Fertility Prediction using Micro Data[*]

Seik Kim
Korea University
seikkim@korea.ac.kr
ORCID: 0000-0002-7605-3817

Yoosik Shin
Hanbat National University
yoosikshin@hanbat.ac.kr
ORCID: 0000-0001-8585-9946

Abstract

This paper presents a novel approach to forecasting fertility patterns using micro data, focusing on individual-level observations to enhance predictions of fertility rates. Our method leverages the advantages of completed fertility rates, which provide more reliable fertility measures when birth rates differ by cohorts. Analyzing detailed birth history outcomes from Census Korea 2% data, we estimate and predict fertility outcomes, encompassing both the timing and total number of childbirths. We find that younger cohorts tend to delay childbirth compared to earlier cohorts, with educational attainment and childbirth history playing a significant role in this trend.

JEL Codes: J13

Keywords: Probabilistic Population Forecasting, Completed Fertility Rate, Total Fertility Rate

[*] The authors are grateful to Jisoo Hwang, Jinyoung Kim, Sangsoo Park, two anonymous reviewers, and seminar participants at Korea Labor Institute, Seoul National University, the annual meetings of the Korean Economic Association and the Seoul Labor Economics Working Group. This study was supported by the Korea University Grant (K2209371). The authors declare that they have no relevant financial interests related to the research described in this paper. Yoosik Shin is the corresponding author.

1. Introduction

The current size and demographic composition of a population play a pivotal role in shaping a nation's future productivity and societal outcomes. Accurate forecasts regarding future population trends, in turn, guide present-day policy formulation, empowering policymakers to make more informed decisions. The structure of the future population hold profound implications across various policy directions, spanning demography (pertaining to age distribution and the working-age population), labor economics (encompassing labor supply and demand dynamics and intergenerational inequality), macroeconomics (impacting economic growth trajectories and government expenditure), and public finance (influencing the allocation of resources towards education, healthcare, and social security systems). Given that population changes are shaped by the interplay of fertility, mortality, and migration patterns, accurate population forecasting necessitates predictions encompassing all three components.

Of the three determinants of population dynamics, this paper presents a strategy that specifically focuses on probabilistic fertility prediction. Fertility holds particular significance for policymakers due to its strong association with future population growth and composition. Governments can directly influence fertility through a range of policies, including childbirth subsidies, maternity and parental leave, and childcare support, among others. In contrast, immigration policies, while important, aim to regulate the influx of immigrants who complement the domestic working-age population and are influenced by fertility projections. Regarding mortality, policies focus on elderly welfare, such as healthcare and social support, rather than directly targeting mortality rates as a primary policy goal. Consequently, mortality becomes an incidental outcome of these policies rather than the primary target.

A widely used indicator for predicting future fertility is the total fertility rate (TFR), but it may be an inadequate predictor if women born in different years exhibit different fertility behaviors. The TFR for a given year is defined by the average number of children a 15-year-old woman would have until she reaches 49, assuming that the age-specific birth rates in that year remain constant throughout her childbearing years.[1] Hence, for the TFR to accurately predict future

[1] The age-specific birth rate is the average number of children born within a year by that age group. The TFR is the sum of all the age-specific birth rates.

fertility, it is essential that the age-specific birth rates in the base year remain unchanged. For instance, in Korea, the annual number of children born has decreased over the decades as younger cohorts of women have fewer births over their lifetimes than older cohorts. When younger cohorts have fewer births over the course of the reproductive years, the TFR will systematically over-predict the future fertility rate because it uses the higher birth rates of older cohorts to predict the future birth patterns of younger cohorts.

The objective of this research is to develop a more reliable predictor for fertility rates, one that takes into account the variations in birth rates across different cohorts. To achieve this goal, it is crucial to have a comprehensive understanding of the trends in birth behavior across cohorts. In this context, the completed fertility rate (CFR) emerges as a useful concept. The CFR represents the average number of children born to women within a specific cohort over the course of their reproductive years. It is realized when a specific cohort reaches the final year of childbearing, typically set to 49. By its nature, the CFR utilizes retrospective records, providing researchers with a means to assess childbearing behavior from previous years. The information obtained from these estimations can then be utilized to make predictions regarding future fertility rates.

This paper develops a method to estimate and predict fertility patterns using micro data. The proposed method makes notable contributions to the existing literature in several key aspects. First, our method enables one to examine how individual-level factors, such as previous birth history or education level, affect the timing and total number of childbirths over a lifetime. By accounting for variations in these factors across cohorts, our approach provides insights into the evolution of fertility outcomes across generations. Second, unlike previous studies relying on aggregate data, our method is based on individual-level data. Its large sample size leads to more accurate projections with narrower confidence intervals. Finally, we predict the birth schedule, capturing not only the total number of completed births but also the timing of each birth. Utilizing a duration model, we can discern how trends in both the ages at childbirth and the number of children vary across cohorts.

We apply our method to the Census Korea 2% data, which offers a large and representative sample of women from different cohorts and their socio-demographic characteristics, allowing us to construct detailed birth history outcomes. Utilizing this data, we find that younger cohorts tend to delay childbirth compared to earlier cohorts, with educational attainment playing a significant role in this trend. Moreover, individuals with a higher number of previous children tend to have longer durations between childbirths, indicating a reduced likelihood of having additional children. The fertility projections for the 1985, 1990, and 1995 cohorts provide insights into future fertility trends. The 1985 cohort is anticipated to delay childbirth more than the 1980 cohort, even though they are expected to have a similar overall number of children. However, the 1990 and 1995 cohorts are projected to have fewer children overall.

The paper proceeds as follows. Section 2 reviews the methods of probabilistic population forecasting and summarizes the contributions of our approach. Section 3 introduces the Census Korea data, presents descriptive statistics, and explains how birth history can be constructed from cross-section data. In Section 4, we generalize the concepts of the TFR and the CFR, ensuring that the fertility measures are comparable and suitable for estimation and prediction purposes. Section 5 discusses the empirical specification, estimates the fertility hazards, and presents the prediction results. Finally, Section 6 offers concluding remarks summarizing the key findings and implications of the study.

2. Methods of Fertility Prediction

2.1. Probabilistic Population Forecasting Methods

There is a large literature dedicated to forecasting fertility (e.g., Alkema et al., 2011; Schmertmann et al., 2014), mortality (e.g., Girosi and King, 2008; Lynch and Brown, 2010), and migration (e.g., Gorbey, James, and Poot, 1999; Bijak, 2010).[2] Some focus on one of the three while others undertake all three aspects within a single study, often employing a shared methodology. For example, the strategy developed by Lee and Carter (1992) to predict mortality is also applied to forecast fertility in Lee (1993). In these papers, mortality and fertility time series data are utilized to estimate and anticipate the respective processes. Likewise, more recent studies such as those conducted by Hyndman and Booth (2006), Raftery,

[2] We exclusively focus on probabilistic forecasting and refrain from utilizing expert-based approaches. Some research combines expert information with stochastic forecast models, as demonstrated by Billari, Graziani, and Melilli (2014), for instance, but our review and discussion center solely on data-driven models.

Alkema, and Gerland (2014) and Wisniowski et al. (2015), among others, also rely on nation-level time series data to project future population trends.[3]

The seminal paper by Lee (1993) on fertility forecasting employs the method used for mortality prediction developed by Lee and Carter (1992). The main goal of his work is to propose a time series model that forecasts long term age-specific fertility rates. After specifying a one-parameter model which accounts for age-time variations in fertility, he estimates the model by using the singular value decomposition approach and predicts the fertility index. The study also imposes upper and lower bounds to avoid negative fertility rates and restricts long-run fertility rates to equilibrate at a specific level using prior information. The model is fitted to fertility data from the United States, and the overall performance of the method is evaluated.

Instead of adopting identification via the singular value decomposition approach, Hyndman and Booth (2006) propose a methodology for stochastic population forecasts using functional principal component analysis. The method extracts the principal components from historical data on fertility, mortality, and migration. These principal components capture the underlying patterns and variability in the data over time. Then, they utilize these components to construct functional time series models, such as functional autoregressive models, to forecast future demographic trends with probabilistic prediction intervals. Additionally, they discuss the application of their methodology to making 20-year forecasts using Australian data for the period 1921–2004 and evaluate the performance of their models using empirical data.

The empirical specification of Lee (1993) is extended by Wisniowski et al. (2015), who employ a Bayesian approach. They first construct a general framework for the projection of mortality, fertility, emigration and immigration by adapting the Lee-Carter model. To forecasts the four population components, they exploit a Bayesian approach which assumes prior distributions for the model parameters. Then, the population components integrate within a cohort component projection model. To obtain posterior distributions, they exploit Markov chain Monte Carlo approach. By applying their method to data from the United Kingdom, they report

[3] In discussions on approaches to population forecasting, Booth (2006) provides a comprehensive overview of studies conducted between 1980 and 2005.

the advantages of the Bayesian approach for population projection.

Methods have been developed to produce country-specific probabilistic population projections for multiple countries. Concerned about the lack of uncertainty evaluation in previous United Nations population projections, Raftery, Alkema, and Gerland (2014) establish Bayesian hierarchical models to forecast future fertility and mortality rates for each country. These rates are measured by the TFR and life expectancy for females and males. The projections yield numerous possible scenarios from the posterior predictive distribution, which are then integrated into a cohort component projection model. They assess their methods using the TFRs and mortality rates of 159 countries.

As fertility, mortality, and migration are determined by their specific processes, previous research has also developed methods or focused on fertility prediction. For instance, Alkema et al. (2011) develop a methodology tailored for probabilistic population projections, specifically focusing on fertility rates. Their approach hinges on a Bayesian hierarchical model designed to generate country-specific TFR projections for all nations. Drawing from the demographic theory of fertility transitions, which delineates pre-transition high fertility, the fertility transition, and post-transition low fertility phases, they model the TFR as a sum of two logistic functions that depend on the current TFR and a stochastic term. They use the United Nations' estimates of TFR spanning 1950 to 2010 for 196 countries to produce probabilistic forecasts of forthcoming fertility trends.

Earlier literature, such as Bloom (1982) and de Beer (1985), has also utilized the CFR, recognizing its advantages over the TFR. These studies note the CFR's robustness, attributed to its freedom from tempo distortion. Models employing the CFR to forecast fertility trends have since been developed. Schmertmann, Zagheni, Goldstein, and Myrskylä (2014) utilize the Human Fertility Database (HFD), combining it with data from other sources to create a final dataset covering 37 countries. The data is aggregated at the cohort level by age for each country. Their proposed method is a Bayesian model that incorporates prior information about patterns over age and time. Cohort schedules are approximated using principal components of HFD schedules in the age dimension, while ensuring smoothness and linearity over short spans in the time dimension. Forecasts of completed cohort fertility for women born in the 1970s and 1980s are provided.

2.2. Our Contribution

Utilizing aggregate data facilitates projections spanning decades or even centuries into the future, albeit with relatively large confidence intervals. While statistical models based on grouped data represent valuable tools for population forecasting, they may not comprehensively account for determinants influencing fertility, mortality, and migration due to the lack of micro data capturing individual behavioral decisions. Factors at the personal level, such as birth history and socio-economic status, can influence fertility decisions but may be overlooked in models relying solely on grouped data. Consequently, complementing purely statistical models with an understanding of individual-level behavioral decisions could enhance the accuracy of population forecasting.

This paper aims to estimate the dynamics of age-specific fertility over cohorts, leveraging the availability of individual-level birth records. Furthermore, our methodology involves developing a future population estimate that not only controls for socio-demographic changes across cohorts but also incorporates individual-level factors such as educational attainment, previous birth, and timing of birth. As highlighted by Lee (1993), disaggregating fertility forecasts by age is crucial for generating accurate predictions of birth numbers in conjunction with population age distributions. However, generating independent age-specific forecasts can be time-consuming and may overlook their strong statistical interdependence.[4] While we acknowledge Lee's rationale, particularly in contexts where micro data are unavailable, our approach capitalizes on individual-level observations to produce age-cohort-specific forecasts tailored to both age and cohort characteristics rather than relying solely on age-specific estimates.

[4] According to Lee (1993), "It is important that fertility forecasts be disaggregated by age, so that they can be used in conjunction with population age distributions to generate forecasts of numbers of births." He continues, "Yet we do not want to generate independent age-specific forecasts, which would be time-consuming and would overlook their strong statistical interdependence."

The comparison between previous methods and our proposed strategy is outlined in Table 1, highlighting several distinctive features. First, our method leverages the advantages of the CFR. As discussed earlier, the CFR is a more reliable measure of fertility when birth rates differ by cohorts. Methods targeting the TFR predict the conventional TFR, which represents the total number of childbirths over a lifetime. Methods targeting the CFR, including ours, predict the fertility schedule, which traces the number of children by age. Predicting the fertility schedule allows us to understand both the number and timing of births. Specifically, our method forecasts the timing of future childbirths for each individual. These individual forecasts are then aggregated to generate cohort-specific predictions.

Second, previous methods use time series models, whereas our approach utilizes a duration model. Previous methods targeting the TFR require time series data of TFR at the national or regional level to predict future TFR. Previous methods targeting the CFR require time series data of birth rates for every age at the national or regional level to predict future fertility schedule. Our approach necessitates the collection of childbirth data for each age of individual mothers. The model specification relies on behavioral outcomes, accounting for variations in fertility not only by age and cohort but also by childbirth history and individual attributes. These factors serve as explanatory variables in our model and are used to predict the future fertility schedule. However, it is worth noting that methods targeting the CFR, including ours, are not suitable for long-term predictions. This is because age-specific birth data or individual birth history data are relatively limited compared to the much longer time series data available for TFR.

Table 1. Differences between the Existing Methods and the Proposed Method

	Previous TFR Methods	Previous CFR Methods	Our CFR Method
Target Variable	Conventional TFR (Number of Childbirth by Age 49)	Fertility Schedule (Number of Birth by Every Age)	Fertility Schedule (Number of Birth by Every Age)
Model Specification	Time Series Model	Time Series Model	Duration Model
Data Requirement	Time Series of Nation/Region-Level TFR	Time Series of Age-Specific Nation/Region-Level Birth Rates	Childbirth History of Individuals from Different Cohorts
Prediction Horizon	Long-Term Prediction (of Several Decades or even Centuries)	Short-Term Prediction (for Cohorts that are 15 years old or older)	Short-Term Prediction (for Cohorts that are 15 years old or older)

3. Data and Descriptive Statistics

3.1. The Census Korea Data

Our main sample is drawn from the 1990, 1995, 2000, 2005, 2010, 2015, 2020 Census Korea 2% data.[5] We collect data on women aged 15 to 45 from each of the seven waves. We limit our sample to women whose relationship with the household head falls into one of the following categories: herself, spouse, child, or spouse of a child. This is because the parent-child relationship is not clearly identified for women outside of these categories. By using the responses regarding the relationship with the household head, we can identify mothers and their children. From the age of each child, we can determine the mother's age at the birth of each child. For example, if a 40-year old mother lives with 15-year old and 10-year old children, she is regarded to have children at age 25 and 30. This step reduces the number of individuals from 1,459,615 to 1,387,480.

While conventional fertility measures typically encompass women aged 15 to 49, our sample is restricted to those up to age 45 to enhance the precision of birth history information. Utilizing Census data on household relationships allows us to construct a woman's birth history based on the ages of co-resident children. Since older cohorts began childbirth earlier, typically in their early 20s, older women in their 40s are more likely to have children who have moved out, potentially leading to underestimations of completed fertility and distortions in birth history records. As we present in Table 3, childbirth among women in their 40s is relatively rare, diminishing the utility of predicting fertility within this age group.

3.2. Descriptive Statistics

Table 2 reports descriptive statistics. The first column of Table 2 presents descriptive statistics. On average, the women in this sample are 30.75 years old and have completed 12.92 years of education. In this sample, 51% of the women are spouse of the household head and 33% of the women are child of the household head, respectively. The proportion of single women is 39%

[5] We also use the 1966, 1970, 1975, …, 1985 Census Korea 2% data to construct historical fertility rates.

and that of married women is 58%.

In more recent waves the average ages are larger, reflecting a decreasing trend in the cohort size. Younger women have higher education level. The average years of education of women increased from 11.08 to 14.43. The proportion of household head is also increased. Only 7% of women were household head in 1990, but 27% of women are household head in 2020. The women in more recent wave are more likely to be single. The proportion of single women rose from 34% to 53%.

Table 2. Descriptive Statistics

	Wave							
Variable	Total	1990	1995	2000	2005	2010	2015	2020
Age	30.75	28.96	29.89	30.84	31.50	31.53	31.43	31.37
Years of Education	12.92	11.08	11.84	12.46	13.51	13.63	14.08	14.43
Relation w/ Head								
Herself	0.14	0.07	0.08	0.10	0.13	0.17	0.18	0.27
Spouse	0.51	0.58	0.60	0.58	0.53	0.46	0.42	0.34
Child	0.33	0.30	0.29	0.30	0.32	0.35	0.38	0.38
Spouse of Child	0.02	0.04	0.03	0.02	0.02	0.02	0.01	0.01
Marital Status								
Single	0.39	0.34	0.33	0.34	0.37	0.42	0.47	0.53
Married	0.58	0.64	0.65	0.63	0.59	0.54	0.50	0.44
Widowed	0.01	0.02	0.01	0.01	0.01	0.01	0.00	0.00
Divorced	0.02	0.01	0.01	0.02	0.03	0.03	0.03	0.03
Observations	1,387,480	202,607	221,707	222,411	202,060	194,365	179,488	164,842

4. Generalization of TFR and CFR

This section generalizes the concepts of the TFR and the CFR. First, the conventional TFR is obtained by the sum of the age-specific birth rates from age 15 to 49 in a given year. We define the generalized version of TFR, $TFR_b(x)$, as the sum of the age-specific birth rates from age 15 to x among women born in year b, where x is a natural number between 15 and 49. For example, suppose that we are interested in the TFR of year 2015. The TFR in 2015 is often considered as the expected number of children a 15-year-old woman would have throughout her lifetime. In 2015, women born in 2000 become age 15. Therefore, the conventional TFR in year 2015 is $TFR_{2000}(49)$.

Second, the conventional CFR is the sum of the age-specific birth rates from age 15 to 49 of a given cohort. To calculate the CFR, the birth history of 49-year-old women is used. We define the generalized CFR, $CFR_b(x)$, as the sum of the age-specific birth rates from age 15 to x for women born in year b, where x is a natural number between 15 and 49. The conventional CFR for women born in 2000 is $CFR_{2000}(49)$.

Table 3 reports $TFR_b(49)$ for $b = 1951, 1955, 1960, 1965, \ldots, 2000$ using the Census data and compare them with officially calculated TFR. The officially calculated TFR is derived by the Statistical Office of Korea using administrative data. The first and second columns present $TFR_b(49)$ using the Census data and officially calculated TFR, respectively. Our measures tend to understate the officially calculated TFR. However, for $b = 1975, \ldots, 2000$, the difference seems negligible.

Next, we compare $TFR_b(49)$ and $TFR_b(40)$ for $b = 1951, 1955, 1960, 1965, \ldots, 2000$ both using the Census data. This is to demonstrate that fertility in the 40s are negligible and that the two estimates are close to each other. This is useful to $TFR_b(40)$ will be directly comparable with $CFR_b(40)$. The first and third columns of Table 3 suggest that the two estimates are almost same except for older cohorts $b = 1951, 1955, 1960$.

Table 3. Generalized Total Fertility Rates from Census and Population

B	(1) $TFR_b(49)$ Census	(2) $TFR_b(49)$ Official	(3) $TFR_b(40)$ Census	(4) $CFR_b(40)$ Census
1951	4.19	4.99	3.92	2.39
1955	3.83	4.53	3.64	1.93
1960	2.81	3.43	2.74	1.75
1965	2.37	2.82	2.35	1.76
1970	1.54	1.66	1.53	1.63
1975	1.51	1.57	1.50	1.50
1980	1.63	1.63	1.62	1.36
1985	1.44	1.48	1.43	
1990	1.07	1.09	1.06	
1995	1.21	1.23	1.19	
2000	1.22	1.24	1.20	

TFRs in columns (1) and (3) are derived by authors' calculation using Census 2% sample.

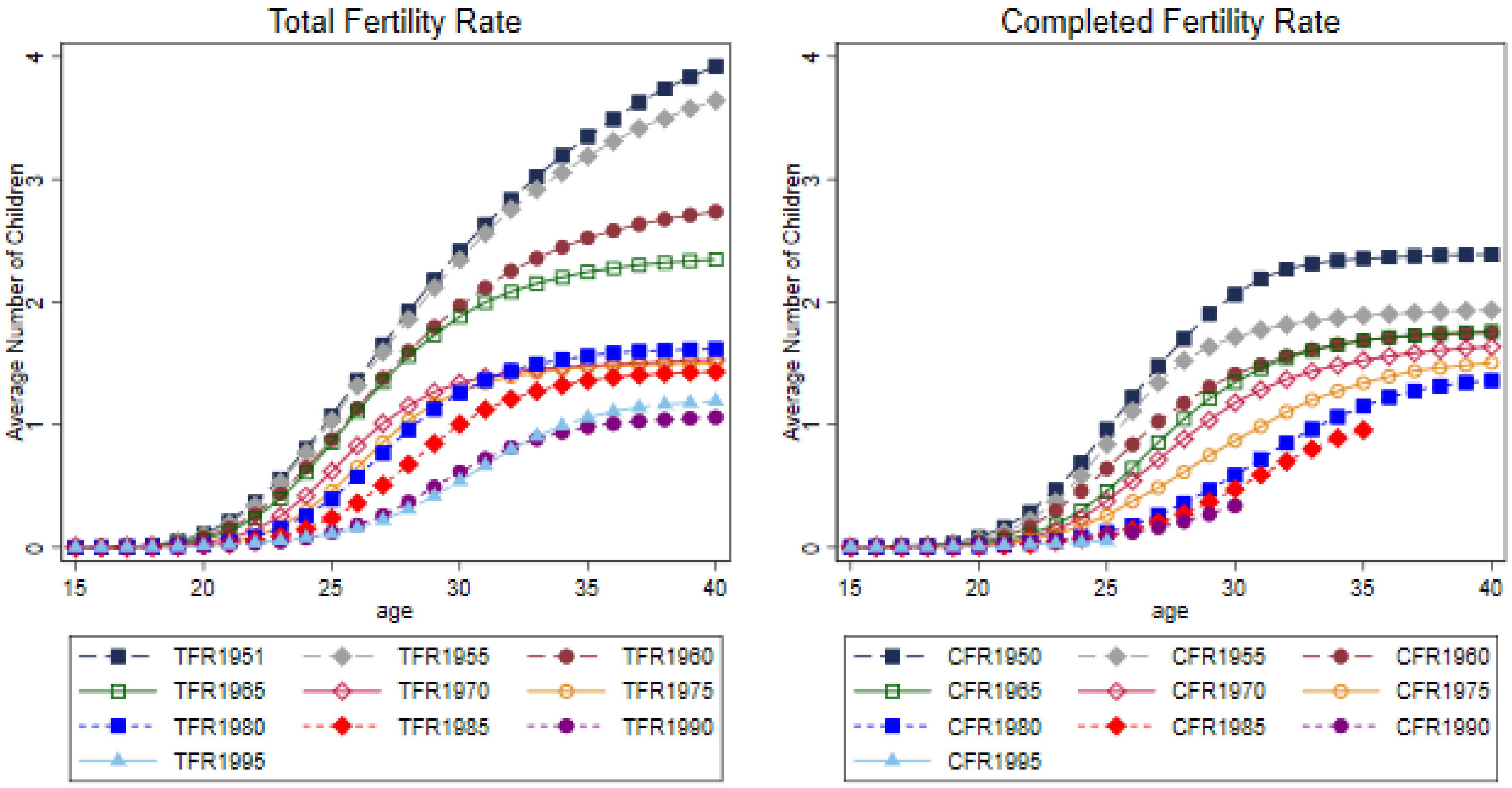
TFRs in column (2) are from Statistical Office of Korea.

CFRs in column (4) are derived by authors' calculation using Census 2% sample.

We compare figures for $TFR_b(40)$ and $CFR_b(40)$ for $b = 1951, 1955,1960, \dots, 1995$. The objective of this exercise is to verify that when there is a decreasing trend in fertility, the TFR will overstate the true fertility. According to our TFR calculation, the 1951 cohort was expected to have 3.92 children by age 40. However, the 1951 cohort actually had 2.39 children on average over the lifetime according to the realized CFR. For the 1955, 1960, and 1965 cohorts, the TFR is an overestimate of actual number of children born by 1.71, 0.99, and 0.59 children, respectively. The discrepancy between the TFR and the CFR narrows for the 1970 and 1975 cohorts but widens for the 1980 cohort. For more recent cohorts, the TFR estimates drop to 1.43 for the 1985 cohort and continue decreasing for younger cohorts, but comparison with the CFR is not possible because the CFR needs to be predicted. Figure 1 illustrates $TFR_b(a)$ and $CFR_b(a)$, respectively, for $a = 15, 16, \dots, 40$ for each $b = 1951, 1955,1960, \dots, 1995$.

The generalized CFR in Figure 1 reveals interesting findings because we can observe the mother's age at birth. For example, the 1960 and 1965 cohorts have the same CFR at age 40 as the two lines meet at age 40. The CFR 1965 line, however, is below the CFR 1960 line between ages 20 and 30. It implies that the 1965 cohort delay having children than the 1960 cohort. Two notable observations are evident in the right panel of Figure 1. First, younger cohorts exhibit lower fertility rates. We observe a decreasing trend in fertility rates, denoted by $CFR_b(40)$, with respect to mother's birth year. Indeed, $CFR_b(age)$ decreases for any given age with respect to b. Second, more of the younger cohorts give birth at older ages compared to earlier cohorts, as evidenced by the steeper birth schedule at older ages for younger cohorts. Delaying childbirth does not alter the completed number of childbirths over an individual's lifetime, but does contribute to a decrease in the population due to children being born later. Overall, younger cohorts tend to have fewer births compared to earlier cohorts, and they exhibit a tendency toward later childbirth.

Figure 1. Generalized Total Fertility and Completed Fertility Rates by Cohorts



5. Empirical Analysis

5.1. Empirical Specification

This section discusses how to predict $CFR_b(x)$ of the future using realized $CFR_b(x)$ of the past. For a woman who has ever given birth, we define t_0 be the age at her most recent childbirth. For a woman who has not experience a childbirth, t_0 is set to be 15. The survival time, t, is defined as the number of years since t_0. Using the natural logarithm of the survival time, $logt_i$, we consider an accelerated failure time model specified by

(1) $$logt_i = \beta_0 + \beta_1 t_{0i} + \beta_2 t_{0i}^2 + \beta_3 t_{0i}^3 + \beta_4 nochild_i + \beta_5 nborn_i + \beta_6 educ_i$$
$$+(\gamma_0 + \gamma_1 t_{0i} + \gamma_2 t_{0i}^2 + \gamma_3 t_{0i}^3 + \gamma_4 nochild_i + \gamma_5 nborn_i + \gamma_6 educ_i) * yob_i + \varepsilon_i,$$

where $nochild$ is a dummy variable that takes on a value of one if there has been no childbirth by t_0., $nborn$ is the cumulative number of children born by t_0, $educ$ is the years of education, yob is the year of birth, and ε_i is the error term. We interact year of birth with all the other control variables to reflect any possible trend by birth cohort.

The error term is assumed to have a logistic distribution, which is equivalent to specifying a loglogistic survivor function. The corresponding hazard rate initially increases, reaches its peak, and then decreases. This property goes well with the observation that births are concentrated at a specific age range. For example, the hazard of having a child increases initially and the chance of having a child later in the life declines.

5.2. Estimation Results

Table 4 presents the estimation results. The age at time zero is set to 15 for individuals who have not given birth before. For those who have previously given birth, the age at time zero corresponds to the age at their most recent birth. Note that a positive coefficient indicates an increase in the time it takes for a childbirth to occur. This would imply a longer duration or delay in having a child, which can be interpreted as a decreased probability or reduced likelihood of having a childbirth over the lifetime.

In Column (1), we present the results from the simplest specification, where our interpretation focuses on the signs of the estimates. A negative coefficient on the age at time zero suggests that the duration of birth becomes shorter with age, indicating an increased likelihood of giving birth as individuals grow older. The positive coefficient on the number of children implies that the duration of birth increases with the number of children. This suggests that mothers with more children are less likely to have additional births. Similarly, education is associated with an increased duration, suggesting that more educated individuals are less likely to give birth.

Column (2) investigates whether females from different cohorts exhibit varying tendencies towards childbearing. The coefficient for the year of birth is positive and statistically significant, suggesting that the younger cohort is expected to take longer to reach childbirth compared to the older cohort at the same age. For instance, when comparing two females born in 1970 and 1980, the duration to childbirth for the younger cohort woman at age 30 is expected to be approximately 19% (=1.9% * 10) longer than that of the older cohort woman at the same age.

In column (3), we extend the specification by interacting the year of birth with age at time zero, number of children at time zero, and education. A notable finding emerges, indicating that the inclination to delay childbirth in response to educational attainment is more pronounced among younger cohorts compared to older cohorts. Accompanied by the fact that younger generation has higher levels of education on average, younger cohorts are less likely give birth of a child than older cohorts. Furthermore, younger cohorts demonstrate a higher tendency to postpone subsequent childbirths when compared to their older counterparts. Consequently, there is a decreased likelihood of giving birth over their lifetime, as compared to the older generation. Adding a cubic polynomial of age at time zero in column (4) does not alter the coefficients for the number of children and education interacted with the year of birth, as observed in column (3).

In column (5), we present the results from the full specification, which additionally includes a dummy variable indicating whether the woman has not experienced childbirth by time zero, along with a variable that interacts this dummy variable with the year of birth. The findings suggest that the duration to first childbirth is shorter than the duration to subsequent childbirths. However, the duration to first childbirth is increasing for younger cohorts. For example, when comparing two females born in 1970 and 1980, the duration to first childbirth for the younger cohort woman expected to be approximately 47% (=4.7% * 10) longer than that of the older cohort woman. We utilize the results from column (5) to predict future fertility rates in the next section.

Table 4. Duration Model Estimates

VARIABLES	(1) $log(t)$	(2) $log(t)$	(3) $log(t)$	(4) $log(t)$	(5) $log(t)$
Age at t_0	-0.133***	-0.134***	-0.059***	-4.148***	-4.255***
	(0.000)	(0.000)	(0.001)	(0.048)	(0.107)
Age^2 at t_0				0.167***	0.163***
				(0.002)	(0.004)
Age^3 at t_0				-0.002***	-0.002***
				(0.000)	(0.000)
No Children by t_0					-1.074***
					(0.077)
Num of Children by t_0	0.753***	0.798***	0.719***	-0.058***	-0.124***
	(0.002)	(0.002)	(0.011)	(0.013)	(0.013)
Education	0.097***	0.060***	-0.067***	-0.046***	-0.047***
	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)
Year of Birth (YoB)		0.019***	0.012***	-0.079***	-0.444***
		(0.000)	(0.000)	(0.006)	(0.014)
Age at t_0 * YoB			-0.001***	0.016***	0.051***
			(0.000)	(0.001)	(0.002)
Age^2 at t_0 * YoB				-0.001***	-0.002***
				(0.000)	(0.000)
Age^3 at t_0 * YoB				0.000***	0.000***
				(0.000)	(0.000)
No Children * YoB					0.047***
					(0.001)
Num Children * YoB			0.000**	0.020***	0.022***
			(0.000)	(0.000)	(0.000)
Education * YoB			0.002***	0.001***	0.001***
			(0.000)	(0.000)	(0.000)
Constant	3.302***	2.486***	2.706***	33.606***	36.720***
	(0.004)	(0.005)	(0.029)	(0.366)	(0.943)
Log(gamma)	-0.651***	-0.653***	-0.653***	-0.853***	-0.863***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2,228,990	2,228,990	2,228,990	2,228,990	2,228,990

The dependent variable is the natural logarithm of the survival time.
The error term follows a logistic distribution.
t_0 is either 15 or the age at the most recent childbirth.
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

5.3. Predicted Fertility

As our duration model assumes loglogistic distribution, the survivor function is given by

(2) $S(t_i) = P(T > t_i) = 1/(1 + (\lambda_i t_i)^{1/\gamma})$ where $\lambda_i = \exp(-x_i\beta)$.

Consequently, the chance that individual i who has not given birth for s years will give birth of a child that year is given by

(3) $h(s_i) = \frac{P(T \le s_i + 1) - P(T \le s_i)}{1 - P(T \le s_i)} = \frac{S(s_i) - S(s_i + 1)}{S(s_i)}$.

The hazard in (3) is used to make prediction. For example, consider the 1990 cohort. Individuals in this cohort are 30 in year 2020. For each individual in that cohort, if she does not give birth of a child in 2020, her age at time zero remains unchanged. However, if she gives birth of a child in 2020, her age at time zero is updated to her current age, 30. In this case, her number of children is updated. For each individual in every cohort in the data, her hazard is calculated by (3). A larger hazard implies a higher chance of having a child in the following year. Using each individual's hazard, we generate uniform random number to determine whether the individual gives birth of a child in 2021. Using the simulated birth outcomes, we repeat the prediction process. This simulation process is essential because the birth outcomes depend on their past outcomes. We use the results in column (5) of Table 4 to make the prediction.

We obtain the simulated standard deviation and the 95% confidence interval as follows. First, we randomly generate estimates from a multivariate normal distribution, with the mean equal to the estimates reported in column (5) of Table 4 and the associated covariance matrix. We then conduct the simulated prediction until age 40 using these estimates. This process is repeated 99 times. Using these simulated fertility values, we obtain the standard deviation and the empirical 95% confidence interval for each prediction.

Figure 2 graphically displays the predicted fertility patterns using the 95% confidence intervals.[6] These predictions suggest several points. First, the 1985 cohort's birth profile

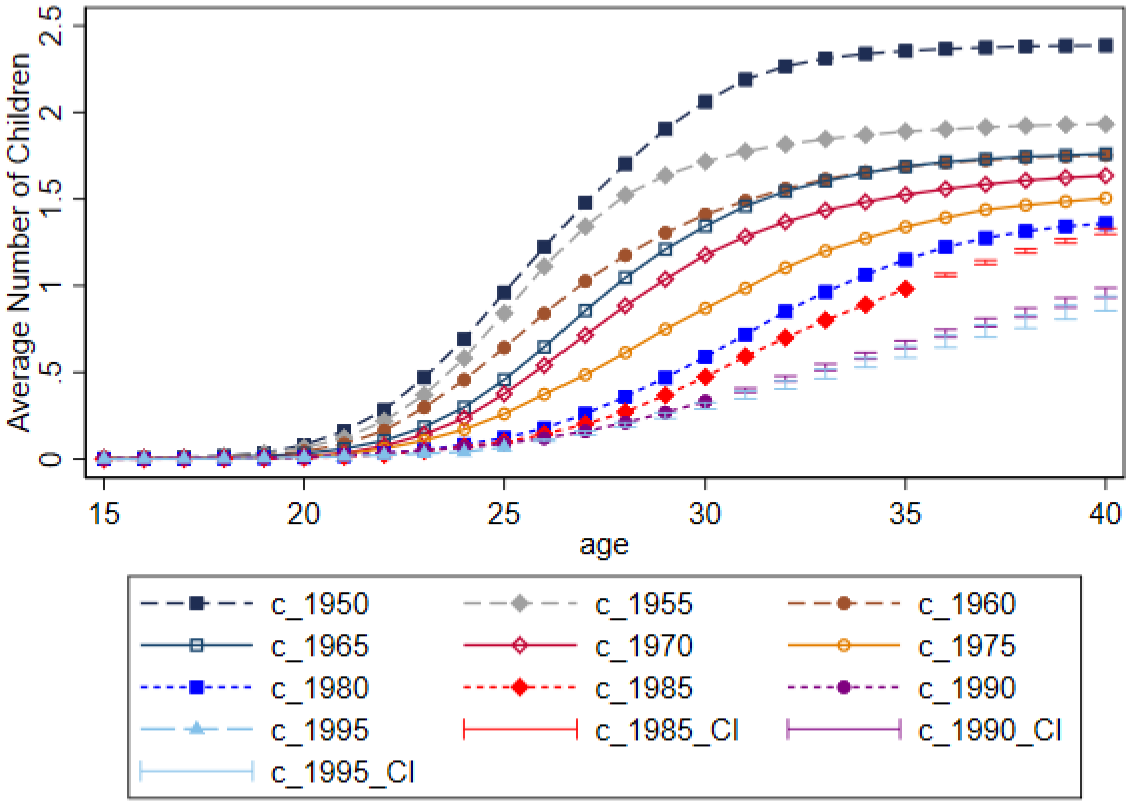
[6] Table A in the appendix presents the predicted fertility schedules and standard deviations for

consistently remains below the 1980 cohort’s birth profile throughout the age range, but by age 40, the 1985 profile gets close to the 1980 profile. The 1985 cohort is expected to have 1.31 children by age 40, similar to the 1980 cohort's prediction of 1.36 children. This convergence from below implies a delay in childbirth among the 1985 cohort compared to the 1980 cohort, rather than a decrease in the number of children over their lifetime.

Second, the prediction results for the 1990 and 1995 cohorts suggest that these cohorts will have a smaller number of children overall. This is evidenced by the fact that the birth schedules of these two cohorts remain substantially below that of the 1985 cohort, and the disparities in fertility rates do not diminish as they age. Specifically, the 1990 cohort is anticipated to have 0.96 children, while the 1995 cohort is expected to have 0.90 children. The widening of confidence intervals with age, particularly for younger cohorts, reflects increasing uncertainty in predictions as individuals age.

the 1985, 1990, and 1995 cohorts.

Figure 2. Realized Birth Schedules and 95% Confidence Intervals of Predicted Births



6. Concluding Remarks

This paper presents a novel approach to forecasting fertility patterns using micro data, focusing on individual-level observations to enhance predictions of cohort-level fertility rates. Our approach contributes significantly to the literature in several aspects. First, our method leverages the advantages of completed fertility rates, which provide more reliable fertility measures when birth rates differ by cohorts. Second, it enables the examination of how individual-level factors, such as individual attributes and birth history, affect fertility schedules. Third, our method is based on individual-level data, resulting in more precise projections with tighter confidence intervals. Finally, we predict fertility outcomes, encompassing both the timing and total number of childbirths.

Utilizing the Census Korea 2% data, we construct detailed birth history outcomes to understand behavioral patterns across different cohorts. We find that younger cohorts tend to delay childbirth compared to earlier cohorts, with educational attainment playing a significant role in this trend. In addition, individuals with a higher number of previous children tend to have longer durations between childbirths, indicating a reduced likelihood of having additional children. The predicted fertility schedules for the 1985, 1990, and 1995 cohorts provide insights into future fertility trends. The 1985 cohort is expected to delay childbirth, but they have a similar number of children compared to the preceding cohort. The 1990 and 1995 cohorts, on the other hand, are projected to have fewer children overall.

REFERENCES

Alkema, Leontine, Adrian E. Raftery, Patrick Gerland, Samuel J. Clark, François Pelletier, Thomas Buettner, and Gerhard K. Heilig (2011). Probabilistic Projections of the Total Fertility Rate for All Countries, Demography, 48(3), 815-839.

Bijak, Jakub. 2010. Forecasting International Migration in Europe: A Bayesian View. Springer Series on Demographic Methods and Population Analysis. Vol. 24. Dordrecht: Springer.

Bijak, Jakub, and John Bryant. "Bayesian demography 250 years after Bayes." Population studies 70.1 (2016): 1-19.

Bloom, D. E. (1982), "What's Happening to the Age at First Birth in the United States? A Study of Recent Cohorts," Demography, 19, 351–370.

Booth, Heather. 2006. Demographic forecasting: 1980 to 2005 in review, International Journal of Forecasting 22 (3): 547–581.

Bryant, John R., and Patrick J. Graham (2013). Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources, Bayesian Analysis, 8(3), 591-622.

de Beer, J. (1985), "A Time Series Model for Cohort Data," Journal of the American Statistical Association, 80, 525–530. [501]

Girosi, F. and King, G. (2008). Demographic Forecasting. Princeton: Princeton University Press.

Gorbey, Susi, Doug James, and Jacques Poot. 1999. Population forecasting with endogenous migration: an application to trans-Tasman migration, International Regional Science Review

22(1): 69–101.

Lee, Ronald D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level, International Journal of Forecasting, 9, 187-202.

Lee, Ronald D., and Lawrence R. Carter (1992). Modeling and Forecasting U. S. Mortality, Journal of the American Statistical Association, 87, 659-671.

Lynch, S. and Brown, J. (2010). “Obtaining multistate life table distributions for highly refined subpopulations from cross-sectional data: A Bayesian extension of Sullivan’s method.” Demography, 47(4): 1053–1077.

Raftery, Adrian E., Leontine Alkema, and Patrick Gerland (2014). Bayesian Population Projections for the United Nations, Statistical Science, 29(1), 58-68.

Raftery, Adrian E., and Hana Sevcikova (2022) Probabilistic population forecasting: Short to very long-term,_International Journal of Forecasting, forthcoming.

Schmertmann, Carl, Emilio Zagheni, Joshua R. Goldstein, and Mikko Myrskylä (2014) Bayesian Forecasting of Cohort Fertility, Journal of the American Statistical Association, 109(506) , 500-513.

Wiśniowski, Arkadiusz, Peter W. F. Smith, Jakub Bijak, James Raymer, and Jonathan J. Forster (2015). Bayesian Population Forecasting: Extending the Lee-Carter Method, Demography, 52(3), 1035-1059.

Appendix

Table A. Predicted Fertility

	C1985		C1990		C1995	
Age	Cumulative Fertility Estimates	SD	Cumulative Fertility Estimates	SD	Cumulative Fertility Estimates	SD
26					0.11	0.0037
27					0.15	0.0058
28					0.20	0.0070
29					0.25	0.0086
30					0.31	0.0100
31			0.40	0.0055	0.37	0.0110
32			0.46	0.0070	0.43	0.0120
33			0.53	0.0083	0.49	0.0129
34			0.60	0.0098	0.56	0.0142
35			0.66	0.0107	0.62	0.0151
36	1.06	0.0044	0.72	0.0109	0.68	0.0160
37	1.14	0.0059	0.79	0.0116	0.74	0.0166
38	1.20	0.0061	0.85	0.0122	0.80	0.0175
39	1.26	0.0067	0.90	0.0133	0.85	0.0184
40	1.31	0.0074	0.96	0.0136	0.90	0.0184