The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.4.421
33(4), 421–437

Development of system of Population projection and driving variation on demography for Korea using R

Jinho Oh[a,1]

[a]Department Mathematical Sciences, HanBat National University

(Received May 20, 2020; Revised June 30, 2020; Accepted July 11, 2020)

Abstract

This paper implemented a method to predict the fertility rate, mortality rate, and international migration rate using the R program, which has been widely used in recent years, that calculates population projection by substituting the results into the Leslie matrix. In particular, the generalization log gamma model for the fertility rate by Kaneko (2003), LC-ER model for mortality rate by Li *et al.* (2013), and functional data model for international migration rates proposed by Ramsay and Silverman (2005) and Hyndman and Booth (2008), Hyndman *et al.* (2013) can be directly demonstrated with R programs. Demography and bayesPop have been introduced as a representative demographic package implemented in R; however, it can be analyzed only for data uploaded to Human Mortality Database (HMD) and Human Fertility Database (HFD) with data changes and modifications requiring application of other data. In particular, in Korea, there is a limitation in applying this package because it is provided only for short-term data in HMD. This paper introduces an R program that can reflect this situation and the different patterns of low fertility, aging, migration of domestic and foreigners in Korea, and derives a population projection for the year 2117.

Keywords: fertility rate, mortality rate, migration rate, Leslie matrix, population projection

1. 서론

통계청 장래인구특별추계 (KOSTAT, 2019)는 인구주택총조사(Ceusus) 결과를 기초로 인구동태통계(vital statistics)와 국제인구이동통계를 활용하여 코호트요인법(cohort component methods)에 의해 2067년까지 사회특성별 인구구조를 제공한다. 코호트요인법은 출생, 사망, 국제이동의 장래 변동수준에 대한 가정을 바탕으로 추계의 시작점이 되는 사회특성별 기준인구(base population)에 출산과 국제순이동은 더하고 사망은 감하는 산식에 의해 다음해 인구를 도출하는 것이다. 최근 합계출산율(total fertility rate; TFR) 1.00명 아래로 떨어진 출산율 저하(low fertility rate)와 연령별 사망률 개선(mortality declined rate)에 따른 고령화(aging)와 내국인과 외국인의 상이한 국제이동 패턴 등과 같은 다양한 미래 인구구조 변화를 반영한 장래인구추계의 수요가 급증하고 있다. 또한 지방소멸, 노노부양, 독거노인 등과 같은 신조어 출현으로 인구구조가 예전과 다른 형태로 변모되고 있다.

더불어 장래인구추계는 국가의 중장기 경제, 사회 발전 계획 수립의 기반이 되고 사회보장체계를 설계하는 재정과 연금 정책 뿐만 아니라 노동, 교육, 산업, 환경, 주택 등의 기초자료로 활용도가 높아지고

[1]Department Mathematical Sciences, HanBat National University, 125 Dongseodaero, Yuseong-gu, Daejeon 34158, Korea. E-mail: jhoh75@hanbat.ac.kr

있다. 따라서 미래 인구구조와 사회변화를 미리 조명하기 위해서는 정밀한 인구추계가 필요하고 누구나 쉽게 직접 인구추계를 도출할 수 있는 통계적 도구(tool)가 필요하고 요구된다. 최근 인구추계 패키지로 소개되고 있는 R 프로그램으로 Hyndman 등 (2019)의 demography와 Ševčíková와 Raftery (2016), Ševčíková 등 (2019)의 bayesPop가 대표적이다.

demography는 Lee-Carter (1992) 모형, 사망률, 출산율, 국제순이동자수를 추정하기 위한 함수적 데이터 모형(functional data model; FDM), 확률론적 인구추계(stochastic Population projection) 등을 구현할 수 있는 R 프로그램이다. 이는 Human Fertility Database (HFD), Human Mortality Database (HMD)에서 출생율과 사망률 자료를 로드하여 추정하고 예측한다. 국제순이동자수는 HMD의 자료에 내재된 형태와 유사한 자료를 국가별로 집계하여 FDM모형에 대입하여 추정한다. bayesPop는 United National Population Division (UNPD)에서 2년간격으로 공표하고 있는 World Population Prospects (WPPs) 자료를 기초로 합계출산율과 기대수명을 베이지안(Bayesian)을 적용한 확률적모형으로 인구변동을 추정하여 세계 200여국의 인구를 제공한다. 이는 세계 모든 국가들이 인구동태통계와 센서스의 자료완비성(completeness)과 품질(quality)이 동일하지 않기 때문에 선진국의 출산, 사망패턴을 사전확률(prior probability)로 개발도상국과 기타국가 등의 인구변동요인 분포를 사후확률(posterior probability)로 간주하고, 이들 국가는 선진국의 출산전이단계, 사망률의 개선 패턴과 유사할 것이다라는 가정하에 베이지안 방법을 적용하여 사후확률을 추정하여 최종적으로 인구를 추계한다.

하지만 이들 모두 HMD, HFD, 그리고 UNPD에서 자체적으로 추정한 WPP 2019의 자료로만 구현이 가능하고 타 출처의 자료를 적용하기 위해서는 자료 가공과 변경이 요구될 뿐만 아니라 우리나라처럼 다양한 인구변동을 반영하는 것이 용이하지 않다.

이에 본 연구는 인구관련 선행논문과 demography, bayesPop을 참고하여 인구추계를 위한 R 프로그램을 소개하고, 인구통계에 관심이 있는 독자들을 위한 활용방법과 구현법을 제공하고자 한다. 그리고 본 연구는 다음과 같은 점에서 선행연구와 다른 차별성을 부여할 수 있다. 첫째 인구추계의 논문과 연구는 그간 많이 진행 되었지만 관심 독자를 위한 자가적(self-in or hands on) 프로그램 소개는 없었다. 이런 제공을 통해 인구관련 독자들에게 인구추계 방법과 관심도를 높일 수 있을 것으로 판단된다. 둘째 우리나라 통계청 국가통계포털(KOSIS)자료를 활용하여 자료 변환 없이 R로 불러와 출산율, 사망률, 국제이동율을 추정, 예측하고 그 결과를 종합하여 코호트요인법에 의해 장래인구추계를 도출할 수 있는 통계적 도구를 안내한다.

본 논문은 총 4개장으로 구성한다. 제 2장은 인구변동 3요인의 예측모형과 장래인구추계 방법론을 소개한다. 제 3장은 R 프로그램을 구현한 장래인구추계와 결과를 제시한다. 끝으로 제 4장은 결언으로 연구요약과 연구의 한계점 제시에 따른 향후연구방향과 시사점 등을 제언한다.

2. 장래인구추계 산정 로직

2.1. 인구변동 3요인 모형과 추정방법

통계청 국가통계포털(KOSIS)의 국내통계-인구, 가구 항목에서 1970–2018년까지의 연령별출산율(age-specific fertility rate; ASFR)과 연령별사망률(age-specific mortality rate; ASMR), 2000–2018년까지의 국제이동률(international migration)을 내국인과 외국인을 구분하여 실측치를 제공한다. 출산율은 가임연령(15–49세)에 대해서 연도별 기간별(period) ASFR, 사망률은 0–99세, 100세 이상에 대해서 연도별 기간별 생명표(life table)와 기대수명(life expectancy), 기대여명(residual expectation of life)이다. 국제이동률은 내국인, 외국인의 상이한 국제이동 패턴으로 분리해서 0–84세, 85세 이상으로 연령별 이동률이다. 더불어 장래인구특별추계 (KOSTAT, 2019)에서 2067년까지 동일한 형태로 인구변동 3요

인(출생, 사망, 국제이동)에 대한 예측 자료를 제공한다.

본 연구는 2018년까지의 실측치와 2067년까지의 예측치(통계청의 실행가능성이 높은 중위수준)를 합친 시계열 자료를 활용하여 2117년까지 약 40년을 연장하는 추계 결과를 제시하고자 한다. 이전 선행연구(Oh, 2018, 2019, 2020)는 2018년 실측치를 토대로 2067년까지의 출산율, 사망률, 국제이동률을 제안한 것이고, 이번 연구는 보다 더 연장해 장기 시계열인 2117년까지 선정해 예측한다. 이는 연금과 재정정책 등 국가 중장기 경제, 사회 발전계획의 기초자료 제공과 장래가구추계 등 인구를 활용한 다양한 주제별 추계의 기초자료 제공 등으로 100년간의 추계결과를 제공하고 있는 통계청의 장래인구추계의 시계열과 동일하다. 그리고 본 연구는 2067년까지 1년 1세 간격으로 인구추계 결과를 제공하고 그 이후는 총인구, 출생아수, 사망자수, 국제순이동만 공표하는 통계청의 국가통계에 인구변동 3요인의 비율의 변화를 추가적으로 제공하는 것이다.

먼저 장래인구추계를 위해 기준인구와 남녀출생성비(ratio of male to female births) 정보가 요구된다. 기준인구는 추계시작점의 주민등록연앙인구(당해연도 7월 1일자 인구)를 의미하며, KOSIS에 제공되고 있다. 그리고 우리나라는 남아선호사상으로 남녀 출생성비를 1.05(이는 통계청이 가정하고 있는 값과 동일)로 간주한다. 다음으로 2068–2117년 인구변동요인(출산, 사망, 이동)의 미래값을 도출하는 것인데 미래는 불확실성(uncertainty)이 내재되어 있으므로 확률론적 추계가 요구된다. 확률론적 추계 방법은 대표적으로 과거예측 오차분석법(historical forecast error method), 통계모형, 전문가 판단법, 인구변화 요인의 확률변수화로 나누어진다 (Lee, 1998). 이들 중 본 연구는 일반적이고 객관적 모형접근인 통계모형으로 추정과 예측하는 방식을 채택한다. 이들에 대한 개별 모형을 아래에 소개한다.

첫째 출산율이다. 우리나라 ASFR 형태가 이단봉(유럽형 ASFR 모형, 20세 초반과 28–32세에 출산율이 높음)과 다르게 시간에 따라 왼쪽으로 치우진 분포(1970–1990년대는 20대 중후반이 출산율이 높음)에서 정규분포(32세를 중심으로 좌우 대칭)로 변화되어가는 양상을 띠고 있다. 이런 변화를 표현하기 위해 필요한 모수는 위치(location), 산포(dispersion), 형상(shape) 모수와 한 코호트가 출산율 경험할 확률을 나타내는 모수가 모형에 포함되어야 한다. Park 등 (2013)논문에서는 유럽형 외에 다양한 출산율 모형을 소개하고 있으며 우리나라 출산율 모형은 혼합정규분포를 따른다고 제안한다. 통계청(KOSTAT, 2010, 2016, 2019)은 위 4개의 모수를 포함하는 Kaneko (2003)가 제안한 일반화 로그 감마 모형(generalized log Gamma model; GLG)을 채택하고 있다.

일반적으로 모수가 많이 포함된 모형은 적합력은 좋은 반면 추정과정이 용이하지 않고 모수 간소화 측면과 우리나라 출산형태가 일본과 유사하고 이단봉이 아닌 점을 고려해 최종적으로 지수족(exponential family)인 GLG 모형을 채택한다. GLG 모형은 출산순위별 출산모형에 확률분포의 개념을 적용하고 모수 추정에 이론적 기반이 있는 통계적 방법을 이용할 수 있는 장점이 있다. 식 (2.1)은 GLG 모형인데 모수 4개(C_i, μ, b, λ)를 포함하고 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 확률이다. μ와 b는 출산연령의 평균과 표준편차, λ는 분포형태를 나타내는 모수이다. 만약 C_i의 출산순위를 고려하지 않는다면 합계출산율과 동일하다. 예측방법은 GLG 모형으로 실측치 시계열만큼 모수를 추정한 후 이 추정된 4개 모수들을 시계열 모형에 적합하여 사용자가 원하는 분석기간에 대해 예측하고 이들 값을 근거로 미래의 ASFR을 산출한다. 따라서 GLG 모형은 유배우의 출생순서(birth-order)에 대한 연령패턴의 규칙성(regularity)을 수학적으로 표현 (Kaneko, 2003)한 것이다.

둘째 사망률이다. 최근 사망률이 낮은 선진국과 우리나라는 예전에 비해 유소년(infant and child)

ASMR 개선은 점진적으로 감소하고, 고령층(65세 이상)의 ASMR 개선은 점점 빨라지는 사망률 개선 교대(declined mortality rate with rotation)현상이 발생하고 있다 (Horiuchi와 Wilmoth, 1995; Li와 Gerland, 2011; Li 등, 2013; Kim과 Oh, 2017). 이는 유소년층과 노년층의 ASMR 감소 패턴이 교대(rotation)되는 현상이고, 기대수명이 증가함에 따라 발생하고 있다. 즉 LC 모형 [$\ln(m_{x,t}) = a_x + b_x k_t + \epsilon_{x,t}$]에서 ASMR 변화를 나타내는 b_x가 시간의 경과에 따라 변화고 있음을 의미한다. Li와 Gerland (2011)는 이와 유사한 현상을 사망률 개선 교대라 명명하고 시간의 변화에 따라 ASMR 개선 교대현상을 반영하기 위해 LC 모형의 b_x에 강건한 순환(robust rotation)을 도입하여 주관적이고 강건한 수정을 간주하는 Lee-Carter method with robust rotation (LC-RR) 모형을 제안한다. 그리고 2년 뒤 Li 등 (2013)은 LC-RR 모형의 회전모델을 유연하게 변화를 줄 수 있도록 식 (2.2)와 같은 Lee-Carter method extended to model the rotation (LC-ER) 모형을 제안한다. 즉, 이는 LC 모형에서 시간의 변화에 관계없이 일정한 b_x에 시간 변수를 고려한 $B_{x,t}$형태로 변환한 것이며, ASMR 개선 패턴이 매년 동일하다는 단점을 보완한 것이다.

그리고 통계청(KOSTAT, 2016)은 사회, 경제적 조건이 유사한 인구 부집단(예:시도별)의 사망률을 고려하는 LL (Li와 Lee, 2005) 모형 [$\ln(m_{x,t,i}) = a_{x,i} + B_x K_t + b_{x,i} k_{t,i} + \epsilon_{x,t,i}$]과 LC-ER를 조합한 LL&LC-ER을 제안(통계청은 이를 LLG 확장 모형으로 명명)한다. LL 모형은 사회 경제적인 조건이 유사한 인구집단의 ASMR 패턴은 장기적으로 공통 사망률로 수렴할 것이라는 가정 하에 기존의 LC 모형을 다중인구로 확장하기 위해 공통사망경향으로 전체 집단의 log(ASMR) 변화정도(B_x)와 시간에 따른 log(ASMR) 수준의 변화(K_t), i번째 인구집단 그룹의 개별 사망 경향으로 개별 집단의 i의 log(ASMR) 변화 정도($b_{x,i}$)와 시간에 따른 log(ASMR) 수준변화($k_{t,i}$)를 나타낸다. 그런데 본 연구는 전체인구를 산정하는 것이므로 식 (2.2)의 LC-ER 모형으로 한정지어 사망률 개선 교대현상을 설명한다.

$$\ln(m_{x,t}) = a_x + B_{x,t} K_t + \epsilon_{t,x}, \tag{2.2}$$

여기서 $B_{x,t}$는 교대 이전의 ASMR인 LC 모형의 b_x와 사망률 개선의 최고점 사망률 b_x^{μ}의 선형 가중평균으로 식 (2.3)과 같이 정의되며, e_0^{μ}는 교대 현상이 끝나는 시점의 기대수명을 나타낸다. LC-ER 모형은 선진 20개국의 사망률 분석에 근거해서 몇 가지 실험적인 결과값을 이용한 몇 가지 주관적인 가정을 활용한다. 식 (2.3)의 ASMR 개선 패턴은 연령 80세에 도달했을 때 교대하기 시작해서 102세까지 지속되며, 이후에는 일정하다고 간주한다. 이는 65세 미만 모든 연령에서 사망률 개선패턴은 동일하며 그 이후에 연령이 증가할수록 작아짐을 의미하고, $B_{x,t}$의 임의의 선형변환으로 K_t는 기대수명이 교대 없이 얻은 LC의 k_t와 비교하여 이 둘의 차가 가장 작은 K_t를 구하기 위해 반복적인 계산으로 도출된다. 이런 과정을 거치면 b_x^{μ}와 $B_{x,t}$을 도출한 후 성별과 K_t값으로 e_0^{μ}을 예측할 수 있다.

$$B_{x,t} = \begin{cases} b_x, & e_0^t < 80, \\ (1 - w_s(t)) b_x + w_s(t) b_x^u, & 80 \le e_0^t < e_o^u, \\ b_x^u, & e_0^u < e_0^t, \end{cases} \tag{2.3}$$

$$b_x^u = \begin{cases} \bar{b}_{15-64}, & 0 \le x \le 64, \\ b_x \times \dfrac{b_{u,60-64}}{b_{65-70}}, & 65 \le x, \end{cases}$$

$$w(t) = \frac{e_0^t - 80}{e_0^u - 80}; \quad w_s(t) = \left[0.5 \left\{ 1 + \sin\left(\frac{\pi}{2} (2w(t) - 1) \right) \right\} \right]^p.$$

그리고 식 (2.3)을 성별에 따라 확장하면 남자(male; m)은 $\ln(m_{x,t,m}) = a_{x,m} + B_{x,t} K_{t,m} + \epsilon_{x,t,m}$, 여자(female; f)는 $\ln(m_{x,t,f}) = a_{x,f} + B_{x,t} K_{t,f} + \epsilon_{x,t,f}$이다.

셋째 국제이동이다. Ramsay와 Silverman (2005)과 Hyndman과 Ullah (2007), Hyndman과 Booth (2008), Hyndman 등 (2013)은 함수적 자료 분석 패러다임(functional data analysis paradigm)을 사용하여 출산율, 사망률, 국제이동률을 모델링하고 예측하기 위한 비모수적 방법인 함수적 자료 모형(functional data model; FDM)을 제안한다. 그들은 관측치에 존재하는 측정오차와 질병, 기아, 전쟁 등으로 인구동태 자료에서 나타나는 불규칙적인 패턴을 교정하기 위해 함수적 자료분석을 이용하여 식 (2.4)와 같은 모형을 구축하고 비모수 평활기법을 이용한다. 제안된 FDM 모형의 구조는 식 (2.4)와 (2.5)와 같다.

$$f_t(x) = \begin{cases} \frac{1}{\lambda}\left(f_t^*(x)^\lambda - 1\right), & 0 < \lambda < 1, \\ \ln\left(f_t^*(x)\right), & \lambda = 0, \end{cases} \tag{2.4}$$

여기서 $f_t(x)$는 시간 t와 연령 x에서 국제이동률을 의미한다. $f_t^*(x)$의 Box-Cox (1964)변형은 $f_t^*(x)$의 값에 따라 증가하는 변동을 줄여주거나 정규화과정으로 λ는 Box-Cox 변형에서 강도를 뜻한다.

$$\begin{aligned} f_t(x) &= s_t(x) + \sigma_t(x)\epsilon_{t,x}, \\ s_t(x) &= \mu(x) + \sum_{i=1}^{J} \beta_{t,j}\phi_j(x) + e_t(x), \end{aligned} \tag{2.5}$$

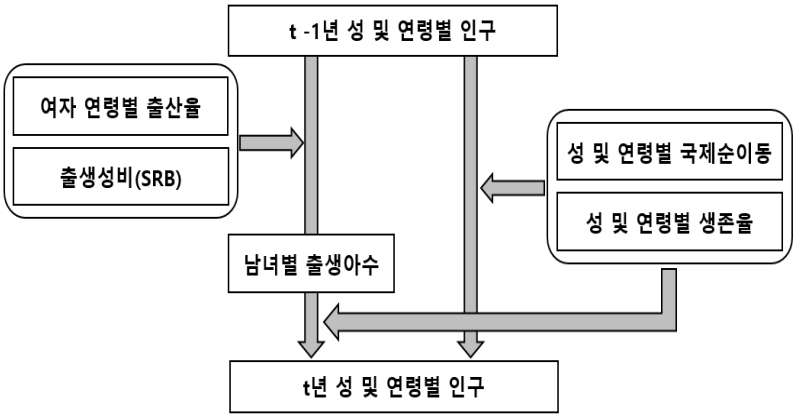
여기서 $\mu(x)$는 $\sum_{t=1}^{n} s_t(x)/n$에 의해 추정된 평균함수로 평활된 연령에 따른 로그국제이동률평균이고, $\beta_{t,j}\phi_j(x)$ $(t = 1, \ldots, n,\ j = 1, \ldots, J)$는 함수적 주성분분석을 사용하여 추정되어지며 $J < n$는 사용된 주성분 수이다. $\Phi = \{\phi_1(x), \ldots, \phi_J(x)\}$는 J개의 함수적 주성분의 집합으로 직교 기저함수(orthogonal basis function)이고 $B = \{\beta_{t,1}, \ldots, \beta_{t,J}\}$는 비상관 주성분 점수(uncorrelated principal component scores)들의 집합으로 시계열 계수를 의미한다.

식 (2.5)에서 $f_t(x)$는 시간 t의 연령 x에 대한 관찰된 로그국제이동률 $\ln f_t(x)$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 로그국제이동률과 평활된 곡선의 차이인 관측오류를 의미한다. 식 (2.5)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. FDM은 첫 번째 주성분에 직교하는 고차원 주성분에 대해서는 다른 시계열 모형들의 주성분 점수가 도출된다. 모든 성분에 FDM 방법은 최적 시계열 모형을 AIC 등과 같은 모형 판별 기준에 의거하여 선택한다. 이 모형에 대한 보다 자세한 설명은 Hyndman과 Booth (2008), Hyndman 등 (2013), Kim과 Oh (2017), Oh (2018)를 참조하면 된다. 참고로 FDM 예측은 $Z = \{f_1(x), \ldots, f_n(x)\}$와 $\Phi = \{\phi_1(x), \ldots, \phi_J(x)\}$의 조건부로 $f_{n+h}(x)$의 h단계 예측치를 구할 경우 식 (2.6)에 의해서 도출 가능하다.

$$\hat{f}_{n+h|n}(x) = E[f_{n+h}(x)|Z, \Phi] = \hat{\mu}(x) + \sum_{j=1}^{J} \hat{\beta}_{n+h|n,j}\phi_j(x), \tag{2.6}$$

여기서 $\hat{\beta}_{n+h|n,j}$은 Hyndman과 Booth (2008)에 의한 지수 평활 또는 ARIMA 모형과 같은 일변량 시계열모형을 활용하여 도출한 $\beta_{n+h,j}$의 h 단계 예측을 의미한다.

지금까지 소개한 인구변동 3요인의 출산율, 사망률, 국제이동률을 구현하기 위한 통계프로그램인 R 구현법과 R에서 제공하는 일부 유용한 패키지 활용 등의 설명과 개요를 3장에서 논하기로 한다.



출처: KOSIS (2019), 장래인구특별추계: 2017–2067

Figure 2.1. A process for population projection by cohort component methods.

2.2. Leslie 행렬을 이용한 코호트요인법

앞 절에서 2117년까지 인구변동 3요인을 추정, 예측하는 방법에 대해 살펴보았다. 이번 절은 이들 요인들을 코호트요인법에 적용해 인구방정식 식 (2.7)에 따라 2068–2117년 인구를 추계한다. 코호트요인법이란 특정 연도의 성 및 연령별 기준인구에 인구변동 요인인 출생, 사망, 국제이동에 대한 장래변동을 각각 추정하여 조합하는 방법이다. 따라서 이 방법은 성 및 연령별로 많은 자료를 필요로 하는 반면에 인구구조를 바탕으로 인구변동 요인을 감안하고 인구학적 해석의 용이성, 가정설정 등의 명확성에 의한 대부분의 공식 인구추계 때 사용(STI, 2013)하고, 인구변동 요인별 미래 수준을 각각 예측한 후, 추계의 출발점이 되는 기준인구에 출생아수와 국제순이동은 더하고, 사망자수는 빼는 인구균형방정식(demographic balancing equation)을 적용하여 다음 해 인구를 반복적으로 산출해 나가는 인구추계 방법으로 정의된다 (Figure 2.1).

$$P_t = P_{t-1} + B_{t-1} - D_{t-1} + NM_{t-1}, \tag{2.7}$$

여기서 P_t는 t년 인구, B_{t-1}는 $t-1$년 출생아수, D_{t-1}는 $t-1$년 사망자수, NM_{t-1}는 $t-1$년 국제순이동이다.

이런 코호트요인법을 가장 손쉽게 구현하고 계산할 수 있는 방법은 Leslie 행렬을 이용하는 것이다. 이와 관련된 연구는 Leslie (1945, 1948), Smith와 Keyfitz (1977), Preston 등 (2001)을 참고하면 된다.

Leslie 행렬 계산과정을 간략히 소개하면, 인구(P_t)를 간략하게 5개의 연령별 그룹으로 구성(0–14, 15–29, 30–44, 45–59, 60+)하자. 앞으로 15년 생존하는 인구를 계산하기 위해서는 식 (2.8)과 같다.

$$\begin{aligned} P_i(t+15) &= P_{i-1}(t)\frac{L_i}{L_{i-1}}, \quad i = 2,3,4, \\ P_5(t+15) &= P_4(t)\frac{L_5}{L_4} + P_5(t)\frac{T_6}{T_5}, \end{aligned} \tag{2.8}$$

여기서 L_i는 생명표의 정지인구(x세의 생존자수가 $x+1$세에 이르기까지의 총인년수(person-years)이고, $L_i = \int_x^{x+1} l_x dt$, l_x는 생존자수), T_i는 x세에서 $x+1$세의 L_x자신을 포함하는 x세 이후의 정지인구의 합계($T_x = \sum_{y=x}^{w-1} L_y$)이다. 그리고 가임기간동안 총출생아수는 모의 연령별 출생(15–45세까지 가

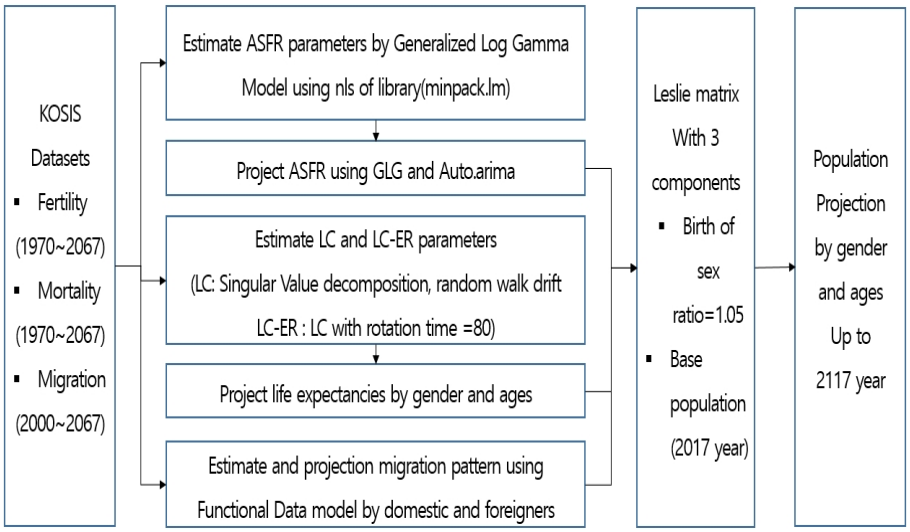


Figure 2.2. Structure of R program for population projection.

정)을 합치면 되므로 식 (2.9)처럼 되고, 첫 15세 연령그룹은 식 (2.10)처럼 유도된다.

$$B[t,t+15] = \frac{15}{2}F_2\left(P_2(t)+P_1(t)\frac{L_2}{L_1}\right)+\frac{15}{2}F_3\left(P_3(t)+P_2(t)\frac{L_3}{L_2}\right), \tag{2.9}$$

$$\begin{aligned} P_1(t+15) &= B[t,t+15]\times\frac{1}{(1+\mathrm{SRB})}\times\frac{L_1}{15}l_0 \\ &= \frac{1}{(1+\mathrm{SRB})}\times\frac{L_1}{2l_0}\left(F_2\frac{L_2}{L_1}P_1(t)+\left(F_2+F_2\frac{L_3}{L_2}\right)P_2(t)+F_3P_3(t)\right). \end{aligned} \tag{2.10}$$

식 (2.10)에서 F_2, F_3는 15–29세, 30–44세의 ASFR, SRB는 남녀 출생성비를 뜻한다. 지금까지의 계산 과정과 식들을 행렬연산으로 표현하면 식 (2.11)처럼 유도된다.

$$\begin{pmatrix} P_1(t+15)\\ P_2(t+15)\\ P_3(t+15)\\ P_4(t+15)\\ P_5(t+15)\end{pmatrix} = \begin{bmatrix} kF_2\frac{L_2}{L_1} & k\left[F_1+F_3\frac{L_3}{L_2}\right] & kF_3 & 0 & 0\\ \frac{L_2}{L_1} & 0 & 0 & 0 & 0\\ 0 & \frac{L_3}{L_2} & 0 & 0 & 0\\ 0 & 0 & \frac{L_4}{L_3} & 0 & 0\\ 0 & 0 & 0 & \frac{L_5}{L_4} & \frac{T_6}{T_5}\end{bmatrix}\begin{bmatrix} P_1(t)\\ P_2(t)\\ P_3(t)\\ P_4(t)\\ P_5(t)\end{bmatrix}. \tag{2.11}$$

시간 t에서 연령 그룹별 인구의 열 벡터를 $P(t)$로 표시하고 시간 t와 $t+15$ 사이의 추계행렬(projection matrix, 이를 Leslie 행렬)을 $L[t,t+15]$로 표기하면, 추계의 일반식은 $P(t+15)=L[t,t+15]P(t)$와 같다. 보다 더 Leslie 행렬에 관심이 있는 독자는 Preston 등 (2001, pp. 118–133)를 참고하면 도움이 된다. 지금까지 인구변동 3요인의 모형과 코호트요인법을 논의하였으며, 이런 구성요소들이 본 연구에서 제안하는 R 프로그램은 어떻게 구성되는지를 Figure 2.2에 정리하여 소개한다.

Figure 2.2는 R 프로그램 프로세스(좌→우)를 보여주고 있으며, 좌측 처음단계는 인구변동 3요인의 KOSIS의 데이터 sets, 중간부분은 출산, 사망, 국제이동, 그리고 Leslie 행렬연산 구조, 우측끝은 인구

Table 3.1. Most important functions of ASFR estimation and prediction

```
# GLG 모형 적합과 모수 초기치 대입을 위한 GLG, GLGmodel 함수
GLG←function(asfr,fer,mu,b,lambda);#ASFR(asfr),출산순위별출산율(fer),평균출산연령(μ),
                                    분산(b),형태모수(λ)
GLGmodel←function(par,xx)          #모수(par), 모수 구분(xx)
# Final : 모형 적합 residual 체크
Final←function(asfr,mu,b,lambda) #ASFR(asfr),평균출산연령(μ),분산(b),형태모수(λ)
# FITGLG: 모수 추정값을 GLG 모형에 대입하여 ASFR 추정치 도출
FITGLG←function(fer,mu,b,lambda) #출산순위별출산율(fer),평균출산연령(μ),분산(b),형태모수(λ)
# Autoari: GLG 모형 모수 4개 예측을 위한 arima
autoari←function(target,a,b,i,k) #모수(target),start(a),end(b),모수구분(i),예측주기(k)
# PREGLG: ASFR 예측치 도출
PREGLG←function(time,fer,mu,b,lambda) # 예측주기(time)
# 주요 함수 활용 사례
coef(GLG(asfr[,i],fer=3,mu=30,b=3,lambda=-0.3))[1:4] #초기치 대입
sim←GLGmodel(glg,x)                                  #nls.lm 패키지 사용을 위한 실측치 저장
residFun←function(p, observed, xx) observed - GLGmodel(p,xx) # residual
coef(Final(asfr[,i],mu_asfr[i],b_asfr[i],lambda[i]))[1:4]      # 연도별 모수 추정값
FITGLG(S2asfr[,1],S2asfr[,2],S2asfr[,3],S2asfr[,4])$glg        # 연도별 연령별 출산율 도출
autoari(S2asfr,1,98,1,50)                                      # 시간별 모수 예측
colSums(PREGLG(50,temp_c,temp_mu,temp_b,temp_lambda)$glg # 시간별 연령별 출산율 예측
```

추계의 결과부분이다. 개별 과정은 사용자 편의를 위해 함수명으로, 중간부분의 인구변동요인들은 모형 적합에 따라 예측값이 도출되고, 이들 예측값들은 Leslie 행렬 연산과정으로 이어져 최종적으로 인구추계를 산출할 수 있도록 구현하였다.

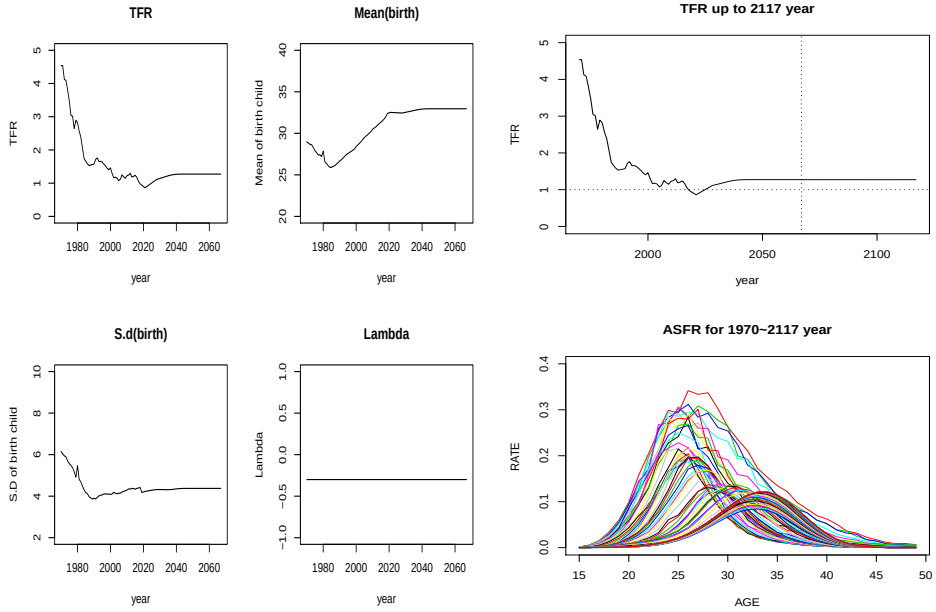
인구변동 3요인의 추정과 예측을 위한 입력값은 KOSIS의 출산, 사망, 국제이동(내국인, 외국인)자료이며, 인구추계를 도출을 위한 입력값은 출산의 GLG 모형, 사망의 LC-ER, 국제이동의 FDM 모형으로 도출된 예측값과 출생성비, 기준인구이다. 기준인구는 주민등록 연앙인구(7.1일자 인구)를 활용하면 된다. 이 모든 자료들은 통계청의 KOSIS에 제공되고 있으며 손쉽게 다운로드 받을 수 있음을 밝힌다.

3. 인구변동 3요인 R 구현과 개요

이번 장에서는 앞장에서 살펴본 인구변동 3요인과 코호트요인법으로 인구추계를 산출할 수 있는 R 프로그램과 일부 R 패키지 활용을 소개한다.

첫째 출산율 R 프로그램은 5단계로 구성된다. 1, 2단계는 출산율 자료 불러오기와 분석하기 위한 list 자료 생성(structure 활용), 3단계는 ASFR을 GLG모형으로 적합하기 위해 GLG, GLGmodel, Final, FITGLG함수를 단계적으로 만들고, 특히 R 패키지의 minpack.lm에서 제공하는 nlsLm함수를 활용하여 추정하고, 4단계는 MAE_glg함수로 적합력(mean absolute error; MAE) 체크와 추정된 모수값들을 토대로 시계열 예측(Auto.arima함수 활용), 마지막으로 추정된 모수값을 활용하여 시간별 연령별 출산율 추정(1970–2067년)하고 PREGLG함수로 만들어 예측(2068–2117년)결과를 출력한다. 이런 일련의 주요 함수를 Table 3.1에 소개하고 Figure 3.1은 1970–2067년 모수 4개의 추정결과와 1970–2117년 ASFR 변화를 보여주고 있다.

둘째 사망률 R 프로그램은 5단계로 구성된다. 1, 2단계는 출산율 1–2단계와 동일하며, 3단계로 사망률 LC 모형과 기대수명은 LC와 LElc함수로, LC-ER의 b_x^{μ}, LC-ER 모형, 그리고 LC-ER의 기대수명은



(a) Four parameters (b) TFR and ASFR up tp 2117

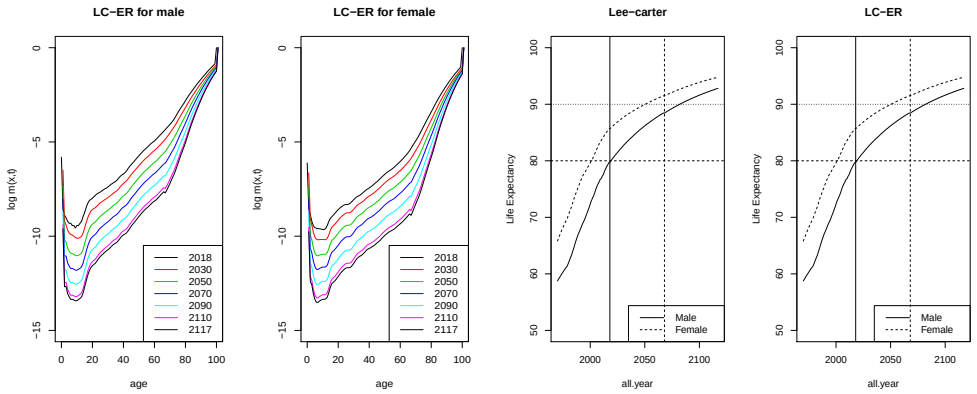
Figure 3.1. The pattern of parameter on fertility model for 1970–2067 years and TFR, ASFR for 1970–2117 years.

bUx, LLG, LELLG함수로 구현하였다. LC함수는 a_x, b_x와 k_t의 모수 추정을 위한 평균(apply(logmx, 1, mean)), singular value decomposition (SVD), b_x와 k_t의 모수제약 스케일링($\sum b_x = 1, \sum k_t = 0$), k_t 예측을 위한 random walk drift부분이 주요 내용으로 구성하였다. LElc함수는 0세의 기대수명, 연령별 기대여명을 산정하기 위해 생존율(survival rate) (Preston 등, 2001, pp. 266–269)을 이용하였다. bUx함수는 식 (2.3)의 b_x^{μ}를 구현하고, LLG함수는 기존의 LC함수의 결과를 활용하여 성별의 교대시점(80세)을 찾아 80세 전과 후로 구분하고, 교대 전은 LC함수와 기대수명, 이후는 LC-ER함수와 기대수명을 따르도록 하였다. 또한 LELLG함수는 LC함수와 유사한 로직으로 구현하였다. 4단계로 적합력(본 프로그램은 mean absolute error (MAE)) 체크와 마지막으로는 모수추정 결과와 ASMR, 기대수명 출력부분이다. ASMR 추정, 예측의 주요 함수를 Table 3.2에 소개하고 Figure 3.2에 2117년까지의 성별의 ASMR 예측과 기대수명을 보여준다.

셋째 국제이동 R 프로그램도 5단계로 구성된다. 1단계는 출산율 1단계와 같고, 2단계는 국제이동의 노출위험(exposure risk, 인구학(demography)에서는 율(rates)은 'occurrence/exposure'의 비율로 정의하는데, 분모의 exposure는 총인년수(person-years))을 산정한다. 즉 인구학에서 율은 'number of Occurrences/Person-years of Exposure to Risk of Occurrence'로 정의한다. 따라서 율은 분자 내에 정의된 기간 내에 발생하는 사건수를, 분모에는 그 기간 동안 개체군에 살았던 사람 년 수의 추정치를 의미하고, 이를 R에서는 '적정상수×apply(국제이동자수, 2, cumprod)'로 구현한다. 3단계는 FDM모형을 적용(R의 demography의 FDM함수 활용, 주성분 개수는 2개), 4단계는 내국인과 외국인을 분리하여 각각의 자료에 FDM모형를 적용하여 결과를 도출한다. 마지막은 이런 일련의 산출과정을 시각적으로 표현하기 위한 것이다. Table 3.3은 국제이동률을 추정, 예측하기 위한 주요함수를 보여주고 있으며, Figures 3.3과 3.4는 2067년까지 내국인과 외국인의 국제이동 동향과 2117년까지의 국제이동자수

Table 3.2. Most important functions of ASMR estimation and prediction

LC 모형, LC 모형 간략 life table
LC←function(mx,trial,targetyear) # mx(사망률), trial(미래 불확실성 대비 시뮬레이션 횟수), targetyear(예측 마지막연도)
LElc←function(mx,lcax,lcbx,lckt,lcmx) # lcax (LC의 a_x), lcbx (LC의 b_x), lckt (LC의 k_t), lcmx (LC의 $log(m_x)$)
LC-ER 모형, LC-ER 모형 간략 life table
bUx←function(lcbx)
LLG←function(mx,lcax,lcbx,lckt,lcmx)
LELLG←function(ma,lcax,lcbx,lckt,lcmx,tempsex,targetyear)
life table : Coale and Demeny (1983) west model, a_0, a_1과 사망확률, 정지인구, 기대수명 등의 생명표
a0←function(m0,sex)
a1←function(m0,sex)
life.table←function(mx,sex,width=1,ax.method="split",radix=100000)
#사망률(mx), 성별(sex), 1세 간격(width=1), 도출의 중앙 분기점(split) 활용, 10만 가상인구
주요 함수 활용 사례
LC(male,1000,2117) #남자 사망률, 1000번 시뮬레이션, 2117년 예측 마지막연도
LElc(male,lcaxm,lcbxm,lcktm,lcmxm) #남자사망률, 추정된 LC의 a_x, b_x, k_t, $log(m_x)$ 대입한 기대수명
LLG(male,lcaxm,lcbxm,lcktm,lcmxm) #LC에서 추정된 값들을 LC-ER 모형에 대입
LELLG(male,lcaxm,lcbxm,lcktm,lcmxm,tempmale,2117) # LC-ER 모형으로 추정된 값들의 기대수명
a0(mx[1],sex) # Coale and Demeny (1983) west model, a_0, a_1
a1(mx[1],sex) # Coale and Demeny (1983) west model, a_0, a_1
life.table(exp(LElc(male,lcaxm,lcbxm,lcktm,lcmxm)$all),"m",width=1,ax.method="split",radix=100000)
#남성(m)생명표 도출, 1세 간격(width=1), a_x도출의 중앙 분기점(split) 활용, 10만 가상인구



(a) Mortality rate by gender (b) Life expectancies by gender

Figure 3.2. Trend of mortality rate for 2018–2117 years and life expectancies by gender for 1970–2117 years.

를 보여준다.

넷째 코호트요인법에 의한 최종 연령별 인구산정이다. 이는 Leslie 행렬로 도출이 가능하며, 본 연구에서는 3단계로 R 프로그램을 구성하였다. 1단계는 출산율, 사망률, 국제이동자수 불러오는 부분이며, 2단계는 코호트요인법을 구현하기 위한 Leslie 행렬연산부분을 출산 경험에 따른 여자와 남자를 구분해

Table 3.3. Most important functions of international migration estimation and prediction

```
# demogdata, demography 패키지를 사용하기 위한 국제이동률 자료 대입
demogdata(data,pop,ages,years,type,label,name,lambda)
# 다변량 시계열 비모수 함수추정 FDM
fdm(data,series,order,ages,max.age,method,lambda,mean,level,transform)
# 시계열 예측 forecast
forecast(object,h=forecast horizon,level=C.I,jumpchoice,method)
# 주요 함수 활용 사례
demogdata(data=d0018mi_m,pop=popm,ages=c(0:max(age)),years=c(2000:2018),type="migration",
          label="migration",name="male")
fdm(dmi_mdata,order=2)
forecast(dmi_m.fit,h=49,jumpchoice="actual")
```

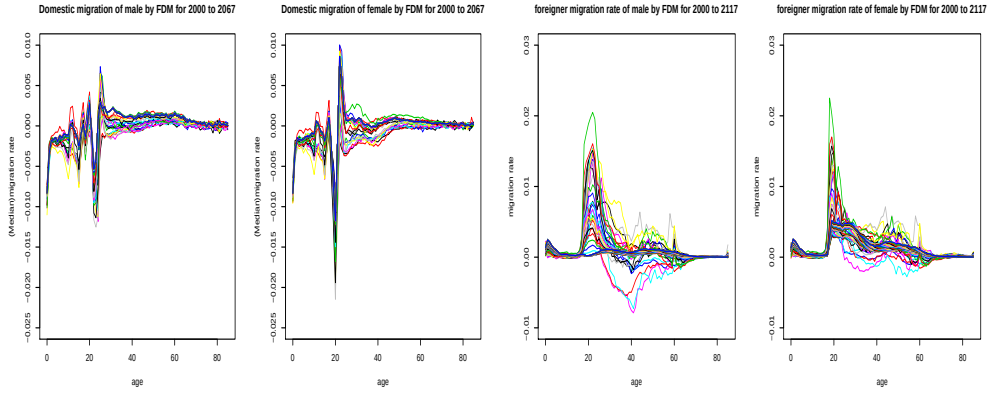


Figure 3.3. Trend of domestic and foreigners migration rate by gender for 2000–2067 years.

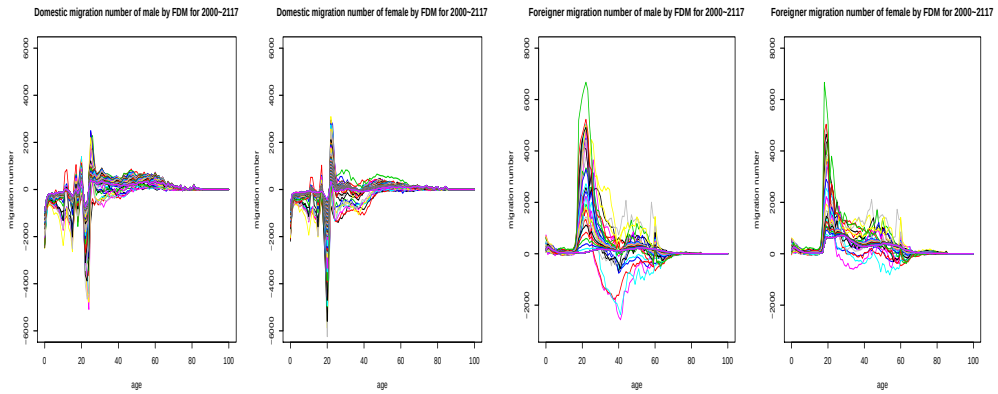


Figure 3.4. Domestic and foreigners migration number by gender for 2000–2117 years.

산출한다. 이 부분은 R에서 proj함수명으로 구성한다. proj에서 요구되는 값은 Nf, Lf, f, If, Nm, Lm, Im, SRB이다. 여기서 Nf는 여성인구수, Lf는 여성 생존자수, f는 출산율, If는 여성 국제이동자수이며 남성은 m으로 명시하였다. 그리고 SRB는 남녀 출생성비로 1.05를 가정한다. 3단계는 Leslie 연산결과

Table 3.4. Most important functions of Leslie matrix for cohort component methods

```
# demogdata, demography 패키지를 사용하기 위한 국제이동률 자료 대입
proj←function(Nf,Lf,f,If,Nm,Lm,Im,SRB) # 코호트 요인법을 Leslie 행렬로 구현하여 인구를 도출하는 함수
# proj 함수 활용 사례
proj(as.vector(datf1[,2]),as.vector(tempLx_fL[,1]),as.vector(tempasfrL[,1]),as.vector(tempmi_fL[,1]),
as.vector(datm1[,2]),as.vector(tempLx_mL[,1]),as.vector(tempmi_mL[,1]),1.05)$p1 # 여성의 proj
```

Table 3.5. Most important functions of R program for population projection

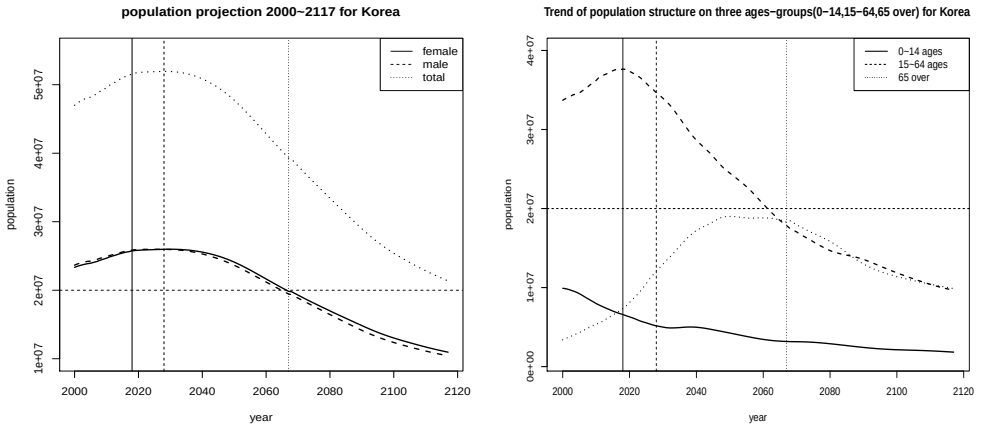
		Function	Goal
Fertility	Estimation	GLG	nlsLm 패키기를 사용하여 일반화 로그 감마모형 적합
		GLGmodel	초기치 대입 이후에 적합 결과
		Final	Estimation and Checking Residual
		FITGLG	모수 추정값을 GLG 모형에 대입하여 ASFR 도출
	Validation	MAE_glg	Mean Absolute Error
	Prediction	autoari	projection for parameters
		PREGLG	Model building by prediction value
Mortality	Estimation, Prediction, & Validation	LC	Lee-Carter Model estimation and prediction
		bUx	b_x^{μ} of LC-ER model
		LLG	LC-ER model
	Life Tables	LElc	Life expectancy for Lee-Carter model
		LELLG	Life expectancy for LC-ER model
		a0	Below age 5 by Coale and Demeny (1983)
		a1	${}_4a_1$ calculation
		life.table	생명표$(q_x, a_x, l_x, d_x, L_x, T_x, e_x)$도출
International Migration	Creation of demogdata	demogdata	demography패키지에서 제공되는 demogdata 객체 생성
	Estimation, Validation	fdm	Functional Data model
	Prediction	forecast	국제이동 연령별 예측
	Combine two demogdata	combine.demogdata	demogdata 실측치와 예측치를 합치는 함수
Cohort Component Method		proj	Leslie 행렬을 이용한 코호트요인법 구현 후 인구추계

와 출력부분이다. Table 3.4는 proj함수와 활용사례를 보여준다.

지금까지 설명한 인구변동 3요인과 코호트요인법의 R 주요함수와 개요는 Table 3.5와 같이 정리된다. 끝으로 2000–2117년까지 성별에 따른 장래인구추이(Figure 3.5)와 결과를 Table 3.6에 정리하였다.

Figure 3.5의 좌측 그림의 실선, 점선, 가는 점선은 2017년, 2028년, 2067년을 표시한 것이며, 2017년은 통계청의 장래인구특별추계의 시작점이면서 기준인구를, 2028년은 남녀 출생 성비를 1.05로 가정했을 때 남녀 비중이 역전되는 연도인 동시에 인구 정점에서 하락추세로 전환되는 지점이며, 2067년은 본 연구에서 2117년까지 예측을 시작하는 시작점을 의미한다. 우측 그림은 일반적인 인구구조 변화를 살펴볼 수 있는 3개 연령별 층인 유소년층(0–14세), 생산가능인구(15–64세), 고연령층(65세 이상)으로 구분하여 2117년까지 추이를 나타낸 것이다. 특히 2018년 이후 생산가능인구가 감소 추이로 전환되고, 고령층은 2060년까지는 증가추세를 보이다가 2060년 중반이후부터 감소하는 경향을 보인다.

인구추계결과 살펴보면 2067년에는 약 3,930만, 2117년에는 2,120만 정도 수준을 나타낼 것으로 예측된다. 2028년 정점인구 5,200만에서 90년 뒤 2117년의 약 2,120만의 수치는 최대치 인구의 40%정



(a) Population (b) Trend of three age-groups

Figure 3.5. Population projection for 2000–2117 years.

Table 3.6. Population projection for 2067–2117 by gender and ages

Age	year											
	2067		2077		2087		2097		2107		2117	
	m	fe	m	fe	m	fe	m	fe	m	fe	m	fe
Total	19454	19840	17144	17453	14807	15298	12811	13335	11435	11899	10374	10757
0∼4	551	522	487	466	403	386	359	345	338	325	292	282
5∼9	538	510	522	500	445	427	369	355	350	338	315	305
10∼14	544	516	543	520	484	464	400	384	355	343	335	324
15∼19	623	592	531	509	519	498	441	424	366	353	347	335
20∼24	698	666	539	521	541	523	481	468	397	389	353	348
25∼29	753	739	621	620	532	533	520	522	442	448	367	375
30∼34	843	834	710	713	553	561	554	561	495	503	411	422
35∼39	905	897	770	774	643	650	554	563	542	550	465	475
40∼44	825	825	853	854	727	732	571	579	572	579	513	520
45∼49	802	808	917	918	785	790	658	665	570	577	558	563
50∼54	1110	1107	836	845	870	875	745	752	589	598	591	596
55∼59	1209	1197	836	850	940	942	806	813	681	688	593	599
60∼64	1216	1193	1120	1115	855	866	891	895	768	772	613	619
65∼69	1634	1557	1232	1216	849	863	951	955	823	826	699	702
70∼74	1916	1747	1225	1207	1113	1117	858	871	896	902	777	780
75∼79	1683	1595	1554	1513	1186	1200	828	857	935	952	815	827
80∼84	1468	1548	1688	1637	1093	1142	1016	107	800	845	847	881
85∼89	1232	1501	1220	1310	1140	1260	18	1034	662	755	776	857
90∼94	641	957	657	873	824	1001	75	737	569	721	482	601
95∼99	228	437	263	450	289	430	287	434	261	394	201	302

주: 단위(천명), m은 male, fe는 female을 의미.

도 수준이고, 이는 저출산과 고령화가 맞물려 인구구조에 영향을 주기 때문으로 판단된다. 남녀추이를 살펴보면 거의 유사한 패턴을 보이고 2025–2030년쯤에 여성이 남자보다 약간 높은 인구 수준을

보이면서 그 추이를 유지하는 것을 알 수 있다. 2067년에 남자는 1,945만, 여자는 1,984만 수준이며, 50년 이후 2117년에는 남자 1,037만, 여자 1,076만으로 예측된다. 이들 수치는 2025년 남자의 정점인구 2,597만, 2030년 여자의 정점인구 2,598만에 비해 2067년은 74.9%(남자), 76.4%(여자)로 75%수준으로 나타나지만, 인구 감소추이가 더해져 2117년은 39.9%(남자), 41.4%(여자)로 40%수준으로 50년만에 35%p가 감소되어 희망적이지 않는 수치가 도출된다. 전반적인 2000–2117년 인구추계는 Figure 3.5에서 확인할 수 있으며, 2028년 이후의 감소 추이가 가파르고 이런 추세가 지속될 경우에는 해외에서 발표된 '인구구조 붕괴, 어려운 인구문제에 직면 등'이라는 연구 결과(GEFIRA, 2018; BBC, 2019)와 일맥상통한 부분이 많다.

4. 결론 및 제언

본 논문은 장래인구추계 방법론과 코호트요인법을 R로 구현하여 실현가능성이 가장 높은 중위수준 관점에서 2117년까지 장래인구를 도출한다. 기존 인구추계 패키지인 demography, bayesPop을 소개하면서, 이들 프로그램으로는 우리나라 인구변동 3요인을 추정하거나 예측이 수월치 않다는 점을 소개했다. 이런 현실을 감안할 때, 본 연구는 통계청 KOSIS의 ASFR, ASMR, 국제이동률 자료를 로드한 후, R에서 제공하는 일부 패키지를 활용한 출산율, 국제이동률의 추정과 예측, R 프로그램으로 Leslie 행렬 구현과 사망률 추정과 예측 결과를 소개한다는 점에서 의의가 있다고 본다. 특히 관심독자를 위해 구글클래스룸(수업코드: s3zf7cu)에 지금까지 소개한 R 프로그램을 상세히 소개하고, 라인별로 각주를 달아 프로그램 설명을 추가하였다. 이는 향후 인구 관련 독자들의 관심도 제고와 후속 연구에 도움을 줄 수 있을 것으로 판단된다.

하지만 본 연구를 수행하면서 몇 가지 한계점과 향후 연구과제는 여전히 남아있다. 첫째 최근 지방소멸, 스마트시티 등 차별화된 인구구조 변화 패턴 등으로 지자체별로 자체 인구추계 수요(지방자치 연구용역 형태, 고양시, 2020; 진주시, 2020; 고령군 2020; 논산시 2020)가 늘어나고 있는 실정이다. 이런 요구를 충족하기 위해서는 본 프로그램을 시도별로 확장하는 작업이 시급하다. 이를 위해 사망률 모형은 LL 모형과 LC-ER 모형의 혼합모형(LL&LC-ER)으로 변환이 요구되고, 국제이동은 17개 지방자치별, 시군구, 읍면동에 맞는 국내인구이동 예측모형으로 전환이 필요하다. 둘째 전체 인구에 대한 인구변동 3요인 통계 자료는 1세, 1년의 형태이지만 시군구, 읍면동의 경우 1세, 1년의 인구동태통계가 구축되어 있지 않고, 대체적으로 5세 또는 10세, 5년 또는 10년 간격으로 수집되는 경우와 일정치 않은 주기가 대부분이다. 특히 시군구, 읍면동의 인구변동 3요인에 대한 자료 품질 뿐만 아니라 완비성도 미비한 수준이다. 이를 보정하기 위해 Beers (1944) 보정계수, spline 방법과, 소지역 추정인 경우 Bayesian 방법도 고려해야 한다. 셋째 보다 완벽하고 사용자 편의를 위한 프로그램이 되기 위해서는 R 프로그램 전부를 함수모듈로 구성하여 제공하는 것이 중요하다. demography나 bayesPop 등을 살펴보면 출산, 사망을 계산하기 위해서 하나의 함수모듈로 제공하고 있다. 그리고 함수명을 클릭하거나 검색하면 보다 상세한 프로그램을 제공하고 있다. 이에 프로그램을 더욱 향상시켜 함수모듈로 구성된 패키지(package)형태로 구축하는 것을 향후연구과제로 남기고자 한다.

지금까지 장래인구추계의 방법론, R 프로그램 구현, 사용법, 결과와 연구의 한계점을 다양하게 제시해 보았다. 우리나라 미래 인구구조를 조명할 때, 초저출산과 급격한 고령화는 자주 언급되는 용어이고 동시에 2000년도 이후 국제이동의 순이동(immigration − emigration)이 양인 입국의 나라로 유지되고 있고 행정수도, 신도시 건설 등으로 국내이동 또한 빈번하다. 이런 여러 현실과 상황을 고려할 때 현재 통계청에서 적용하고 있는 전체에서 부분인구(Top-Down, 전체인구→시도별, 시군구, 읍면동)로 산출하는 방식 뿐만 아니라 부분에서 전체(Bottom-up, 전체인구←시도별, 시군구, 읍면동)를 동시에 도출하는 방식을 도입해야 한다. 더불어 이런 쌍방 방식을 손쉽게 구현할 수 있는 프로그램 구축이 더욱 필

요하다고 판단된다.

References

BBC (2019). South Korea's population paradox, https://www.bbc.com/worklife/article/20191010-south-koreas-population-paradox.

Beers, H. S. (1944). Six-Term Formulas for Routine Actuarial Interpolation, *American Institute of Actuaries*, **33**, 245–260.

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

GEFIRA (2018), The collapse of the South Korean population: the countdown has begun, https://gefira.org/en/2018/01/11/childless-south-korea/.

Horiuchi, S. and Wilmoth, J. R. (1995). Aging of Mortality Decline, Rockefeller University, New York.

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, T., Tickle, L., and Maindonald, J. (2019) demography, Forecasting Mortality, Fertility, Migration and Population Data, https://cran.r-project.org/web/packages/demography/index.html.

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

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.

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.

KOSTAT (2010). Population projection 2010∼2060.

KOSTAT (2016). Population projection 2015∼2065.

KOSTAT (2019). The Special population projection 2017∼2067.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association*, **87**, 659–671.

Lee, R. D. (1998). Probabilistic approaches to population forecasting, *Population and Development Review*, **24**, 156–190.

Leslie, P. H. (1945). The use of matrices in certain population mathematics, *Biometrika*, **33**, 183–212.

Leslie, P. H. (1948). Some further notes on the use of matrices in population mathematics, *Biometrika*, **35**, 213–245.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method, *Demography*, **42**, 575–594.

Li, N. and Gerland, P. (2011). Modifying the Lee-Carter Method to Project Mortality Changes up to 2100, the Population Association of America 2011 Annual meeting-Washington, DC, Session 125, formal Demography I: Mathematical Models and Methods.

Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee-Cater method to model the rotation of age patterns of mortality decline for long-term projections, *Demography*, **50**, 2037–2051.

Oh, J. H. (2018). A comparison between the real and synthetic cohort of mortality for Korea, *The Korean Journal of Applied Statistics*, **31**, 427–446.

Oh, J. H. (2019). Forecast and identifying factors on a double dip fertility rate for Korea, *The Korean Journal of Applied Statistics*, **32**, 463–483.

Oh, J. H. (2020). Stochastic population projections on an uncertainty for the future Korea, *The Korean Journal of Applied Statistics*, **33**, 185–201.

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.

Preston, S. H., Heuveline, P., and Guillot, M. (2001). Demography, Measuring and Modeling Population

Processes, Blackwell.

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

Ševčíková, H., and Raftery, A. E. (2016). bayesPop: Probabilistic Population Projections, *Journal of Statistical Software*, **75**, 1–25.

Ševčíková, H., Raftery, A., and Buettner, T. (2019). bayesPop: Probabilistic Population Projection, https://cran.r-project.org/web/packages/bayesPop/index.html.

Smith, D. P. and Keyfitz, N. (1977). *Mathematical Demography: Selected Papers*, Springer Verlag, Berlin.

STI (Statistical Training Institute) (2013). Basic and Application of demography.

R를 활용한 인구변동요인 산정과 인구추계 시스템 개발

오진호[a,1]

[a]한밭대학교 공과대학 수리과학과

(2020년 5월 20일 접수, 2020년 6월 30일 수정, 2020년 7월 11일 채택)

요 약

본 논문은 최근에 널리 사용되고 있는 R 프로그램으로 출산율, 사망률, 국제이동률을 예측하고 이들 결과를 Leslie 행렬에 대입해 인구추계 산출하는 방법을 소개한다. 특히 Kaneko (2003)가 제안한 출산율의 일반화로그감마모형, Li 등 (2013)의 사망률 LC-ER 모형, Ramsay와 Silverman (2005)가 제안한 국제이동률의 함수적데이터모형을 시현할 수 있도록 하였다. 최근 R로 구현된 대표적인 인구추계 패키지로 demography, bayesPop가 소개되고 있으나, 이는 Human Mortality Database (HMD), Human Fertility Database (HFD)에 업로드된 자료에 한에서만 분석이 가능하고 기타 데이터를 적용하기 위해서는 자료 변경과 수정이 요구된다. 특히 우리나라의 경우 HMD에 단기간의 자료로만 제공되어 있어 이 패키기를 적용하기에는 한계점이 있다. 이에 본 논문은 이런 실정과 한국의 저출산, 고령화, 내국인, 외국인 국제이동률 상이패턴을 반영할 수 있는 R 프로그램을 소개하고, 2117년까지의 인구추계를 도출하였다.

주요용0어: 출산율, 사망률, 국제이동률, Leslie 행렬, 인구추계

[1](34158) 대전광역시 유성구 동서대로 125, 한밭대학교 공과대학 수리과학과. E-mail: jhoh75@hanbat.ac.kr