

A scale- and time-independent technique for estimating total fertility rates from age-sex distributions *

Mathew E. Hauer ^{1*} *University of Georgia*

Carl P. Schmertmann ^{2,3} *Florida State University*

Recent methodological advances in indirect migration and mortality estimation for small populations^{1,2,3} have revealed important patterns, but accurate estimation of fertility for small or specialized populations remains elusive. The primary fertility index for a population, the total fertility rate (TFR), requires accurate data on births disaggregated by mother's age. TFR is thus incalculable for the many areas and time periods that lack such information. Here we discuss a universal methodological framework for estimating TFR that uses inputs as minimal as the age-sex structure of a population. The implied total fertility rate (iTFR) accurately estimates fertility from a population's age-sex structure in a wide range of scales, time periods, and even species. We also discuss two extensions of the iTFR, called xTFR and BayesTFR, that offer improved accuracy with minimal additional data requirements. To demonstrate the utility of this approach, we produce the first complete county-level map of US fertility, reconstruct historical TFRs for three European countries up to 150 years prior to the collection of any birth records, and estimate TFR for the United States conditioned on household income, a variable that is not recorded in US birth records. Given its parameter-free nature, the method captures fundamental qualities that govern fertility with a near universal applicability across space and time. We anticipate that our methodological framework will be a starting point for more sophisticated fertility analyses in previously inestimable sub-populations, time periods, and geographies, significantly expanding our ability to understand the nature of fertility.

* Corresponding author. hauer@uga.edu. p: 706-542-9369.

¹ Carl Vinson Institute of Government, University of Georgia. 201 N. Milledge Ave. Athens, GA USA 30602.

² Department of Economics, Florida State University.

³ Center for Demography and Population Health, Florida State University.

*The data and code that supports this analysis are available in the supplementary materials.

Main Text

Fertility is the primary engine of global population change⁴ and is linked to the United Nation's Sustainable Development Goals for female education, child and maternal mortality, gender equality, and reproductive health⁵. The total fertility rate (TFR; the expected number of children born over a complete reproductive lifetime) is widely regarded as one of the most important components of population change and has been extensively used throughout the 20th Century in a range of applications^{6,7,8,9,10,11,12,13,14,15}.

Although the conventional technique for calculating TFR is straightforward, it requires data on births disaggregated by mother's age. This makes TFR incalculable in several situations: (i) for countries and regions that lack detailed birth records, (ii) for historical populations without vital event registration, such as the United States prior to 1933, (iii) for small-area populations when reporting agencies mask birth records for privacy reasons, and (iv) for any subpopulation not identified on official birth records, such as the women in a specific income decile, religion, tribe, or caste. The need for disaggregation of births by mother's age thus limits fertility analysis mainly to large populations in contemporary countries with good vital registration systems.

Numerous indirect estimation techniques have been proposed to circumvent these limitations^{16,17,18}. These methods are often regression-based, and they rely on covariates – such as mean age at marriage, percent of women ever married, etc. – that may be absent from census data and therefore need to be collected in surveys. The resulting scale- and time-dependent estimates are often inaccurate,^{19,20} and, like TFR, limited to areas, time periods, and populations with the requisite population and survey data.

Here we discuss a modeling framework that overcomes these problems and demonstrate the near universal applicability of a parameter-free estimation method.

We first approximate TFR using a variant of the general fertility rate (the total observed births divided by the total number of reproductive age women), in which we substitute the

number of enumerated children for recorded births. Then, based only on accurate population counts by age and sex, we develop more complex versions that can accurately estimate total fertility rates over numerous scales, time periods, sub-populations, and even species.

Demographic calculations in the Supplementary Material show that the expected number of surviving children of both sexes, per woman in reproductive age groups at the end of an n -year period is the product of three factors: $\frac{C}{W} = s_0 \cdot \bar{p} \cdot TFR$. These three factors are:

1. Child mortality, s_0 . This is the expected fraction of children still alive among those born in the past n years, derived from person-years in a life table population with radix $l_0 = 1$ as ${}_nL_0/n$.
2. A population-weighted index of the age-specific fertility schedule, \bar{p} . This represents the average share of lifetime fertility that women in the population experienced over the past five years, after a small adjustment for the possible mortality of adult women. It depends on the relative age pattern of fertility from menarche to menopause, $\phi_{15} \dots \phi_{45}$ for human populations, on the current age-specific populations of women ($W_{15} \dots W_{45}$), and on the potential mortality of women of childbearing ages ($L_{15} \dots L_{45}$).
3. The total fertility rate, TFR.

One can thus calculate TFR as

$$TFR = \frac{1}{s_0} \cdot \frac{1}{\bar{p}} \cdot \frac{C}{W}$$

the product of the child/woman ratio and two additional factors — a child mortality multiplier $1/s_0$ and an age structure multiplier $1/\bar{p}$ (equation: 4).

If women of reproductive age are uniformly distributed over seven five-year age groups, $a = 15 \dots 45$, then $\bar{p} \approx 1/7$ and the age structure multiplier is approximately 7. We show that the simplest approximation, child-mortality near zero ($s_0 \approx 1$) and a uniform age-distribution of women across the reproductive age groups ($\bar{p} \approx 1/7$), still produces accurate estimates

of total fertility within 0.25 children, making the approach parameter free without loss of significant accuracy (supplementary materials and Figure 1). This leaves the child/woman ratio as the dominant input to determining fertility levels. We refer to this simplest approximation as the implied or intrinsic total fertility rate (iTFR), representing the fertility rate implied by the age-sex structure of a given population^{20,21}. We calculate implied fertility as

$$iTFR = \frac{\beta - \alpha}{n} \cdot \frac{nC_0}{W} \quad (1)$$

where α and β are the minimum age at menarche and maximum age at menopause, respectively, nC_0 is the number of children under age n , and W is the number of women between ages α and β .

Equation 1 resolves the limitations of the conventional technique for calculating TFR, freeing the estimation of TFR from the burden of detailed birth records. Instead, the approach produces estimates with as few inputs as basic age-sex counts – common data collected in censuses. It also has no other tunable parameters allowing for a near universal deployment of the method. Of course the method does require accurate population data, making it more in-common with the demographic accounting equation than more traditional estimation methods. Thus, (eq. 1) represents the parameter-free, fundamental equation of implied fertility. Finally, the implied fertility rate can be applied within stochastic frameworks, estimating both TFR and its variance as well as the sensitivity of TFR calculation under various mortality regimes.

To test the flexibility of our measure and its ability to estimate TFR in a variety of data situations, we present three derivations of the implied fertility rate (see supplementary material) with increasing complexity and subsequent data requirements — a deterministic formulation based on the algebraic rearrangement, iTFR (equation: 1)(**Fig. 1, left column**), a regression-based formulation conditional on the age-structure (essentially incorporating $1/\bar{p}$ into the estimation), xTFR (equation: 10)(**Fig. 1, middle column**), and a complete Bayesian approach based on child- and maternal-mortality, the female reproductive

Table 1: **Summary statistics of the three derivations of the iTFR using data from the HFD and HMD.** APE is the Absolute Percent Error and RMSE is the Root Mean Squared Error. Overall, the methods produce TFR estimates on average within less than 1/10th of a birth.

Method	50th Percentile APE	90th Percentile APE	50th Percentile Absolute Error	90th Percentile Absolute Error
iTFR	4.7%	11.33%	0.089	0.247
xTFR	3.23%	8.31%	0.06	0.191
BayesTFR	2.59%	6.68%	0.051	0.146

age-structure, and potential underenumeration of children and reproductive-age women (incorporating both $1/\bar{p}$ and $1/s_0$), BayesTFR (equation: 18)(Fig. 1, right column). These deviations allow us to quantify the source of error in equation 1.

We verify the accuracy of the three derivations using data from the Human Fertility Database and Human Mortality Database^{22,23} for nearly 1,800 fertility schedules between 1891 and 2015 across 30 countries (see Table 2 in the Supplementary Materials for a complete list). This dataset comprises only complete, official vital event statistics and is the most complete and accurate collection of observed fertility data compiled to date (Fig. 1). We find good agreement between our implied fertility rates and the observed TFRs for all three methods (Fig. 1, a, b, c). Typically demographic parameters increase in accuracy as population sizes increase due to the law of large numbers, and they tend to be more accurate in more recent time periods due to improved data collection practices. However, we find that error rates are independent of population size and scale (Fig. 1, d, e, f) and do not vary across time (Fig. 1, g, h, i), suggesting scale- and temporal- independence uncommon in other indirect methods. Additionally, all three methods on average estimated total fertility within 0.1 observed births, with 90% of estimates from the poorest performing method within 0.25 births (Fig. 1, j, k). Finally, 90% of estimates were within 11% of the observed value, regardless of derivation, with the average estimates within 5% of the observed TFRs. Taken together, these results suggest a robust method with applicability across space and time, with better performance with additional model parameters. Even the simplest derivation with the least data requirements, implied fertility rate itself, produces accurate estimates of the TFR.

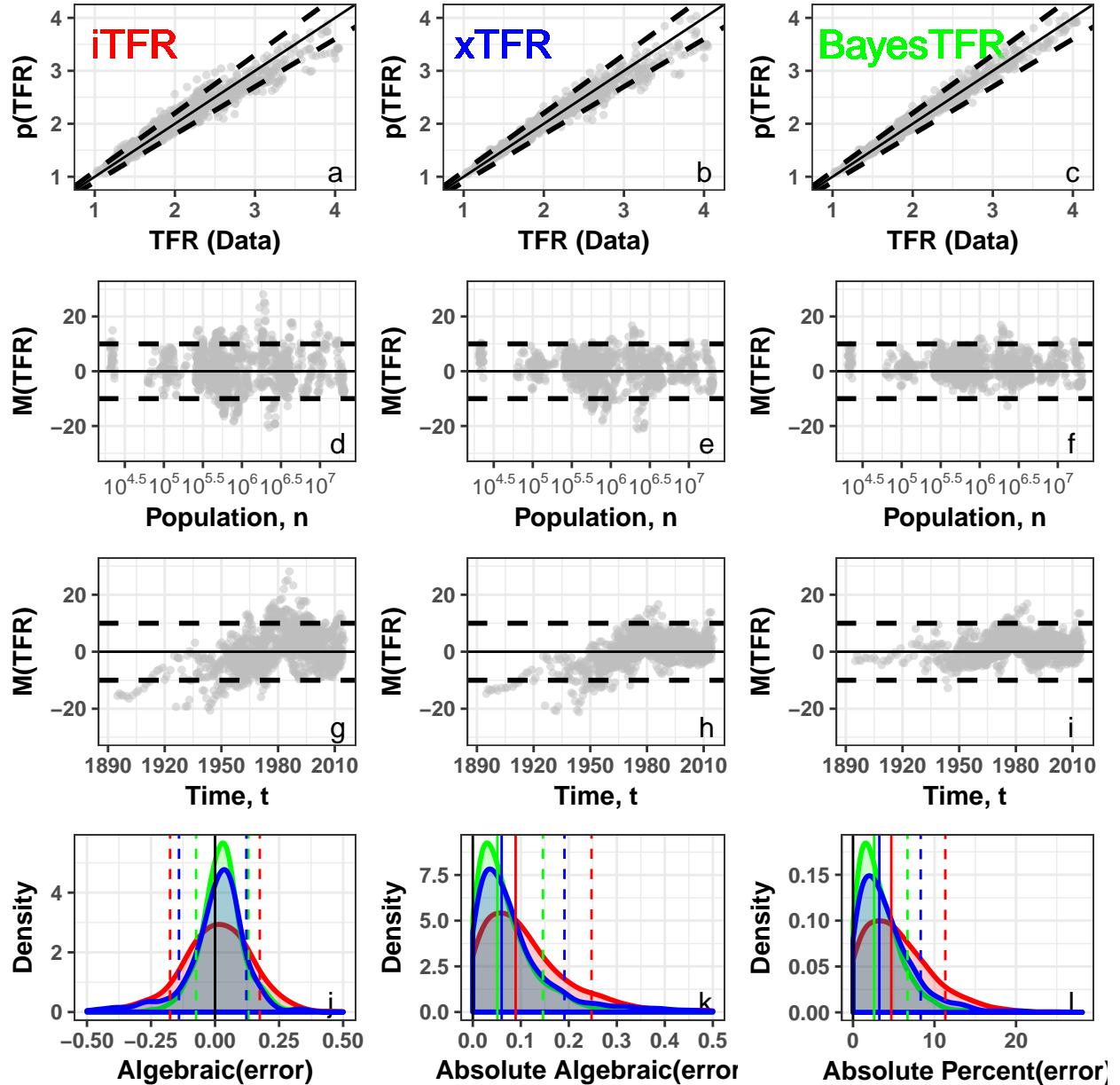


Figure 1: **The implied Total Fertility Rate.** We compare the performance of three formulations of the implied total fertility rate against observed total fertility rates. (a, d, g) in the first column use $p(TFR) = n(C/W)$; (b, e, h) in the second column use $p(TFR) = x(C/W)$ where x is derived from the ratio of women aged 25-34 to all women of childbearing ages; (c, f, i) in the third column estimates TFR conditional on C, W, k, q_5, β . (a, b, c) are scatter plots of the estimated TFR against the observed 5-year average TFR from fertility schedules in the Human Fertility Database for 1891-2015. The solid line is $Y=X$, and the dashed lines represent $\pm 10\%$ of $Y=X$. (d, e, f) demonstrate the percent error, $M(TFR)$, for each method against population sized n . The dashed lines represent 10%. (g, h, i) plot the percent errors, $M(TFR)$, for each method for the period 1891-2015, t . The dashed lines represent 10%. (j) plots the distribution of algebraic errors for each method. The dashed lines correspond to the 10th and 90th percentile errors for each method. (k) plots the distribution of absolute algebraic errors. (l) plots the distribution of absolute percent errors for each method. The solid lines correspond to the 50th percentile error, and the dashed lines correspond to the 90th percentile errors for each method. Regardless of formulation, the iTFR demonstrates scale and time independence to a high degree of accuracy.

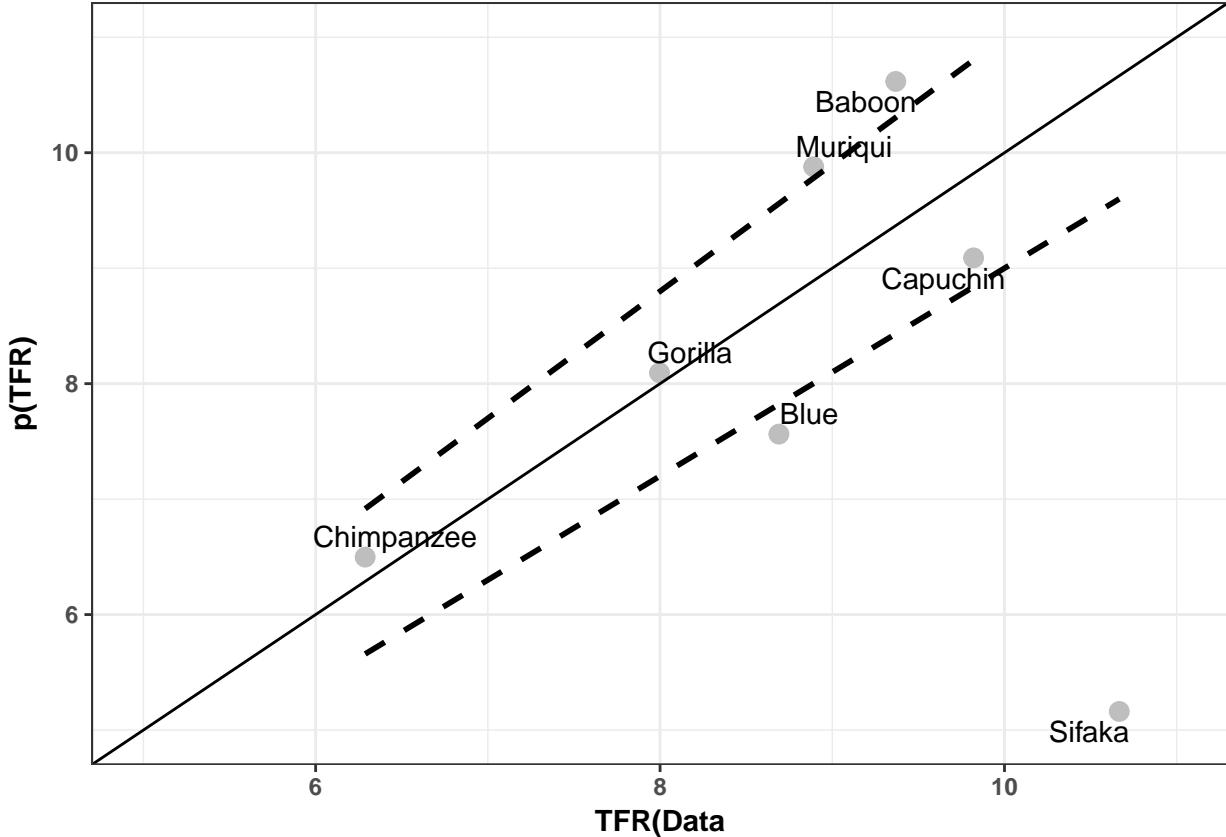


Figure 2: **iTFR using Primate Fertility Data.** We test the iTFR using data from seven wild primate species^{24,25}. These species exhibit markedly different menarche and menopause, length of fertility schedule, and fertility tempo from each other and from humans. We plot the observed TFR against the estimated TFR using the iTFR formulation. The solid line is $Y=X$, and the dashed lines represent $\pm 10\%$ of $Y=X$. Overall the iTFR performs remarkably well across the seven primate species, losing very little accuracy for species with markedly different fertility characteristics.

Table 1 reports the summary statistics for all three derivations. All three methods produce TFRs within approximately one-tenth of a child, with the best performing method producing estimates within one-twentieth of a child. Ninety percent of estimates are within one-quarter of a birth and have an average error rate less than 5%. This is significant improvement upon previous indirect estimation methods²⁰.

To test the generalizability of the method, we examine the accuracy of the implied fertility derivation in seven non-human primate populations (**Figure 2**). These populations vary substantially in population size and fertility patterns, allowing us to assess whether the

methods work across species. In contrast to humans with a fertility length of 35-40 years and menarche and menopause typically demarcated at 15-49 years old, menarche among the seven primate species ranged from a low of three years old with Sifakas (*Propithecus verreauxi*) to a high of 11 years old with Chimpanzees (*Pan troglodytes*), and fertility length ranged from a low of 22 years for Capuchins (*Cebus capucinus*) to a high of 45 years for Chimpanzees. These species display fertility intervals, fertility schedules, and TFRs that differ greatly from humans. We find that implied fertility accurately estimates total fertility among these species (**Figure 2**). Remarkably, this simple method with quite limited data requirements can accurately estimate total fertility across such diversity. Our results with nonhuman primates suggest the method captures fundamental patterns that govern fertility and that it can nearly universally be applied to a wide span of previously inestimable populations, time periods, geographies, and possibly species.

The three panel plots in (**Fig. 3**) demonstrate the flexibility of the method, depending on data availability, and how it greatly expands our ability to estimate fertility across geographies, time periods, and subpopulations. The National Center for Health Statistics does not publish fertility information for US Counties with populations fewer than 100,000 for privacy reasons. As a result, fertility can be estimated in only 500 of the approximately 3000 US Counties, significantly hindering the examination of sub-national fertility patterns. We use data collected from the 2010 US Census to produce robust estimates for all US counties using the xTFR method. We are then able to create the first complete county-level map of US fertility (**Fig. 3, a**). We also extended our analysis of the Human Mortality Database for three additional example countries: France, the Netherlands, and Sweden (**Fig. 3, b**). The Netherlands began collecting detailed birth records in 1950, France in 1946, and Sweden in 1891. However, these countries collected both mortality and age-sex data considerably earlier (Netherlands in 1850, France in 1816, and Sweden in 1751). By using the BayesTFR method, we can reconstruct historical TFRs to create a time series of fertility data well before birth record collection began, significantly improving our ability to explore historical

fertility patterns from up to 250 years ago. In (**Fig. 3, c**) we use the basic iTFR method to estimate TFRs by household income in the United States using data from the US Census Bureau's American Community Survey (ACS). The United States does not record fertility or mortality data by household income, precluding the use of the BayesTFR method. In addition, the ACS data have more statistical accuracy when estimates are pooled, precluding the use of the xTFR method, due to pooling women by age. However, by linking age-sex data to household income, we estimate TFR by income, a subpopulation previously unmeasured.

Researchers can tailor the method they use to the data available or the research question due to the flexible data requirements, a rare feature in methodologies, and can do so with confidence in the accuracy of the resultant estimates. As we show in **Fig 3.**, the applications of such a robust, simple method cannot be understated. The implied fertility rate opens the door for sophisticated fertility analyses in many previously inestimable populations of interest to sociologists⁹, economists²⁶, anthropologists²⁷, epidemiologists²⁸, and population geographers²⁹. The parameter-free, scale-, time-, and species- robust technique can estimate fertility rates even in areas where such data are not collected systematically as it relies only on age-sex data – ubiquitous basic data collected in censuses across scale and time.

We anticipate this method will open up new lines of inquiry into human fertility patterns. The global demographic transition is typically examined post-1950³⁰; however, using age-sex structure data, researchers can examine the global demographic transition in greater detail further back in time. Relatedly, re-creations of historic human fertility rates rely on estimates of energy balance and the relative metabolic load or a universal density-dependent demographic model³¹. Using the methods proposed here, anthropologists could re-create historic human fertility rates directly from skeletal remains or cemetery information while accounting for underenumeration and child-mortality. Finally, there has been an increased demand for high-resolution gridded population datasets related to climate change research^{32,33}. Our methods could be used to estimate fertility levels and change as either inputs for gridded

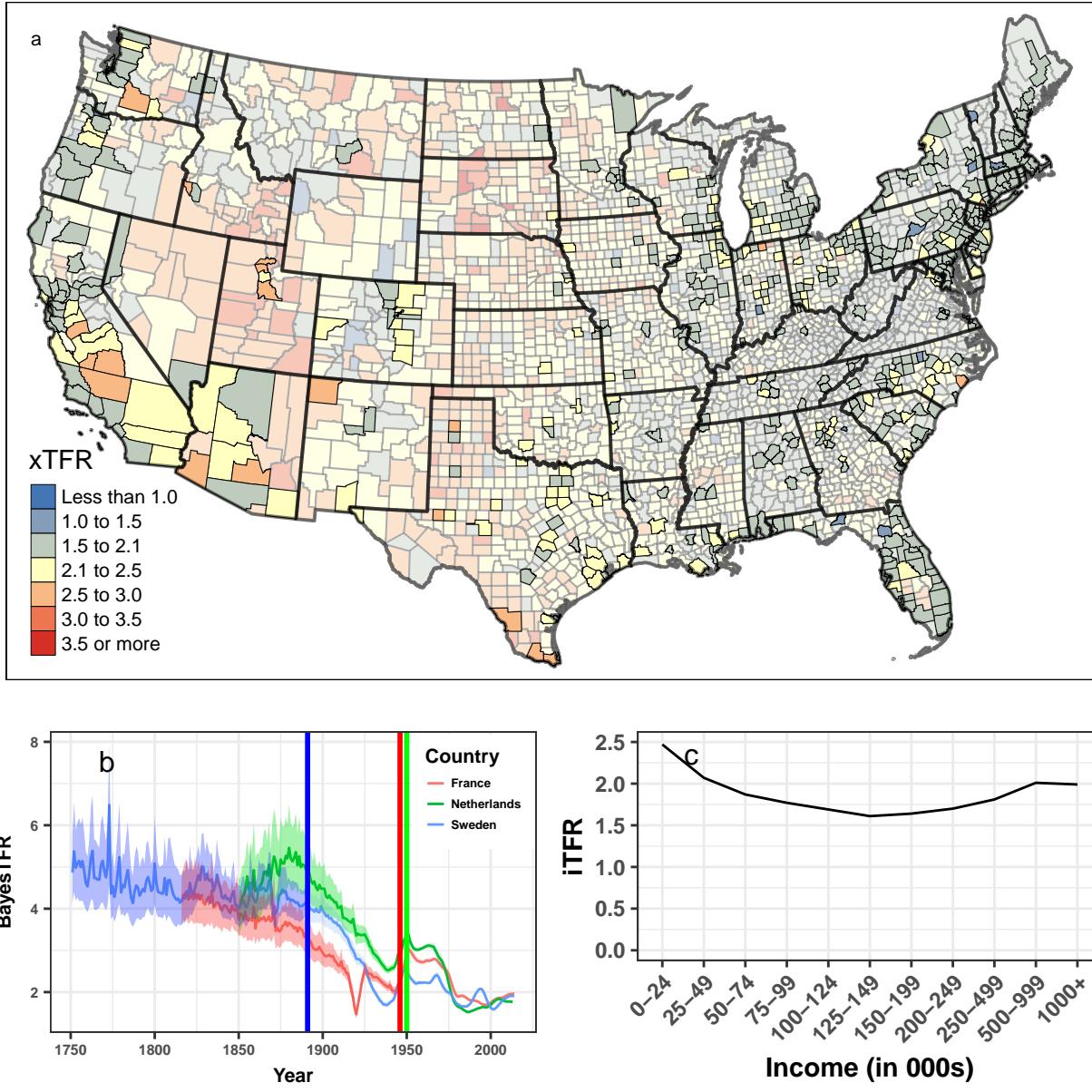


Figure 3: Revealing latent fertility patterns. We demonstrate that the iTFR framework can be used in a variety of situations. (a) uses the xTFR method to estimate total fertility rates in US Counties based on Census 2010 data. The National Center for Health Statistics masks fertility data for counties with fewer than 100,000 people for privacy reasons making fertility estimates possible for only 524 of the approximately 3000 US counties. The darker counties are the counties with actual fertility data. (b) uses the BayesTFR method to estimate historic fertility rates in three European countries using data from the Human Mortality Database. The vertical lines correspond to when birth registration began in each country. The shaded regions represent the 90th percentile. (c) uses the iTFR method on household income data from the Census Bureau’s American Community Survey. Birth registration is not recorded by income group, but we are able to estimate fertility rates using household survey data. These three examples represent previously inestimable populations by geography, time period, and sub-population and demonstrate the methodological flexibility of the iTFR framework.

population projections or for the production of gridded fertility level datasets due to the demonstrated scale independence.

Competing Interests

The authors declare that they have no competing financial interests.

Table 2: Human Fertility Database Countries and years of data availability.

Country	Data Availability	Country	Data Availability
Austria	1951-2014	Netherlands	1950-2012
Belarus	1964-2014	Norway	1967-2014
Bulgaria	1947-2009	Portugal	1940-2015
Canada	1921-2011	Russian Federation	1959-2014
Chile	1992-2005	Slovakia	1950-2009
Czech Republic	1950-2014	Slovenia	1983-2014
Estonia	1959-2013	Spain	1922-2014
Finland	1939-2015	Sweden	1891-2014
France	1946-2013	Switzerland	1932-2014
Germany	1990-2013	Taiwan	1976-2014
Hungary	1950-2014	Ukraine	1959-2013
Iceland	1960-2010	England & Wales	1938-2013
Italy	1954-2012	Scotland	1945-2013
Japan	1947-2014	Northern Ireland	1974-2013
Lithuania	1959-2013	United States of America	1933-2014

Supplementary Information

Data:

Human Fertility Database. Data on fertility patterns come from the Human Fertility Database²². These data represent the most complete and accurate historical pattern of human fertility currently available. The fertility rates are entirely based on official vital statistics and are not modelled. The HFD covers fertility schedules for 34 countries between 1891 and 2015. In total, the HFD contains 1,870 country-years of total fertility rates.

Human Mortality Database. Data on mortality rates come from the Human Mortality Database²³. These data represent the most complete and accurate historical pattern of human mortality currently available. Mortality rates come from official death counts from vital statistics, census counts, birth counts, and population estimates from varying sources. Q_x values and population counts used in this analysis were gathered from this data source.

Primate data. Primate age-specific fertility data come from Bronikowski et al^{24,25}. These data contain female age-specific fertility estimates for seven wild primates: sifaka (*Propithecus verreauxi*) in Madagascar, muriqui (*Brachyteles hypoxanthus*) in Brazil; ca-

puchin (*Cebus capucinus*) in Costa Rica; baboon (*Papio cynocephalus*) and blue monkey (*Cercopithecus mitis*) in Kenya; chimpanzee (*Pan troglodytes*) in Tanzania; and gorilla (*Gorilla beringei*) in Rwanda. The primate species were continuously monitored for at least 29 years. The fertility data are estimated in single-year intervals.

US County Fertility.

The implied Total Fertility Rate

We propose two derivations or extensions of the implied total fertility rate from the main text and describe the formulation in greater detail.

TFR is the expected number of children born over a complete reproductive lifetime at current age-specific rates. It is notated as

$$TFR = \int_{\alpha}^{\beta} f(x) dx \quad (2)$$

where $f(x)$ is the fertility rate (births per woman-year) at exact age x and $[\alpha, \beta]$ is the range of ages with non-zero fertility. In practice scholars approximate $f(x)$ with a step function that has a constant rate F_x within each n -year age interval $[x, x + n]$. F_x values are estimated as ratios of annual births to women in each age group (B_x) to their mid-year populations (W_x).

TFR is then calculated as

$$TFR = n \cdot \sum_{x=\alpha}^{\beta-n} \frac{B_x}{W_x} \quad (3)$$

Fertility calculations imply that the expected number of surviving children of both sexes per woman in reproductive age groups at the end of an n -year period, which we call K_a is

$$\begin{aligned}
K_a &= \left[\frac{L_{a-5}}{L_a} \cdot F_{a-5} + F_a \right] \frac{L_0}{2} \\
&= TFR \cdot \frac{L_0}{5} \cdot \frac{1}{2} \left(\frac{L_{a-5}}{L_a} \cdot \phi_{a-5} + \phi_a \right) \\
&= TFR \cdot s_0 \cdot p_a
\end{aligned} \tag{4}$$

where F_a is the average fertility rate over exact ages $[a, a + 5]$; ϕ_a is the fraction of total fertility occurring in age group a ($_5F_a/TFR$); L_a is the expected person-years lived in age group a , in a life table with a radix $l_0 = 1$; s_0 is the expected fraction still alive among children born in the past five years ($l_0/5$); W_a is the observed women in age group a ; and W is the total number of women at childbearing ages [15, 50].

We can decompose the right-hand side of K_a into three multiplicative factors: the total fertility rate, child survival, s_0 , and age-structure fertility, p_a .

If we examine the expected total number of surviving 0-4 year olds in a given population that contains women of childbearing ages 15-49:

$$C = \sum_{a=15}^{45} W_a K_a = TFR \cdot s_0 \cdot \left(\sum_{a=15}^{45} W_a p_a \right) \tag{5}$$

then the number of surviving children becomes a product of the fertility rates, child survival, and tempo fertility ($W_a p_a$) or the timing of childbearing.

We then further refine equation (5) to produce the child/woman ratio:

$$\frac{C}{W} = TFR \cdot s_0 \cdot \left(\sum_{a=15}^{45} \frac{W_a}{W} p_a \right) = TFR \cdot s_0 \cdot \bar{p} \tag{6}$$

where \bar{p} is the population-weighted average of lifetime fertility.

We can now rearrange equation (6) to solve for TFR such that

$$TFR = \frac{1}{s_0} \cdot \frac{1}{\bar{p}} \cdot \frac{C}{W} \quad (7)$$

If women of childbearing ages in a given population are uniformly distributed across the fertility interval, then $\bar{p} = n$, where n is the width of the fertility interval. If we assume near-zero child-mortality ($S_0 = 1$) and that women of childbearing ages are uniformly distributed across the fertility interval ($\bar{p} = n$). We can then further simplify equation (7) into the implied total fertility rate, or the fertility rate that is implied by the age-sex structure of a given population. Here we have broadened the equation to include all possible widths of the fertility interval.

$$iTFR = \frac{\beta - \alpha}{n} \cdot \frac{nC_0}{W} \quad (8)$$

This derivation is remarkably similar to the general fertility rate ($B_t/{}_40W_{10}$) and age-specific fertility (${}_nF_x = {}_nB_x/{}_nW_x$) with children substituted for births in both equations. This allows the iTFR to convert the C/W into ${}_nF_x$ and by extension, produce the indirect equivalent of the total fertility rate that is implied or intrinsic to the observed population age-sex structure freeing the estimation of TFR from its heavy data requirements.

xTFR

If women of reproductive age are uniformly distributed over the fertility interval, then the multiplier, n , will perfectly capture the width of the fertility interval. However, in practice, women are unlikely to be exactly uniformly distributed over the interval. We solved for the ideal multiplier in 1,859 fertility schedules in the HFD using women aged 15-49 and children aged 0-4 (ideal multiplier=7) and find multipliers ranging from a low of 4.52 (Taiwan 1986) to a high of 11.29 (France 1946). They are within 10% of the ideal multiplier (6.3-7.7) in 68.7% of the country-years. Ninety percent of all multipliers fall between 6.02 and 7.79.

However, fertility schedules typically follow a skewed distribution toward younger ages^{34,35} and demonstrate a common shape across human populations with most fertility occurring between 20 and 34. We observed a strong relationship between the proportion of women aged 25-34 and the multiplier (Figure S1., r-squared=0.388). Other age groups had considerably lower r-squared values (15-24 = 0.163, 20-29 = 0.3492, 25-34 = 0.388, 30-39 = 0.179, 35-44 = 0.108, 40-49 = 0.096). We then modify the iTFR to account for this new information:

$$xTFR = x \cdot \frac{C}{W} \quad (9)$$

$$x = \beta_0 + \beta_1 X_i$$

where X_i is the proportion of women aged 25-34 observed in population i . This formulation improves the estimates of total fertility (Fig. 1) but requires additional information about the distribution of women in the fertility interval.

We test a derivation of the xTFR in a multiple regression based on the proportion of women in each five-year age group,

$$xTFR = x \cdot \frac{C}{W} \quad (10)$$

$$x = \beta_0 + \beta_1 X_i + \beta_2 X_j \dots \beta_7 X_7$$

However, we find that using the population aged 25-34 produced almost identical estimates with minimal loss of accuracy (supplementary figure X).

Bayes TFR

We extend our previous approach into a Bayesian framework relating the number of children aged 0-4 to women aged 15-49. The iTFR formulation outlined in sup. mat. section 2 does not account for the potential error associated with any of the given parameters. The xTFR incorporates additional information from the age structure to improve the estimates

but does not account for infant mortality or possible estimation errors in the number of women of childbearing ages. s_0 and \bar{p} cannot not be truly known and could be subject to random errors in measurement.

We deconstruct the estimation of total fertility from equation (7) into TFR , C , W , s_0 , and \bar{p} producing a five-parameter model with four of the parameters deriving from the age-structure – C , W , \bar{p} , and TFR). The fifth parameter, s_0 , is not information embedded within the age structure and must be supplied in addition to the age structure. These parameters can be broadly categorized as fertility (TFR , \bar{p}), mortality (s_0), and age structure (W) with the result of these quantities being the expected number of surviving children (C).

Fertility Parameters

We decompose the fertility schedule for 5-year age groups into two components: level and shape

$$(F_{10}, F_{15}, \dots, F_{45}) = \frac{TFR}{5} \cdot (0, \phi_{15}, \dots, \phi_{45}) \quad (11)$$

where TFR is the total fertility rate and $\phi_{15\dots45}$ is the proportion of lifetime fertility that occurs in age group a . Fertility is negligible before age 15; thus, $F_{10} = 0$. The proportions $\phi_{15\dots45}$ are rewritten into indices, such that $\gamma_a = \ln(\frac{\phi_a}{\phi_{15}})$ for $a = 15\dots45$ such that $\phi_a(\gamma) = \frac{\exp(\gamma_a)}{\sum_z \exp(\gamma_z)}$ sum to one.

The γ indices are modelled as $\gamma = m + X\beta$ where m and X are constants derived from empirical data (see section X.X) and β are shape parameters. These three fertility parameters (TFR , β_1 , β_2) yield eight 5-year fertility rates ($F_{10}\dots F_{45} : \beta \rightarrow \gamma \rightarrow \phi$ and $\frac{TFR}{5} \cdot \phi = F$).

We use a completely uninformative, improper prior for TFR : $f(TFR) \propto 1$. We assign higher probability to more typical fertility patterns by building the prior for β coefficients from information in the HFD and the US Census Bureau's International Database³⁶ to create priors of the shape of the fertility schedule by age. We calculate γ indices for F_a empirical schedules (n=226 from the IDB, n=411 from the HFD), and then performed a singular value decomposition on the (de-meaned) 6x637 γ array. This produces a model in which each of the

637 columns of γ could be well approximated by the mean vector plus a weighted sum of two principal components: $\gamma_i \approx m + X\beta_i$. We scale the two columns of X so that β_i coefficients have zero means, unit variances, and zero covariances over the empirical data $i = 1 \dots 637$.

These calculations produce constants $X = \begin{matrix} 0 & 0.27 & 0.54 & 0.73 & 0.88 & 1.04 & 1.52 \\ 0 & 0.32 & 0.51 & 0.51 & 0.35 & 0.05 & -0.72 \end{matrix}'$, with

which we use the prior

$$\beta \sim N(0, I_2) \quad (12)$$

with support restricted to the range [-2,+2] for each β coefficient, in order to better mimic the HFD distributions. When we examine the X matrix, we find that β_1 affects the mean age of childbearing and β_2 affects the variance: If β_1 is higher, fertility is postponed; if β_2 is higher, fertility is concentrated in fewer age groups.

Mortality Parameters

We model child and adult mortality with a two-parameter relational mortality model³⁷. The probability of death before age five (q_5) and a shape parameter k with typical values between -2 and +2 index the mortality schedule. The model uses fixed constants $\{a_x, b_x, c_x, v_x\}$ derived from mortality schedules in the Human Mortality Database:

$$\ln \mu_x(q_5, k) = a_x + b_x [\ln q_5] + c_x [\ln q_5]^2 + v_x k \quad , \quad x = 0, 1, 5, 10 \dots 45 \quad (13)$$

Mortality rates μ_0 and μ_1 refer to age intervals $[0, 1)$ and $[1, 5)$, μ_x refers to 5-year age intervals $[x, x + 5)$ for all other age groups. $q_5 = 1 - l_5$ is a model parameter, meaning there are no $\{a_x, b_x, c_x, v_x\}$ constants for calculating μ_1 ; instead, μ_1 is calculated by $\mu_1 = -\frac{1}{4} [\mu_0 + \ln(1 - q_5)]$.

We convert log mortality rates into life table person-years, L_a , for 5-year age intervals $[a, a + 5)$ using standard demographic approximations. Survival probabilities to exact ages are $l_0 = 1$, $l_1 = e^{-\mu_0}$, $l_5 = l_1 \cdot e^{-4\mu_1}$, and $l_x = l_{x-5} \cdot e^{-5\mu_{x-5}}$ for $x = 10 \dots 45$. Life table

person-years are $L_0 = \frac{1}{2}(l_0 + l_1) + \frac{4}{2}(l_1 + l_5)$ and $L_a = \frac{5}{2}(l_a + l_{a+5})$ for $a = 5 \dots 45$. Thus the two mortality parameters (q_5, k) yield 10 L_a values $(L_0, L_5 \dots L_{45})$.

To account for the uncertainty in the mortality estimates, we use a beta distribution of q_5 (denoted as \hat{q}_5). The prior is defined as

$$q_5 \sim Beta[a(\hat{q}_5), b(\hat{q}_5)] \quad (14)$$

where $a(\hat{q}_5)$ and $b(\hat{q}_5)$ are chosen so that $P[q_5 < \frac{1}{2} \min(\hat{q}_5)] = P[q_5 > 2 \max(\hat{q}_5)] = .05$. This assigns a 90% prior probability that under-five mortality q_5 is between one-half and twice the observed q_5 value.

For the shape parameter k , we use the prior

$$k \sim N(0, 1) \quad (15)$$

which centers the distribution at zero and has a low probability of falling out of the $[-2, +2]$ range. We assume that mortality parameters q_5 and k are independent.

The values for parameters (TFR, β, q_5, k) imply specific values K_a in equation 4. The expected number of children to the W_a women observed in age group a is $W_a K_a$, and the observed number of surviving children may be modeled as $C_a \sim Poisson(W_a K_a)$. It is reasonable to assume that C_a values are statistically independent, conditional on fertility and mortality rates, so that their sum $C = \sum_a C_a$ is also a Poisson random variable. Thus,

$$C|W, TFR, \beta, q_5, k \sim Poisson \left[\sum_{a=15}^{45} W_a K_a(TFR, \beta, q_5, k) \right] \quad (16)$$

Posterior Distribution of TFR

The posterior for parameters conditional on data is

$$P(TFR, \beta, q_5, k|C, W) \propto L(C|W, TFR, \beta, q_5, k) f_\beta(\beta) f_q(q_5) f_k(k) \quad (17)$$

where the likelihood on the right-hand side is the Poisson likelihood in (16), and the f functions represent the prior densities implied by (12), (14), and (15), respectively. Note that the improper flat prior for TFR does not affect the posterior distribution.

The marginal posterior for TFR , which expresses the relative probabilities of alternative fertility levels given the number of children C and the counts of women $W_{15} \dots W_{45}$, is

$$P(TFR|C, W) \propto \int L(C|W, TFR, \beta, q_5, k) f_\beta(\beta) f_q(q_5) f_k(k) d\beta dq_5 dk \quad (18)$$

In practice, we sample from the full posterior distribution in (17) by applying Markov Chain Monte Carlo (MCMC) methods. Specifically, we program the model in the *Stan* MCMC language³⁸, as implemented in the *rstan* package in *R*^{39,40}. We use the empirical density of the sampled TFR values to estimate the marginal posterior of TFR in (18).

References

1. Abel, G. J. and Sander, N. *Science* **343**(6178), 1520–1522 (2014).
2. Simini, F., Gonzalez, M. C., Maritan, A., and Barabasi, A.-L. *Nature* **484**(7392), 96–100 (2012). 10.1038/nature10856.
3. Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Mackenbach, J. P., van Lenthe, F. J., Mokdad, A. H., and Murray, C. J. *JAMA Internal Medicine* (2017).
4. Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., and Lalic, N. *Science* **346**(6206), 234–237 (2014).
5. Abel, G. J., Barakat, B., KC, S., and Lutz, W. *Proceedings of the National Academy of Sciences* **113**(50), 14294–14299 (2016).
6. Yule, G. U. *Journal of the Royal Statistical Society* **69**(1), 88–147 (1906).
7. Lutz, W. and KC, S. *Science* **333**(6042), 587–592 (2011).
8. Shenk, M. K., Towner, M. C., Kress, H. C., and Alam, N. *Proceedings of the National Academy of Sciences* **110**(20), 8045–8050 (2013).
9. Lesthaeghe, R. *Proceedings of the National Academy of Sciences* **111**(51), 18112–18115 (2014).
10. Myrskyla, M., Kohler, H.-P., and Billari, F. C. *Nature* **460**(7256), 741–743 (2009). 10.1038/nature08230.
11. Lee, R. and Mason, A. *Science* **346**(6206), 229–234 (2014).
12. Kaiser, J. *Science* **333**(6042), 548–549 (2011).
13. Conley, D., Laidley, T., Belsky, D. W., Fletcher, J. M., Boardman, J. D., and Domingue, B. W. *Proceedings of the National Academy of Sciences* **113**(24), 6647–6652 (2016).
14. Jones, J. H. and Tuljapurkar, S. *Proceedings of the National Academy of Sciences* **112**(29), 8982–8986 (2015).
15. Currie, J. and Schwandt, H. *Proceedings of the National Academy of Sciences* **111**(41), 14734–14739 (2014).
16. Bogue, D. and Palmore, J. *Demography* **1**(1), 316–338 (1964).
17. Hanson, W. R. *Wildlife Monographs* (9), 3–60 (1963).
18. Rele, J. *Fertility analysis through extension of stable population concepts*. Institute of International Studies, University of California, (1967).
19. Tuchfeld, B. S., Guess, L. L., and Hastings, D. W. *Demography* **11**(2), 195–205 (1974).

20. Hauer, M., Baker, J., and Brown, W. *PloS one* **8**(6), e67226 (2013).
21. Schmertmann, C. and Hauer, M. (2017).
22. Database, H. F. , Available at www.fertilitydata.org (accessed on 1 May 2017).
23. Database, H. M. , Available at www.mortality.org or www.humanmortality.de, (data downloaded on 1 May 2017).
24. Bronikowski, A. M., Cords, M., Alberts, S. C., Altmann, J., Brockman, D. K., Fedigan, L. M., Pusey, A., Stoinski, T., Strier, K. B., and Morris, W. F. *Scientific data* **3** (2016).
25. Bronikowski, A. M., Cords, M., Alberts, S. C., Altmann, J., Brockman, D. K., Fedigan, L. M., Pusey, A., Stoinski, T., Strier, K. B., and Morris, W. F. 03/01/ (2016). *Scientific Data*.
26. Hotz, V. J., Klerman, J. A., and Willis, R. J. *Handbook of population and family economics* **1**, 275–347 (1997).
27. Greenhalgh, S. *Situating fertility: Anthropology and demographic inquiry*. Cambridge University Press, (1995).
28. Paulson, R. J., Boostanfar, R., Saadat, P., Mor, E., Tourgeman, D. E., Slater, C. C., Francis, M. M., and Jain, J. K. *Jama* **288**(18), 2320–2323 (2002).
29. Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. *Scientific data* **2**, 150045 (2015).
30. Bongaarts, J. *Science* **282**(5388), 419–420 (1998).
31. Bocquet-Appel, J.-P. *Science* **333**(6042), 560–561 (2011).
32. Jones, B., Oneill, B. C., McDaniel, L., McGinnis, S., Mearns, L. O., and Tebaldi, C. *Nature Clim. Change advance online publication* (2015). [rsquor].
33. Cincotta, R. P., Wisnewski, J., and Engelman, R. *Nature* **404**(6781), 990–992 (2000). 10.1038/35010105.
34. Coale, A. J. and Trussell, T. J. *Population index* , 185–258 (1974).
35. Schmertmann, C. *Demographic Research* **9**, 81–110 (2003).
36. United States Census Bureau. (2016).
37. Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., and Sawyer, C. *Population studies* **66**(1), 1–28 (2012).
38. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. *Journal of Statistical Software* **76**(1), 1–32 (2017).

39. Stan Development Team. (2016). R package version 2.14.1.
40. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2016).