Modeling the Age Pattern of Fertility: An Individual-Level Approach

Daniel Ciganda ✉[1,2] and Nicolas Todd[3]

[1]Max Planck Institute for Demographic Research
[2]Statistics Institute, University of the Republic, Montevideo, Uruguay
[3]Musée de l'Homme - CNRS

Abstract

Macro-level modeling remains the dominant approach in many demographic applications, including population projections. Individual-level models tend to require larger amounts of data and very often cannot be estimated. The approach we introduce in this article attempts to overcome these limitations. Using likelihood-free inference techniques, we show that it is possible to infer the parameters of an individual-level model of the reproductive process from a set of aggregate fertility rates. By estimating individual-level quantities from widely available aggregate data, this approach can contribute to a better understanding of reproductive behavior and its driving mechanisms. It also allows for a more direct link between individual-level and population-level processes. We illustrate our approach using data from three natural fertility populations. The models we introduce can be readily applied in biodemography, to study the reproductive process in non-human populations.

Keywords—

1 Introduction

Modeling the age distribution of fertility rates is an essential step in a number of demographic applications. Some of this applications only require a very accurate fit to an observed schedule, as in the generation of single-age rates from grouped data. This type of problems are usually handled with non-parametric models, typically based on spline functions. Other applications require models that can fit the data well at the same time they provide a well-defined, ideally small set of parameters. Preferably, these parameters should represent quantities that can be interpreted in "demographic" terms. This typically means ages associated with relevant characteristics of the fitted curve, like the age at which the curve becomes no zero, the age at which it peaks, or the age at which it goes back to zero.

The interpretability of parameters is particularly relevant in a forecasting context, where it is useful to associate a change in the value of the parameters with an underlying behavioral process, such as fertility postponement or the diffusion of contraceptive methods.

One of the issues with most parametric models is that the interpretation of their parameters can be elusive (Hoem et al., 1981). Even in the best cases, the relationship between mechanisms and parameters is ambiguous and indirect. After all, these models are defined one level above the level at which behaviors are observed.

A potential solution to this problem is to model at the individual level. Such an approach would involve four steps: 1) develop a model of the reproductive process; 2) use this model to generate synthetic data and compute a set of simulated age-specific fertility rates (ASFR); 3) estimate the parameters of the model by minimizing the distance between simulated and observed rates.

Although all of the building blocks have been available for a while, this approach has not yet been systematically explored in full.

The tools and ideas required for the first of its steps have received the attention of a relatively small but significant group of researchers that contributed to the development of the first analytical models of the reproductive process (Gini, 1924; Henry, 1953; Sheps et al., 1973, among others).

These ideas where later explored using simulation methods, with the objective of incorporating more complex, although basic characteristics of the reproductive process like age-dependent fecundability, or to represent change over time in fertility rates (see: Ridley and Sheps, 1966; Barrett, 1971; Le Bras, 1993). The use of simulation methods allowed researchers to complete step 2, by generating synthetic age-specific fertility rates from individual level models and drawing a direct connection between behavior and aggregate demographic indicators.

These contributions, however, were largely theoretical. Parameter values where borrowed from previous studies, or calibrated through trial and error until the simulated data would fit a given target distribution.

Interest in these type of models dwindled after the 1980's, just at the same time the first statistical methods that would enable inference on complex simulation models were starting to emerge Rubin (1984).

Likelihood-free inference methods, such as Approximate Bayesian Computation (ABC), have been researched extensively since then, providing a robust statistical framework to estimate the parameters of computational models in a number of scientific fields Beaumont (2019). What we attempt to show in the remainder of this paper, is that these advances in statistical computing provide the missing piece to the program outlined above, allowing to infer quantities that describe the reproductive process at the individual-level from aggregate fertility rates. More generally, the approach we introduce here provides a clear roadmap to fully integrate behavioral mechanisms in the modeling and forecasting of fertility trends.

All the results presented in this article are fully reproducible using the code and data available at `https://github.com/dciganda/comfert_natural`.

2 Model

The simplest form of the reproductive process is the one in which the outcomes of the process, the number and timing of births, are not a function of individual preferences. Louis Henry called this "natural fertility", in an attempt to characterize a process that is exclusively determined by physiological factors.

To illustrate our approach we use data from three natural fertility populations: Hutterites communities from 1860 to 1914, the French-Canadian population from 1700 - 1750, and a subsample of the French population from 1680 to 1760 (see Section 3 for details).

A model of the reproductive process in a natural fertility context requires three basic inputs: The moment when the process starts, typically defined as the age at marriage; the risk or probability of a conception, known as fecundability; and the length of the period in which women are not able to conceive following childbirth, know as post-partum amenorrhea.

We represent the reproductive experience of a cohort of women from birth to age 50 using a discrete-time simulation model, in which time advances at fixed increments of one month. The age (in months) at marriage for a woman i is simulated from a lognormal distribution with mean μ_m and standard deviation σ_m.

Every month, married women who are neither pregnant nor amenorrheic are exposed to the risk of a conception. We model this pattern using two polynomial basis:

$$\phi(x) = \phi_1 \cdot (3 * x_s^3 - 6 * x_s^2 + 3 * x_s) + \phi_2 \cdot (-3 * x_s^3 + 3 * x_s^2) \tag{1}$$

Where $f(x)$ is the fecundability of women during the reproductive window from age 10 to 50, and x_s is the reproductive window scaled in the 0-1 range. The coefficient that multiplies the first polynomial basis, ϕ_1, influences the height of peak fecundability, while ϕ_2 influences the pace at which fecundability decays until permanent sterility is reached at menopause. To determine whether or not woman i conceives in a given month, a Bernoulli trial is simulated with probability $\phi(x_i)$.

Figure 1 presents three patterns associated with different combinations of parameter values for the risk of conception.

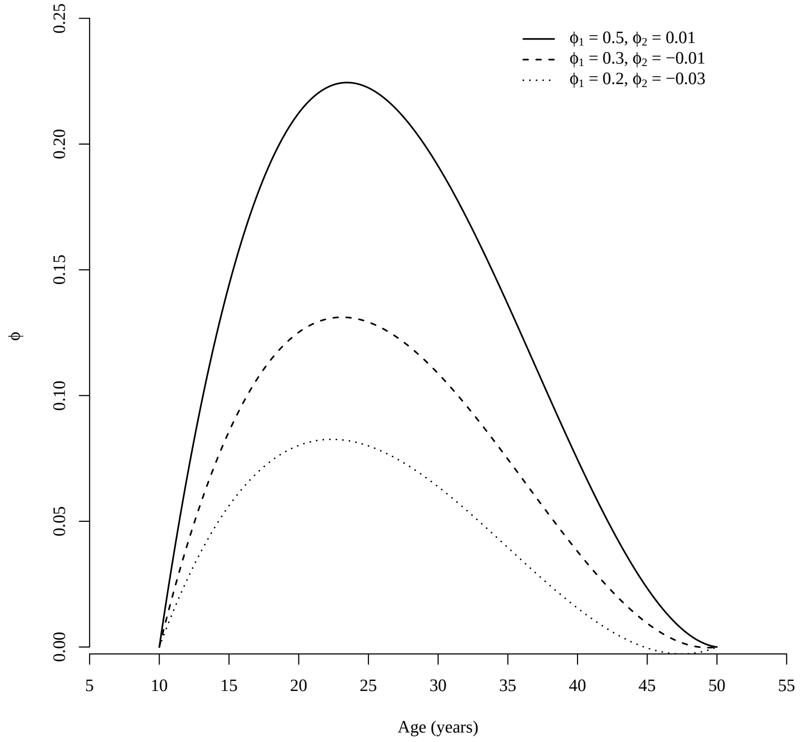


Fig. 1 | Age Distribution of Fecundability for Different Combinations of Parameter Values. Although it depends on only two parameters, the model used to represent the evolution of the risk of conception across the reproductive life course captures a wide range of patterns.

After a conception is observed, women enter a non-susceptibility state (no risk of conception) for the duration of the pregnancy, 9 months, plus the duration of post-partum amenhorrea, δ, which is estimated together with the rest of the parameters of the model.

The model outputs individual reproductive trajectories, from which a set of simulated age-specific fertility rates ${}_sf(x)$ can be computed. We then leverage the distance between simulated and observed rates, ${}_of(x)$, to estimate the parameters of our model. This procedure is explained in more detail in Section 3.5, after we describe the data used in our analysis.

3 Data

We use 3 high quality natural fertility dataset which provide full maternity histories plus information on the dates of other relevant events in a woman's life course like marriage, death, separation, and death of spouse.

3.1 Hutterites

The Hutterites are a Anabaptist community, originated in the XIV century in the Tyrolean Alps. After a long history of migration in Europe and Asia, they relocated to North America at the end of the XIX century where they still live today in self-sustained, largely isolated colonies.

Like many religious communities, the Hutterites oppose the use of birth control methods (Lee and Brattrud, 1967; Ingoldsby and Stanton, 1988). What made them stand out, however, was how strictly they adhered to these beliefs, at least until the second half of the 20th century. The marital fertility of the Hutterite cohorts born until the early 1900s remained close to the theoretical maximum, with an average of around 10 children per woman, providing researchers with an exceptional opportunity to study the reproductive process under natural fertility conditions (Eaton and Mayer, 1953).

Another reason why the Hutterites became the gold standard for natural fertility research was their custom of keeping detailed family records. These records, personally checked for consistency by colony preachers, were made available for various scientific studies in the 1950s and 60s (Eaton and Mayer, 1953; Mange, 1964). These data was extended in the course of one of these studies through follow-up interviews, which resulted in complete maternity histories for 562 families Sheps (1965).

3.2 French-Canadian

BALSAC is a longitudinal population database that contains information on individuals and families who lived in Quebec from the 17th century to the present. The information used in this article, that concerns birth cohorts before 1750, was gathered using family reconstitution methods from parish registers (Vézina and Bournival, 2020). Given the high quality of these records, the information on XVII Century Quebec populations has become an important reference in the literature on natural fertility (see: Clark et al., 2020; Larsen and Yan, 2000; Eijkemans et al., 2014, among others).

3.3 Henry

Another important dataset in the research on natural fertility is the Enquete Louis Henry. The information on this dataset was gathered by Louis Henry... It contains data on life events for a sample of 378 parishes for the period 1670 to 1829 from parish registers and administrative records. The families in a sub-sample of 40 rural parishes have been fully reconstituted.

3.4 Sample Sizes

To keep things simple, we model a process without death or union dissolution. Therefore, our data belongs to "intact" marriages, i.e., marriages that do not dissolve by death or separation before the woman reaches age 50. Table 1 contains the remaining sample sizes after we exclude censored trajectories and records with incomplete information. Records with partial information were recovered in some cases by imputing the month or day of events. The inclusion of imputed data does not significantly affects our results.

While our simulated data belongs to a single birth cohort, in order to get reliable estimates our empirical data is obtained from multiple cohorts. To minimize potential biases we kept a range of birth cohorts that are as homogeneous as possible with respect to the age at first marriage and the total fertility rate (see appendix X). The cohorts included for each population are also displayed in Table 1

Table 1: Sample Information for the Three Natural Fertility Datasets

	Nr. Marriages	Nr. Births	Cohorts
Hutterites	161	1726	1860 - 1914
XVIII Century Quebec	14303	110772	1722 - 1730
XVII - XVIII Century France	3235	18623	1680 - 1760

Figure 2 shows the observed age-specific fertility rates for the three natural fertility populations. As expected, the Hutterites show consistently high fertility from age 20, with an average number of children per woman in these cohorts of 10.76. The Hutterite data is also unique in terms of its pattern, with a visible hump around ages 24-26, followed by a decline that is slow until age 35 and accelerates afterwards. The French-Canadian cohorts have the second largest fertility in the three populations, with an average of 7.7 children per woman, followed by the French cohorts with an average of 5.75 children per woman. The shape of the ASFR is more conventional in these last two populations, characterized by a rounded top, which in the case of the French-Canadian cohorts is found around ages 28 and in the case of the French cohorts around age 32. This is not surprising considering the mean ages at marriages for these cohorts which is 20.4 for the Hutterites, 21.1 for the French-Canadian, and 24.1 for the French cohorts.

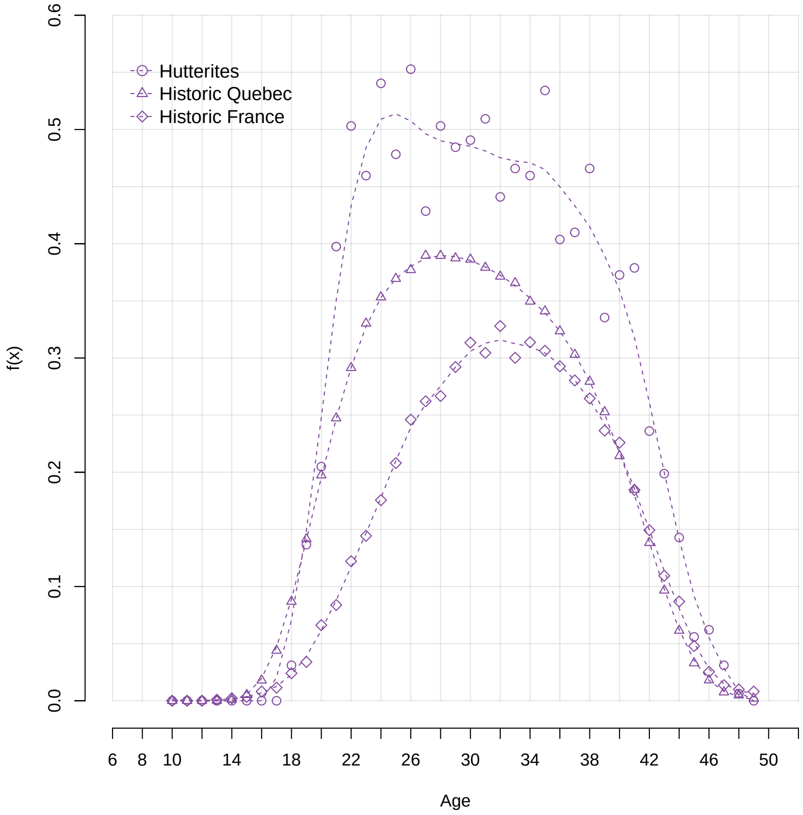


Fig. 2 | Age Specific Fertility Rates for Three Natural Fertility Populations.

3.5 Estimation

As previously mentioned, our primary goal is to demonstrate that the lack of individual-level data does not have to be an impediment to understanding the behaviors and mechanisms underlying fertility processes.

When only aggregate data is available, the probability of individual observations cannot be calculated, and model parameters cannot be estimated using a maximum likelihood approach. Instead, we rely on Approximate Bayesian Computation (ABC), a popular likelihood-free estimation approach for the estimation of computational models.

The core idea behind ABC is to estimate posterior distribution without explicitly calculating the likelihood function. Instead, ABC algorithms simulate data from the model and accept parameter values when the simulated data is sufficiently close to the observed data, based on a chosen distance metric.

We use the basic ABC rejection algorithm, which draws parameter samples from the prior distribution and rejects those outside a predefined distance threshold (Tavaré et al., 1997; Pritchard et al., 1999). While this method is straightforward to implement, it requires a large number of simulations to obtain a small set of accepted samples.

To improve the efficiency of the approach, a Gaussian Process (GP) regression adjustment was used in a second step. Here, previously accepted parameter values were modeled as a function of the simulated summary statistics to obtain a better approximation of the posterior distribution.

```
Algorithm: ABC + Regression Adjustment
Input: Prior distribution $\pi$, model $\mathcal{M}$ for simulation, observed data $y_{\text{obs}}$, distance
  threshold $\epsilon$.
Output: Adjusted samples from the posterior distribution $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_N\}$.
begin
    Draw parameter samples, $\theta^*$, from $\pi$
    Simulate data, $y^*$, from $\mathcal{M}(\theta^*)$
    Compute distance between $y^*$ and $y_{\text{obs}}$
    Accept $\theta^*$ if the distance is less than $\epsilon$
    With accepted parameters $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N\}$ and corresponding simulated data
      $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$:
    begin
        Fit a Gaussian Process regression model:
            $\hat{\theta} = m(\hat{y}) + \varepsilon$
        Adjust the values of the accepted parameters:
            $\tilde{\theta}_i = m(y_{\text{obs}}) + \varepsilon_i$
                $= m(y_{\text{obs}}) + (\hat{\theta}_i - m(\hat{y}_i))$
    end
    Return $\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_N\}$ as samples from the adjusted posterior distribution.
end
```

4 Results

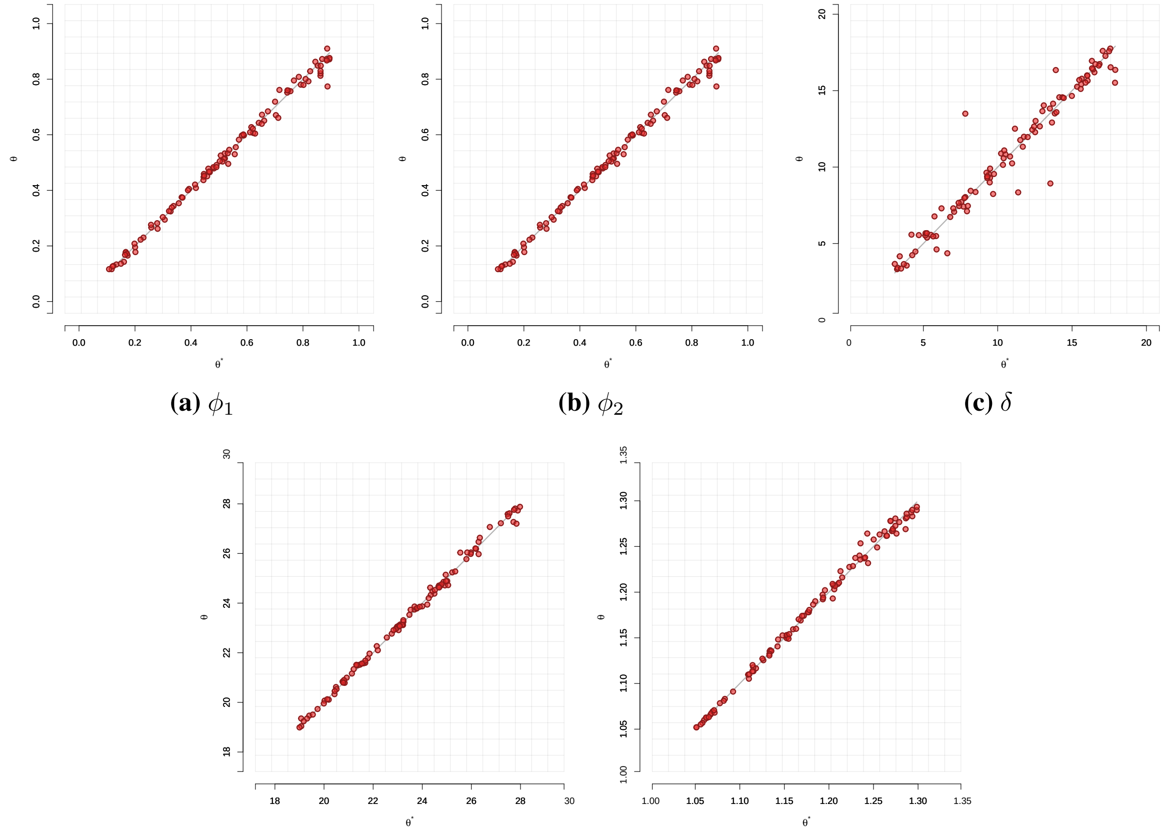
4.1 Model Validation

We use cross-validation to ensure that the parameters in our model can be correctly identified from a set of age-specific fertility rates. This procedure makes use of the set of parameter values θ^* and their corresponding simulated rates y^*. From this set we randomly select the i^{th} simulation as our validation set. The simulated rates in the validation set are taken as pseudo-observed values. Taking into account all simulations except the validation set, the model parameters are then estimated using the procedure described in the previous section. This process is repeated for a subset of 100 randomly chosen simulations and the prediction error is computed as:

$$\varepsilon_{\mathrm{p}} = \sum_{i} \left(\frac{(\theta_i^* - \theta_i)^2}{\mathrm{Var}(\theta_i)} \right)$$

The results of the cross-validation exercise are shown in Figure 3. They show that the estimation procedure consistently returns parameter values θ that are very close to the true parameters that generated the data θ^*. The parameter with the smallest prediction error is μ_m and the parameter with the largest error is δ. This is not unexpected as a change in the value of μ_m produces a shift of the entire distribution of rates that is very distinct from the effect of a value change in any other parameter. A change in the value of δ, on the other hand, produces an effect on the rates that overlaps to a certain extent to the effect of a change in the value of ϕ_1 (see appendix A for some animations illustrating the effect of each parameters on the schedule of age-specific fertility rates).

Fig. 3 | Results of Model Cross-Validation. The location of each of the red dots displayed in the figure is defined by the mean of the estimated posterior distribution in the y axis and the value of the parameter that generated the simulated results used as pseudo-observed data during the validation process.



(a) ϕ_1

(b) ϕ_2

(c) δ

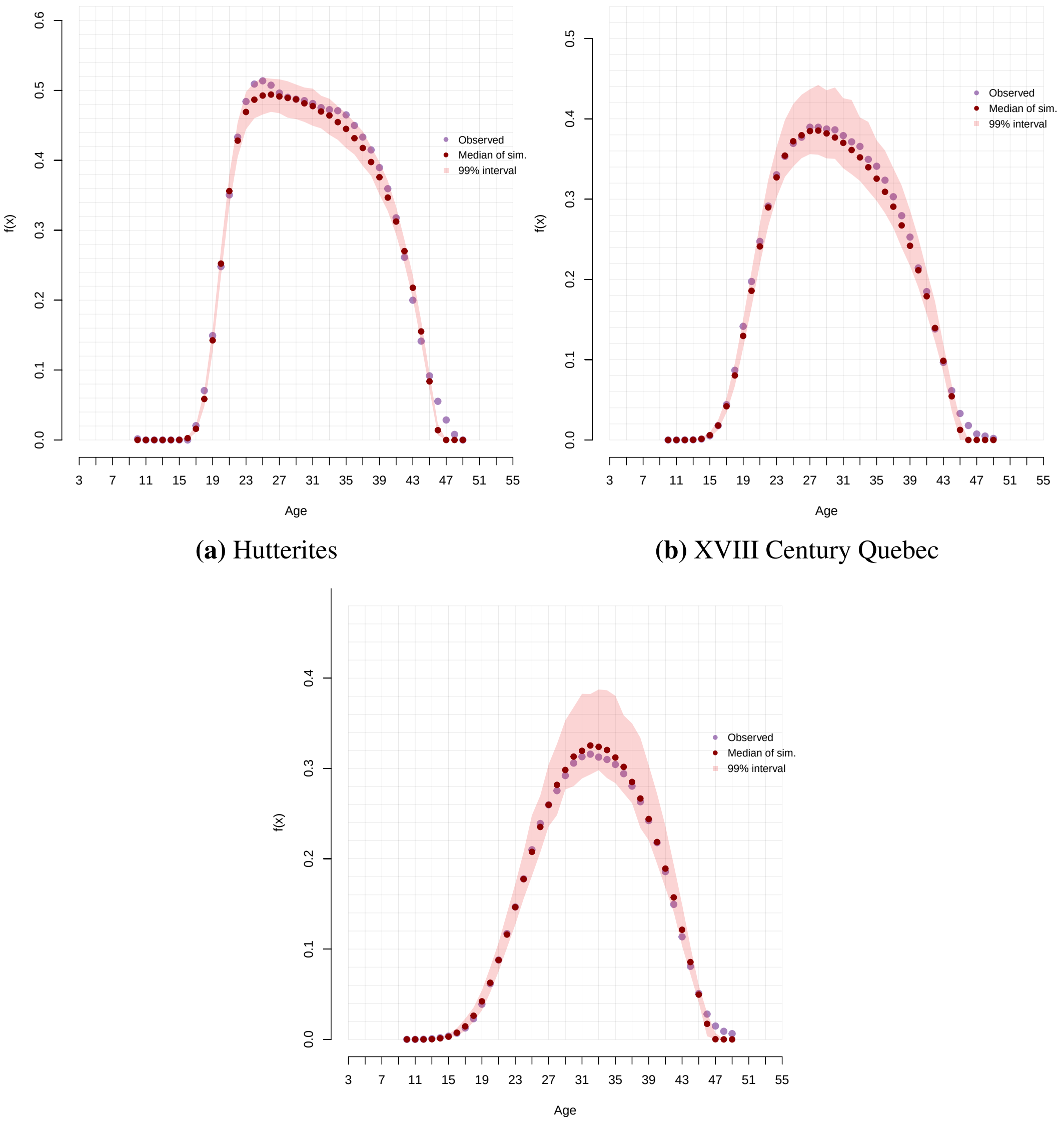
(d) μ_m

(e) σ_m

4.2 Model Fit

Figure 4 depicts the model's posterior predictive distribution against observed ASFR for the three populations analyzed. The model accurately captures the general characteristics of the data, such as the shape, location of peaks, and rate of decline of the three distributions. In fact, the fit is remarkable in all three cases, except for the rates at the very end of the reproductive age window (ages 47 to 50), which the model tends to underestimate.

Fig. 4 | Observed vs. Simulated Age-Specific Fertility Rates with 95% Credible Intervals.



(a) Hutterites

(b) XVIII Century Quebec

(c) XVII - XVIII Century France

A Appendix

Fig. 5 | Effect of ϕ_1 on Simulated Age-Specific Fertility Rates

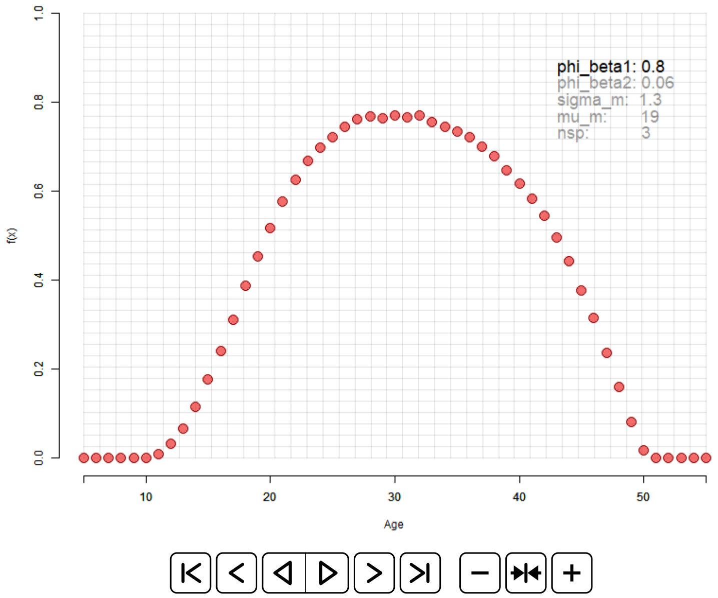


Fig. 6 | Effect of μ_m on Simulated Age-Specific Fertility Rates

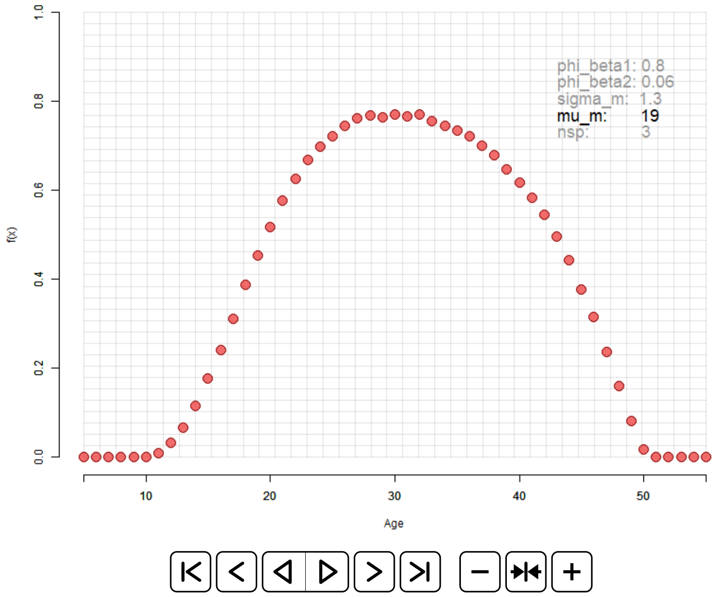
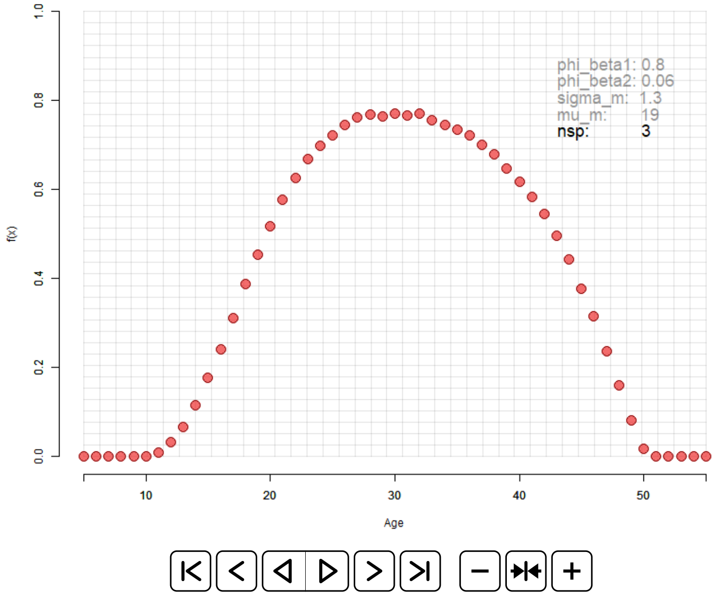


Fig. 7 | Effect of δ on Simulated Age-Specific Fertility Rates



References

Barrett, J. C. (1971). Use of a fertility simulation model to refine measurement techniques. *Demography 8*(4), 481–490.

Beaumont, M. A. (2019). Approximate bayesian computation. *Annual review of statistics and its application 6*, 379–403.

Clark, G., N. Cummins, and M. Curtis (2020). Twins support the absence of parity-dependent fertility control in pretransition populations. *Demography 57*(4), 1571–1595.

Eaton, J. W. and A. J. Mayer (1953). The social biology of very high fertility among the hutterites. the demography of a unique population. *Human biology 25*(3), 206.

Eijkemans, M. J., F. Van Poppel, D. F. Habbema, K. R. Smith, H. Leridon, and E. R. Te Velde (2014). Too old to have children? lessons from natural fertility populations. *Human Reproduction 29*(6), 1304–1312.

Gini, C. (1924). Premières recherches sur la fécondabilité de la femme. In North-Holland: (Ed.), *Proceedings of the International Mathematical Congress.*, Volume Vol. 2., Toronto.

Henry, L. (1953). Fondements théoriques des mesures de la fécondité naturelle. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute 21*(3), 135–151.

Hoem, J. M., D. Madsen, J. L. Nielsen, E.-M. Ohlsen, H. O. Hansen, and B. Rennermalm (1981). Experiments in modelling recent danish fertility curves. *Demography 18*(2), 231–244.

Ingoldsby, B. B. and M. E. Stanton (1988). The hutterites and fertility control. *Journal of Comparative Family Studies 19*(1), 137–142.

Larsen, U. and S. Yan (2000). The age pattern of fecundability: an analysis of french canadian and hutterite birth histories. *Social biology 47*(1-2), 34–50.

Le Bras, H. (1993). *Simulation of change to validate demographic analysis.* Oxford England Clarendon Press 1993.

Lee, S. and A. Brattrud (1967). Marriage under a monastic mode of life: A preliminary report on the hutterite family in south dakota. *Journal of Marriage and the Family*, 512–520.

Mange, A. P. (1964). Growth and inbreeding of a human isolate. *Human Biology 36*(2), 104–133.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution 16*(12), 1791–1798.

Ridley, J. C. and M. C. Sheps (1966). An analytic simulation model of human reproduction with demographic and biological components. *Population Studies 19*(3), 297–310.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.

Sheps, M. C. (1965). An analysis of reproductive patterns in an american isolate. *Population studies 19*(1), 65–80.

Sheps, M. C., J. A. Menken, and A. P. Radick (1973). *Mathematical models of conception and birth*. University of Chicago Press Chicago.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from dna sequence data. *Genetics 145*(2), 505–518.

Vézina, H. and J.-S. Bournival (2020). An overview of the balsac population database: past developments, current state and future prospects. *Historical Life Course Studies 9*, 114–129.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.