는 이들 조건에서 도출된 TFR 추이와 감소함수 궤적을 도시한 것이다.

Table 3.3은 3가지 출산율 모형 예측 결과를 보여주는데, 통계청 TFR과 베이지안 결과는 모수, 비모수 예측 결과보다 높은 값을 보인다. 출산수준의 순위는 2030년까지는 비모수, 모수, 베이지안, 통계청 TFR 순이다. 이러한 결과 도출의 주된 이유는 최근 출산율 자료 반영유무와 예측시 연구자가 경험적으로 정의하는 가정 때문이다.

Table 3.3. A comparison of TFR on various fertility models

	2020	2025	2030	2035	2040	2045
모수(generalized log gamma model; GLG model)	1.04	1.07	1.11	1.15	1.18	1.21
비모수(functional dada model; FDM)	1.05	1.06	1.07	1.10	1.14	1.19
베이지안(Bayesian hierarcical model; BHM)	1.18	1.15	1.12	1.16	1.19	1.22
통계청 TFR	1.24	1.28	1.32	1.36	1.38	1.38
범위(max–min)	0.20	0.22	0.25	0.26	0.24	0.19

TFR = total fertility rate.

Table 3.4. The results of various fertility models

	Birth-Order	C	μ	b	λ	MAE
GLG	First	ARIMA$(1,0,0)$ -154.342	ARIMA$(0,0,2)$ 116.507	ARIMA$(1,0,0)$ -204.281	ARIMA$(2,0,2)$ -1875.527	0.0000
	Second	ARIMA$(1,0,0)$ -151.168	ARIMA$(0,0,2)$ 115.046	ARIMA$(1,0,0)$ -193.557	ARIMA$(1,0,0)$ -1875.010	0.0000
	Third	ARIMA$(2,0,2)$ -185.886	ARIMA$(0,0,2)$ 111.173	ARIMA$(1,0,1)$ -181.337	ARIMA$(1,0,0)$ -1874.948	0.0002
	Fourth over	ARIMA$(1,1,1)$ -137.988	ARIMA$(1,0,2)$ 30.714	ARIMA$(1,0,0)$ 38.913	ARIMA$(1,0,0)$ -1865.572	0.0002
FDM	Order	MAE	설명력			
	1	0.081	70.0%			
	2	0.026	$96.8\% = 70\% + 26.8\%$			
	3	0.016	$98.9\% = 70\% + 26.8\% + 2.1\%$			
	4	0.014	$99.2\% = 70\% + 26.8\% + 2.1\% + 0.3\%$			
Bayesian	MAE	AR(1) model				
	0.212	$f_{c,t+5}(c) - 1.5 = 0.886(f_{c,t} - 1.5)$, $\sigma_\epsilon^2 = 0.120^2$				

주) 설명력은 직교 기저함수들의 설명력(percentage variantion due to basis function)을 의미함.
MAE = mean absolute error; GLG = generalized log gamma; FDM = functional data model.

끝으로 지금까지 설명한 개별 출산율 모형들의 모수 적합과 MAE, MAPE, 베이지안 AR(1) 모형식 등은 Table 3.4에 요약하였다. Table 3.4는 GLG 모형의 출산순위(1–4아 이상) 모수적합 결과와 적합도, 비모수 FDM의 적합도와 주성분수에 따른 설명력, 그리고 베이지안 모형의 적합도와 출산 전이 3단계 예측의 AR(1) 모형을 제시하고 있다. 적합도 측면에서는 모수적 모형이 우수한 것으로 판단된다. 보다 자세한 결과비교는 3.2절에서 살펴본다.

3.2. 결과 비교

최근 3년간 합계출산율을 반영한 모수와 비모수 예측결과는 2020년에 1.04–1.05명에 근접함을 보인 반면 2016년 정기적인 추계로 최근 데이터를 반영하지 못한 통계청 TFR과 WPP 2017 자료의 베이지안 결과와는 차이를 보인다.

GLG모형을 활용하여 TFR을 살펴본 모수적 접근은 2045년 1.21명 수준까지, 비모수적 접근은 동일 연도 1.19명으로 GLG 결과보다는 약간 낮게 예측되었으나 베이지안 방법은 2020년 1.18명 수준에서 2045년 1.22명 수준으로 도출되어 모수와 비모수 결과보다 높은 수준을 보인다.

이들 모형에 대해 추정 모수 개수의 계산 효율성, 모형 적합도와 정확도, 그리고 결측치에 대한 강건도 관점에서 비교해 보고자 한다.

첫째, 추정 모수 개수의 계산 효율성 측면이다. GLG모형은 추정 모수 4개, FDM모형은 평활과 주성분 수가 1이상이라면 추정 모수는 적어도 4개 이상, 베이지안모형은 최소 모수 12개 이상을 추정해야 한다. 따라서 효율성과 편의성 측면에서는 모수, 비모수, 베이지안 순이다.

둘째, 모형 적합도와 정밀성(precision)이다. 비교를 위해 MAE를 활용한다. GLG모형, FDM모형, 베이지안모형의 MAE 수치는 $2.295 * 10^{-9}$, 0.026, 0.212이다. 우리나라 ASFR자료를 기준으로 고려했을 때 베이지안모형보다는 모수, 비모수 모형의 적합도가 상대적으로 우수하다.

셋째 강건도이다. 비모수는 자료 결측치나 이상치가 있을 경우 평활방법으로 추세를 유지하는 방법론이므로 자료 완비성에 강건한 방법으로 알려져 있다. 베이지안도 자료 품질이나 부재의 경우에 고려된 방식이다. 하지만 모수적 접근은 과거의 추세가 미래에도 유지된다는 가정에 입각해서 추정하기 때문에, 만약 결측이 존재한다면 두 적합치 보다는 우수하지 못할 수 있다.

따라서 TFR 예측에 있어 자료 완비성이 높고 예측에 대한 계산 효율성을 고려한다면 모수적 방법, 이를 충족하지 못한 경우라면 비모수와 베이지안 방법을 적용하는 것을 제안한다.

4. 결론 및 제언

본 논문은 변동이 작지 않고 2005년 TFR 수준으로 회귀하는 우리나라 출산율에 대해서 3가지 출산율 모형을 적용한 결과를 비교하여 어느 방법이 더 우수하고 합리적인지 살펴보았다.

예측치 비교에서는 대체적으로 통계청 TFR이 가장 높고, 베이지안, 모수, 비모수 순으로 나타났다. 2017년 TFR 1.05명 수준이 판단기준일 때 모수, 비모수적 모형으로 도출된 TFR 예측값이 합리적이다. 그리고 모수추정이나 계산 효율성과 적합도 관점에서는 모수적 방법이 타 방법보다 우수한 것으로 드러났다. 단, 출산율 자료 완비성이 높고 품질이 우수한 경우에만 성립한다. 만약 출산율 자료의 결측이나 이상치가 발생하는 경우는 비모수적 방법, 데이터 완비성과 출산율 자료 품질이 떨어진다면 베이지안 방법을 고려해야만 한다.

본 연구를 통해 한계점과 몇 가지 향후 연구방향을 도출할 수 있다. 한계점으로는 첫째, GLG모형의 C_1과 C_2 출산순위 예측이다. 본 연구는 출산율 미래 추이의 복원성 가정으로 정상시계열을 간주하여 분석하였다. 만약 미래 출산율이 지속적인 감소를 보인다면 출산율 전이 2, 3단계가 비정상 시계열에 해당되므로 예측 결과는 다를 것이다.

둘째, 베이지안의 MCMC 과정에 대한 명확성이다. Raftery 등 (2014)의 논문에서도 사전분포는 밝히고 있으나 각 모수들의 사후분포는 제시하지 않고 있다. 이들에 대한 상세한 생성과정과 방법은 향후연구로 남겨야 할 부분으로 판단된다.

다음으로 향후 연구방향이다.

첫째 통계청은 17개 지방자치별 출산율 추계를 실시하고 공식통계를 제공하고 있다. 하지만 읍면동 규모의 소지역 인구추계 결과는 제공하고 있지 않고 있어 각 지방자치별로 추계를 실시하고 있다. 이러한 현실에서 제안할 수 있는 하나의 방법은 베이지안 방법론이다. 예를 들어 경기도 합계출산율은 경기도 각 소지역(시, 군, 읍, 면, 동) 출산율과 유사할 것이라는 가정 하에서, 상위 큰지역 출산율을 사전분포로 생성하고 이를 통해 소지역에 대해서 사후분포를 생성해 추계하는 방법을 제안해 볼 수 있다.

둘째, 우리나라처럼 변동이 큰 출산율의 경우에는 출산율 패턴이 미래에도 계속되어야 한다는 가정에 기초한 모수적 접근 방식에 개선이 필요하다. 출산연령의 상승과 ASFR 중심축이 비대칭에서 대칭으로 전환되고 있으므로 GLG 모형이 아닌 서로 다른 분포가 합쳐진 출산율 모형을 제안해 볼 수 있다.

셋째, 일부지역에 특별한 사건이나 정책 등으로 출산율이 갑자기 높거나 낮아질 경우, 그 이후 추이는

앞 시기와 동일하지 않을 가능성이 높다. 이러한 경우에 평활방법을 적용한 비모수 방법을 적용해 이상 패턴의 영향을 최소화하여 합리적인 결과를 도출하는 연구가 필요하다.

넷째, 출산율 예측은 최근 시계열에 민감하다. 특히 변동이 큰 경우 예측값 도출시 최근 데이터에 의존하는 경향이 강하다. 모수, 비모수적 방법의 2016년, 2017년 ASFR을 반영한 결과와 WPP 2017 베이지안 옵션 설정에서 그 결과는 확인 되었다. 따라서 인구동태자료나 행정자료를 확보하여 실시간 시계열 구축이 중요하다.

부록: Bayesian hierarchical models for Korea fertility by R program

```
library(demography);library(bayesTFR);library(WPP2017)
setwd("C:/User/user/Desktop/data/bayesian"); getwd()
simulation.dir←file.path(getwd(),"bayesian")
# Starting a new Simulation #
   # 출산전이 3단계와 한국의 출산율 추이를 고려한 Δc4의 최대치, 최소치, 간격을 조정 #
m1←run.tfr.mcmc(nr.chains=2,iter=1000,output.dir=simulation.dir,replace.output=TRUE,
Trangle-c4.low=0.9,Trangle-c4.low=1.1,Trangle-c4.trans.width=1.0)
# Continuing an existing simulation #
m2←continue.tfr.mcmc(iter=1000,output.dir=similation.dir)
# Accessing existing MCMC results #
m3←get.tfr.mcmc(sim.dir=simulation.dir); m3.chain2←trf.mcmc(m3.chain.id=2)
# Generating new predictions and Accessing an existing projection #
pred1←tfr.predict(sim.dir=simulation.dir,end.year=2050,mu=1.5,min.tfr=0.8,burnin=1000,
nr.traj=100,verbose=TRUE)
   # 출산전이 3단계 후, 예측을 위한 목표값(1.5명), 최소값(0.8명), 미래 불확실성 궤적(100) #
pred2←get.tfr.prediction(sim.dir=simulation.dir)
# Results, Summary, and Diagnose #
summary(m3,country="Republic of Korea",par.names=NULL,thin=5,burnin=1000)
summary(pred2,country="Republic of Korea")
tfr.trajectories.plot(pred2,country="Republic of Korea",pi=c(95),nr.traj=100)
DLcurve.plot(country="Republic of Korea",mcmc.list=m3,burnin=300,pi=c(95),nr.curves=100)
tfr.diagnose(sim.dir=simulation.dir,thin=5,burnin=1000,express=FALSE,
country.sampling.prop=NULL,keep.thin.mcmc=FALSE,verbose=TRUE)
```

References

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). probabilistic projections of the total fertility rate for all countries, *Demography*, **48**, 815–839.

Basten, S. A., Coleman, D. A., and Gu, B. (2012). Re-examining the fertility assumptions in the UN's 2010 World Population Prospects: Intentions and fertility recovery in East Asia? *Presented at the Annual Meeting of the Population Association of America; San Francisco.*

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation*, John Wiley & Sons, New York.

Chandola, T., Coleman, D. A., and Horn, R. W. (1999). Recent European fertility patterns: fitting curves to 'distorted distributions, *Population Studies*, **53**, 317–329.

Eom, J. M. and Kim, K. W. (2013). A study on forecasting total fertility rate using female first marriage rate, *Journal of the Korean Data Analysis Society*, **15**, 1261–1272.

Evans, M. D. R. (1986). American fertility patterns: a comparison of white and nonwhite cohorts born 1903–1956, *Population and Development Review*, **12**, 267–293.

Hadwiger, H. (1940). Eine Analytische Reprodutions-Funktion fur Biologische Gesamtheiten, *Skandinavisk Aktuarietidskrift*, **23**, 101–113.

Hoem, J. M., Madsen, D., Nielsen, J. L., Ohlsen, E., Hansen, H. O., and Rennermalm, B. (1981). Experiments in modelling recent Danish fertility curves, *Demography*, **18**, 231–244.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration, *International Journal of Forecasting*, **24**, 323–342.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: the product-ration method with functional time series models, *Demography*, **50**, 261–283.

Hyndman, R. J. and Ullah, S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach, *Computational Statistics & Data Analysis*, **51**, 4942–4956.

Jun, K. H. (2006). Development of fertility assumptions for the future population projection, *Korean Journal of Population Studies*, **29**, 53–88.

Kaneko, R. (2003). Elaboration of the Coale-McNeil Nuptiality model as the generalized log gamma distribution: a new identity and empirical enhancements, *Demographic Research*, **9**, 223–262.

Kang, K. H. and Ahn, H. S. (2006). Functional data analysis of temperature and precipitation data, *The Korean Journal of Applied Statistics*, **19**, 431–445.

Kim, K. W. and Jeon, S. B. (2015). Scenario analysis of fertility in Korea using the fertility rate prediction model, *The Korean Journal of Applied Statistics*, **28**, 685–701.

Kim, S. Y. and Oh, J. H. (2017). A study comparison of mortality projection using parametric and non-parametric model, *The Korean Journal of Applied Statistics*, **30**, 701–717.

Kim, S. Y., Oh, J. H., and Kim, K. W. (2018). A comparison of mortality projection by different time period in time series, *The Korean Journal of Applied Statistics*, **31**, 41–65.

KOSIS (2011, 2016). Population Projections (2010∼2060), Population Projections (2015∼2065).

Lutz, W., Skirbekk, V., and Testa, M. R. (2006). The low-fertility trap hypothesis: forces that may lead to further postponement and fewer births in Europe, *Vienna Yearbook of Population Research*, 167–192.

Neal, R. M. (2003). Slice Sampling, *The Annals of Statistics*, **31**, 705–767.

Park, Y. S., Kim, M. R., and Kim, S. Y. (2013). Probabilistic fertility models and the future population structure of Korea, *The Korean Association for Survey Research*, **14**, 49–78.

Peristera, P. and Kostaki, A. (2007). Modeling fertility in modern populations, *Demographic Research*, **16**, 141–194.

Raftery, A. E., Alkema, L., and Gerland, P. (2014). Bayesian population projections for the United Nations, *Statistical Science*, **29**, 58–68.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G. K. (2012). Bayesian probabilistic population projection for all countries. In *Proceedings of the National Academy of Sciences of United States of America*, **109**, 13915–13921.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed), Springer-Verlag, New York.

Ryder, N. B. (1990). What is going to happen to American fertility?, *Population and Development Review*, **16**, 433–454.

Ševčíková, H., Alkema, L., and Raftery, A. E. (2011). bayesTFR: an R package for probabilistic projections of the total fertility rate, *Journal of Statistical Software*, **43**, 1–29.

Ševčíková, H., Alkema L., Raftery, A. E., Fosdick, B., and Gerland, P. (2018). *Package 'bayesTFR'*, R demography.

UN (2017). World Population Prospects 2017.

모수, 비모수, 베이지안 출산율 모형을 활용한 합계출산율 예측과 비교

오진호[a,1]

[a]통계청 통계개발원

(2018년 6월 28일 접수, 2018년 8월 3일 수정, 2018년 10월 24일 채택)

요 약

최근 2017년 우리나라 합계출산율은 1.05명로 2005년 1.08명 수준으로 회귀하는 현상을 보이고 있다. 1.05명은 인구대체선(2.1명), 안전선(1.5명)과도 거리가 먼 초저출산 수준이고 마치 초저출산 덫에 빠질 우려가 있다. 이에 합계출산율의 합리적인 예측과 이를 통한 출산정책에 유용한 자료를 제공하는 것은 그 어느 때 보다도 중요하다. 그동안 다양한 통계적 방법으로 합계출산율 추이를 예측하였는데, 데이터 완비성이 높고 품질이 좋은 경우 모형 접근인 모수적 방법, 데이터 추이가 단절되거나 변동이 심한 경우 평활과 가중치를 적용한 비모수적 방법, 데이터 부족과 품질 등으로 선진국의 출산율 3단계 전이현상을 참고하여 이들의 사전분포를 활용하는 베이지안 방법 등이 적용되어 왔다. 본 연구는 최근 변동이 심한 우리나라 출산율에 모수, 비모수, 그리고 베이지안 방법을 적용하여 추정과 예측을 실시하고 도출된 결과 비교를 통해 적합성과 타당성 측면에서 어떤 방법이 합리적인지 모색하고자 한다. 분석 결과 합계출산율 예측값 순위는 통계청 합계출산율이 가장 높고, 베이지안, 모수, 비모수 순으로 나타났다. 2017년 TFR 1.05명 수준을 감안할 때 모수, 비모수모형으로 도출된 합계출산율 예측값이 합리적이다. 또한 출산율 자료 완비성이 높고 품질이 우수할 경우 계산 효율성과 적합도 관점에서 모수적 추정과 예측 접근 방법이 타 방법보다 우수한 것으로 도출되었다.

주요용어: 합계출산율, 초저출산 덫, 모수적 모형, 비모수적 모형, 베이지안 모형

본 논문은 통계청의 공식견해가 아니며 저자의 개인적인 연구결과임을 밝힙니다.

[1](35220) 대전광역시 서구 한밭대로 713 통계센터 통계개발원 6층 통계분석실. E-mail: comet123@korea.kr