Boom, echo, pulse, flow

Tim Riffe*[1], Kieron Barclay[1], Christina Bohk-Ewald[1], and Sebastian Klüesener[2]

[1]Max Planck Institute for Demographic Research
[2]Bundesinstitut für Bevölkerungsforschung

August 30, 2018

Abstract

Human population renewal starts with births. Since births can happen at any time in the year and over a wide range of ages, demographers typically imagine the birth series as a continuous flow. Taking this construct literally, we visualize the birth series as a flow. A long birth series allows us to juxtapose the children born in a particular year with the children that they in turn had over the course of their lives, yielding a crude notion of cohort replacement. Macro patterns in generational growth define the meandering path of the flow, while temporal booms and busts echo through the flow with the regularity of a pulse.

1 Introduction

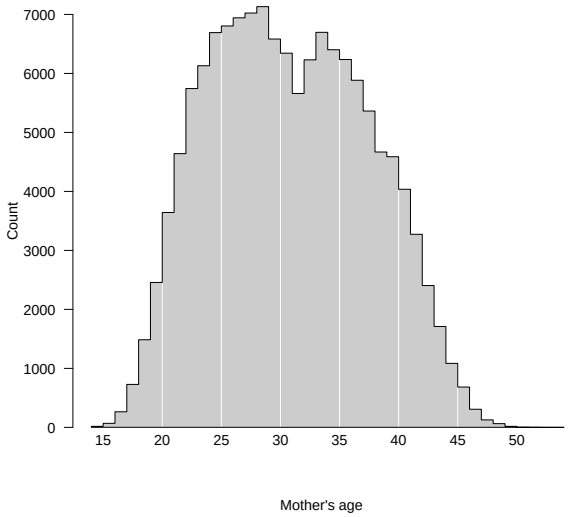
Usually we think of fertiltiy as an age-regulated process. In any case it is bounded by menarche and menopause, both of which are anchored to age. These anchors may move, but not far or fast. And between these bounds, at least within acceptably homogenous subpopulations, fertility patterns appear to conform to some regular schema, best captured by fertility rates. In this treatment, we retreat from rates, the material of projections, to babies, the raw material of population renewal. We use birth count data from the Human Fertility Database. (2017) for Sweden, covering a total of [TODO: 241] ocurrence years. These data include preliminary data for the period 1775 to 1890, which we have graduated and adjusted. We describe those steps in Appendix A. A picture of the births in a year is for demographers most instinctively broken down by the age of mothers who gave birth in that year, Fig. 1a, or by the year of birth of mothers Fig. 1b. These two distributions are essentially identical, but appear as mirror images if chronological time is enforced in x.

If one disposes of a long-enough time series of births classified by mothers' year of birth, then one may further examine and break down the full reproductive career of the cohort of individuals born in a particular year. Since the childbearing of a cohort is spread over a synchronous span of ages and years, the classification by age (Fig. 2a) or year (Fig. 2b) yields identical and redundant distributions.

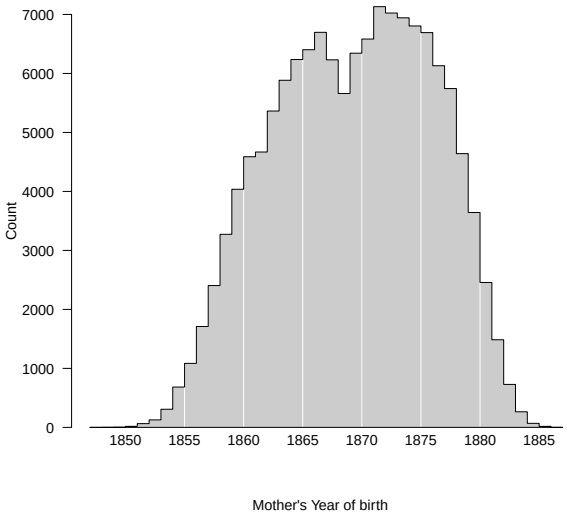
The births in a year are classified by mothers' cohort, i.e. cohort *origins* in Fig. 1b, whereas the births *from* a cohort are classified *to* time in Fig. 2b. The two distributions are different in kind, but relatable and both on a common scale. A fuller representation of their relationship would place them as two disjoint distributions on the same timeline, as in Fig 3.

The two distributions in Fig. 3 are related, and of comparable scale, but different in kind. The x coordinate of the left distribution is indexed to mothers's birth cohort, whereas the x coordinate of the right distribution is indexed to child cohort, ocurrence year. In this way the x coordinates belong to grandmothers and grandchildren, where the *ego* generation is 1900. These are two quantities that we may wish to compare in various ways to get a better feel and understanding of the Swedish birth series.

*riffe@demogr.mpg.de

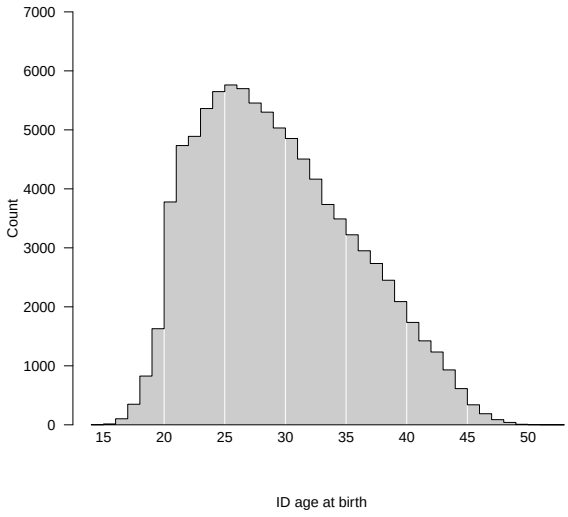


(a) Births in 1900 by age of mother

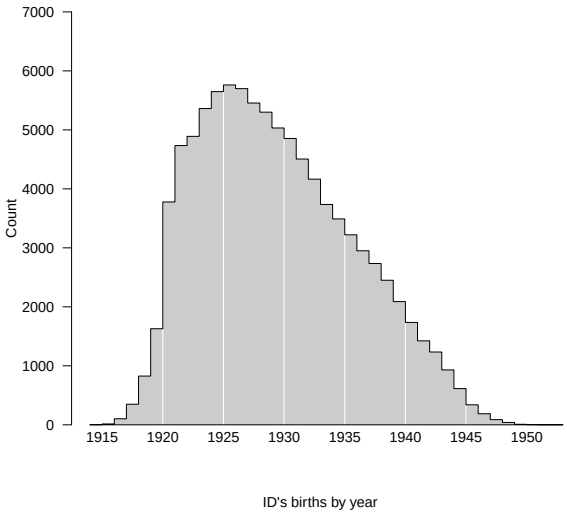


(b) Births in 1900 by year of birth of mother

Figure 1: Births in a year structured by mothers' age versus mothers' year of birth are a reflection over y and shift over x.



(a) Births from mothers born in 1900 by age of mother



(b) Births from mothers born in 1900 by year

Figure 2: Births of a cohort structured by mothers' age versus mothers' year of birth are a reflection over y and shift over x.

For the case of these Swedish data, we have 241 such distribution pairs, making single-axis rendering impractical. An honest attempt might look like Fig. 4, where we reflect the Fig. 3 left distribution over y (**A**), keeping the Fig. 3 right-side distribution on top (**B**). These two distributions are linked by the year 1900, which of course overlaps with neither of them. In this representation, **A** and **B** are re-drawn for each possible ego year (1775-2016), and therefore imply a large sequential set of overlapping distributions. Each 20th distribution is highlighted, but despite attempts to make this graph legible, i) the high degree of overlapping and ii) the spatial dissociation of each **A** — **B** pair makes the intended comparison difficult

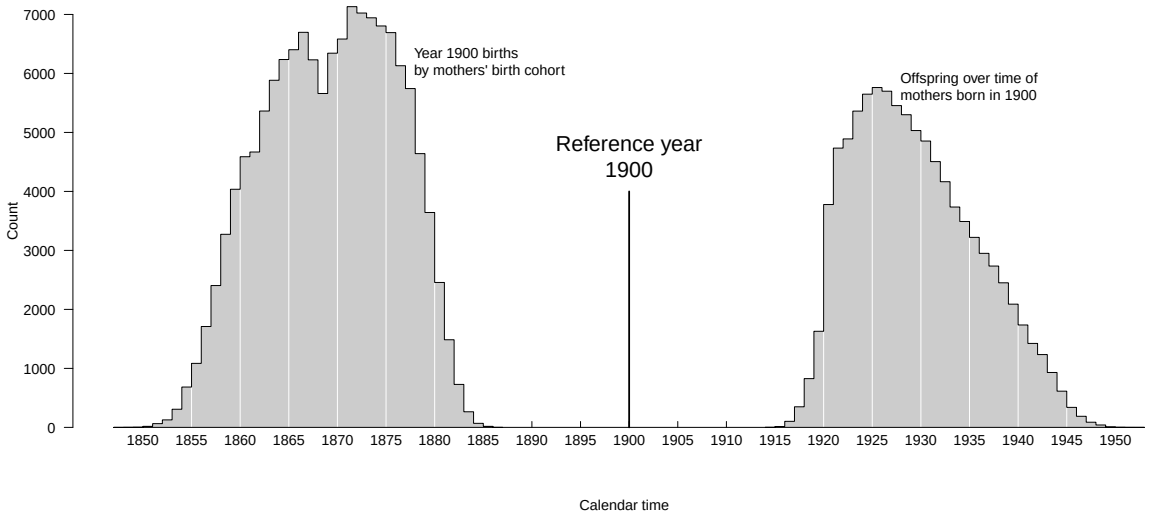


Figure 3: Births from mothers born in 1900 by year

over the series.

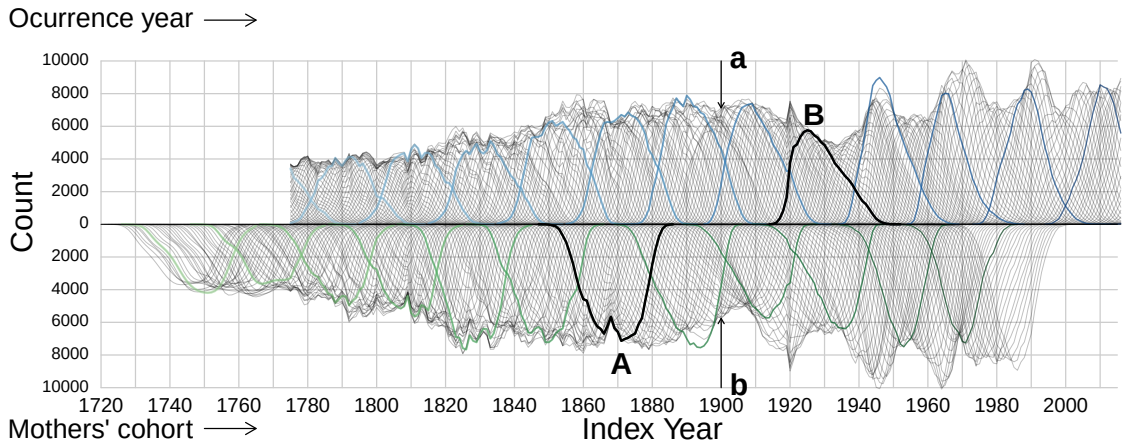


Figure 4: Two time series of birth count distributions. The top series is composed of offspring distributions of mother cohorts over time, indexed to ocurrence years. The bottom series is composed of the offspring of a year indexed to mothers' birth cohorts. **B** is the offspring of mothers from the 1900 cohort indexed in x to ocurrence year, and **A** are the births ocurred in 1900 indexed in x to mothers' birth cohorts. The cross-section **a** gives **A** and the cross-section **b** gives **B**.

Fig. 4 produces at least two noteworthy artifacts that we may wish to preserve or clarify. 1) First order differences in the top series appear to cascade into the lower series— This merely points out that larger cohorts have more offspring than smaller neighboring cohorts and vice versa, sudden fertility rate changes notwithstanding. 2) The composition of **A** in the bottom series is implied by the cross-section **a** of the top series, and the composition of **B** is implied by the cross-section **b**. To reiterate: **A** are all the births in 1900 indexed back to mothers' cohorts. Each possible *slice* of **A** comes from a different top distribution as it crosses the year 1900 (and vice versa for the bottom). **a** and **b** are in a sense already juxtaposed for us, as they share an x coordinate. The "problem" with the cross-sections **a** (**b**) is that each *slice* of the corrsponding distribution **A** (**B**) is perfectly overlapped, such that it is just about impossible to imagine what **A** might look like if presented only with **a** and its surroundings.

In this way the two distributions that we might wish to compare for a given ego year are already available at a like coordinate, but comparison is stifled by overplotting. If instead we stack the slices that are indecipherably overlapped in **a** (and likewise for **b**) we get something like that shown in Fig. 5, cumulative birth distributions.[1] Here the total bar length is proportional to the total cohort (offspring) size, and stacked bins reflect 5-year mother cohorts (ocurrence years). From this representation it is clear that mothers born in the 20 years between 1860 and 1880 produced the bulk of the 1900 cohort (86%), which itself produced the majority of its offspring in the 20 years between 1920 and 1940 (90%). It is also quite visible that the 1900 cohort did not replace itself in a crude sense: 138,139 babies formed a cohort whose mothers gave birth to 95,379 babies over their life-course, a crude replacement of 69%. Other perspectives on reproduction that account for survival and attrition of the mother cohort through migration would give a more optimistic assessment. The key feature of Fig. 5 is that the two distributions that were disjoint in Fig. 3 and hard to pick out in Fig. 4 can now be associated at a common x coordinate. This virtue allows us to view the time series of Fig. 4 with greater clarity and perhaps reveal some macro properties of the history of Swedish natality.

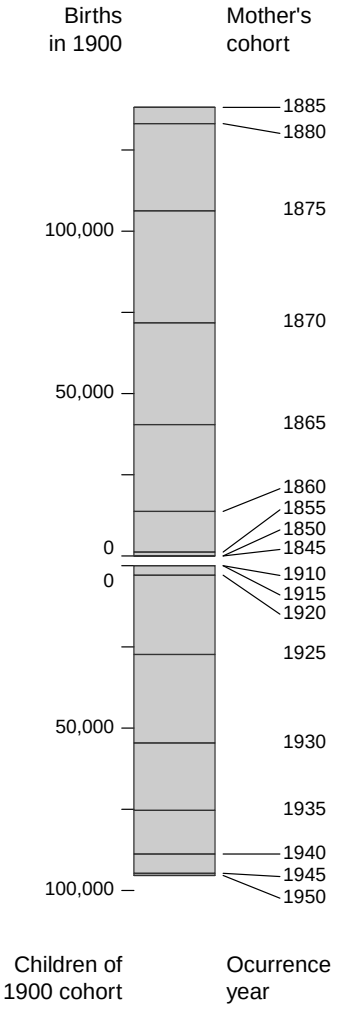


Figure 5: The 1900 cohort as a composite bar with its offspring reflected over y. The size of each bar stacked in the top composition is proportionate to the area of its corresponding polygon in the left distribution of Fig. 3. The size of each bar stacked in the lower composition is proportionate to the area of its corresponding polygon in the right distribution of Fig. 3.

Fig. 6 is a depiction of the exercise of Fig. 5, cohort bars on top reflected with offspring bars on the bottom. Equal bounded bins from Fig. 5 are joined into continuous regions. For the top region, filled polygons represent the births of mothers from quinquennial cohorts, spread over time. For the bottom region, filled polygons represent the mother-cohort origins of the births in quinquennial periods. The darkness and saturation of polygon fill colors are approximately proportional to the total births in the polygon (and therefore true to grayscale printing), whereas hue is irrelevant. In this way, darkness and saturation on the top are proportional to total height (with respect to baseline) on the bottom, and vice versa.

The meandering baseline of Fig. 6 is proportional to a smoothed time series of the crude cohort replacement rate. We overlay a horizontal line to indicate periods of approximate growth, replacement, and contraction. Periods where the meandering x-baseline is above this line (ca 1780 to 1860) indicate crude growth, and periods below the horizontal reference line (ca 1870 to 1930) indicate crude generation contraction.

[1] Young mothers are on top and older mothers on bottom for both distributions. It would also make sense to plot increasing (or decreasing) ages eminating out from the centerline in both directions.

[fold-out figure 4×a4 paper size at 100% in separate pdf, about here.]

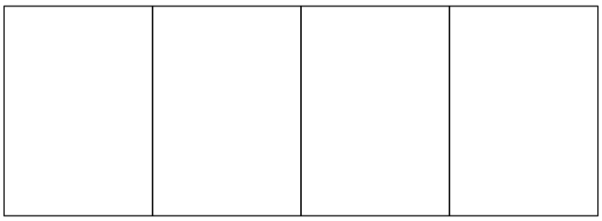


Figure 6: A time series of the same graphical construct as presented in Fig. 5. The x axis now meanders proportional to a smoothed time series of the crude cohort replacement rate. Fill color darkness and saturation are approximately proportional to the total number of births in each birth distribution. The birth series now appears as a flow, but reveals echoes in cohort and offspring size, an odd periodicity in recent decades, and a long term dampening of the crude replacement rate. A 5-generation female lineage is annotated atop to serve as a guide.

To aid the viewer with interpretation, we overlay a known lineage of five female generations,[2] where x position is exact to the year, y position in the top region is matched to the mothers cohort, and y position in the bottom is matched to daughters' year of birth. Wider horizontal spacing between generations over time indicates increasing ages at maternity within this lineage (increasing from 23 to 39).

2 Discussion

[TODO: To be continued...] Several macro features come to the fore in this visualization. These are either known features of the Swedish birth series, or else merit further study. Echoes, booms, why is boom periodicity a recent phenomenon? Is there a dose-response to vertical reverberation in first derivative (can this be referred to as first or second order?) features.

3 Analytical perspectives

[TODO: this sort of thing ought to inspire discussion, but the stuff presently in this section may not belong in the paper.] A few macro patterns can be extracted from the data structure implied by the matrix $\mathbf{B}(c,t)$.

3.1 Analytical vignette 1

An example of macro patterns that can be extracted from this data structure include two-distribution location distance statistics, as demonstrated in Fig. 7. For each reference year in the range 1775 to [TODO: 1970] we have the mothers' cohort distribution and the next-generation childbirth year of ocurrence distribution. Since both distributions are derived from the same table of counts by year of ocurrence and mothers' age, it will help to use a simple notation, where the reference year is denoted with r, and following the index position convention $B(c,t)$ where mothers' cohort c takes the first and year of ocurrence t the second position, respectively. In this way $B(c,r)$ are the births in year r to mothers from cohort c and $B(r,t)$ are the births in year t to mothers from cohort c (offspring). In this way, the weighted mean intergenerational lag $\overline{m2}$ is defined as

$$\overline{m2}(r) = \frac{\sum\sum(B(c,r) * B(r,t)) * (c-t)}{\sum\sum B(c,r) * B(r,t)} \quad , \tag{1}$$

[2] This lineage can be located in the public domain on `https://www.geni.com/people/Karin-Ottolina-Landsten/6000000022470480183`.

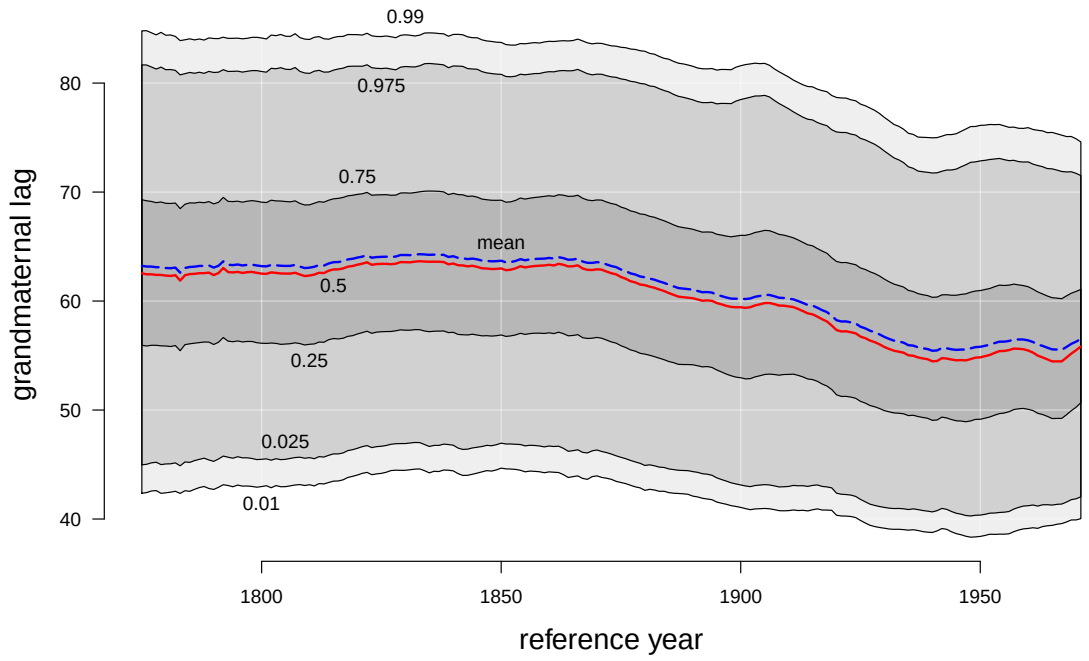


Figure 7: Time lag from mothers' cohort to next generation offspring year of birth, the *grandmaternal lag*, referenced to central (ego) cohorts. Decimals indicate birth distribution quantiles, where the red line indicates the median, and the blue dashed line the mean. The lag decreased broadly by all measures by ca 7-8 years in the 100 ego years from 1850 to 1950. The interquarile, 95% and 99% ranges also compressed by about 1 year in the same period.

and like-weighted distance quantiles may also be derived, as displayed in Fig. 7.

From Fig. 7 we learn that the mean two-generation maternal lag decreased in the mean and all quantiles in the 100 years from 1850 to 1950 by around 7 to 8 years. The interquartile, 95% and 99% spreads all decreased by about 1 year over the same period, and by more than another year in the following 20 years.

3.2 Analytical vignette 2

One of the most immediately visible features of Fig. 6 is the propagation of first differences in $B(t)$ to $B(c)$. The 1920 cohort is a particularly visible example: There were 23560 more births in 1920 than in 1919, an increase of 20.4%, and mothers from the 1920 cohort also gave birth to 20.7% more babies than the 1919 cohort. Fig. 8 displays the relationship in proportional first differences between matched birth cohort and offspring size. For the most part, the size of such structural echoes is maintained 1:1 in cohort offspring.

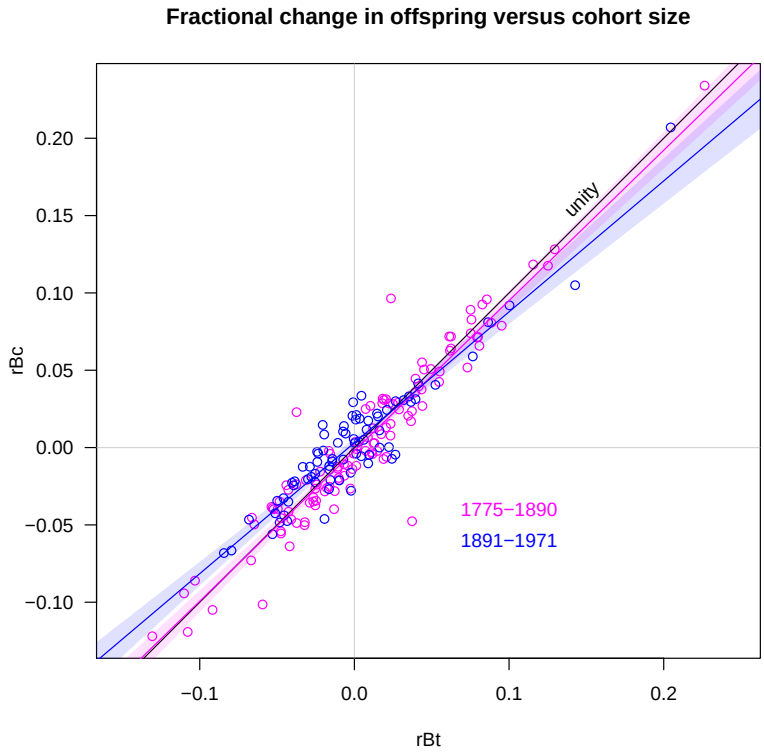


Figure 8: A roughly 1:1 does-response relationship in relative size of structural echo.

References

Joop de Beer. A time series model for cohort data. *Journal of the American Statistical Association*, 80 (391):525–530, 1985.

Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). online, 2017. Available at www.humanfertility.org (data downloaded on [June, 2017]).

Aiva Jasilioniene, DA Jdanov, T Sobotka, EM Andreev, K Zeman, VM Shkolnikov, JR Goldstein, D Philipov, and G Rodriguez. Methods protocol for the human fertility database. Technical report, Max-Planck-Institude for Demographic Research and the Vienna Institute for Demography, 2015. URL `https://www.humanfertility.org/Docs/methods.pdf`.

Marius D. Pascariu, Silvia Rizzi, and Maciej Danko. *ungroup: Penalized Composite Link Model for Efficient Estimation of Smooth Distributions from Coarsely Binned Data*, 2017. URL `https://github.com/mpascariu/pclm`. R package version 0.8.3.

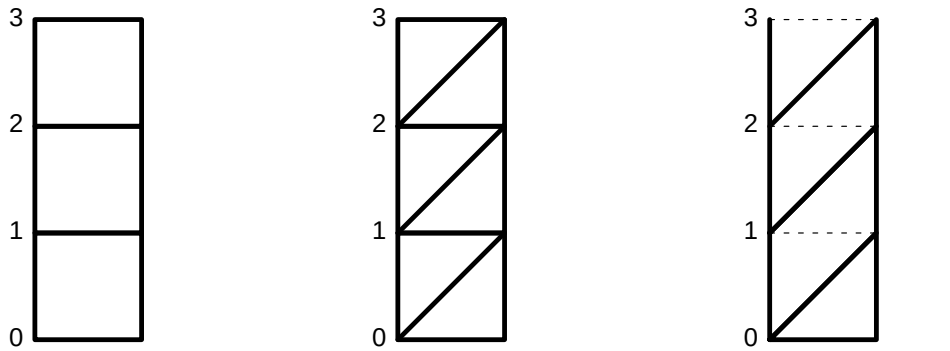
Silvia Rizzi, Jutta Gampe, and Paul HC Eilers. Efficient estimation of smooth distributions from coarsely grouped data. *American journal of epidemiology*, 182(2):138–147, 2015.

A Data sources and adjustments

Data presented here are from three separate series. The first contains birth counts in the period-cohort Lexis shape, `SWEbirthsVV.txt`, as produced by the Human Fertility Database. (2017) according to the Methods Protocol (Jasilioniene et al. 2015). This file contains births by calendar year and mother birth cohort for the years 1891 until 2016, and we use it as-is. A second file contains births for ocurrence years 1775 until 1890. These data are age-period classified, and given in a mixture of age classes, with a predominance 5-year age classes (especially for ages 20-50), but also sometimes single ages (especially for ages 15-19), and time-varying top and bottom open ages. A third time series derives from a projection of cohort fertility for the cohorts born [TODO: 1970-2016] [Check years].

A.1 Adjustments to historical data

It is this second file, with data covering years 1890 and earlier, that we have adjusted in four main steps. First, births of unknown maternal age were redistributed proportionally to the distribution of births of known maternal age. Second, counts were graduated to single ages using the graduation method proposed by Rizzi et al. (2015) and implemented in `R` in the package `pclm`[3]. Third, counts were shifted into period-cohort Lexis bins assuming that half of the births in each single age x bin go to the lower triangle of age $x+1$ and half to the upper triangle of the age-reached-during-the-year (PC) parallelogram at age x, as diagramed in Fig. 9.



(a) AP square bins (b) Split evenly to triangles (c) Regroup to PC bins

Figure 9: The count regrouping procedure for years 1776 to 1890, step three of data adjustment. Data are graduated to single ages (Fig. 9a), then split in half (Fig. 9b) and regrouped to period cohort (PC) bins (Fig. 9c).

At this stage data are binned and Lexis-conformable with HFD data for years 1891 and forward. With data processed as of step three, one could produce two time series represented in Fig. 6, with a subtle artifact visible in Fig. 10. In area **A** of this figure, birth counts in age bins have been graduated using the previously mentioned pclm method, which has the usually-desired artifact of smoothness. For the affected range of years, mother cohorts are identified via the identity $C = P - A - 1$.[4] Since age patterns of counts are smooth, these sum in Lexis diagonals to a smooth time series of cohort total offspring, as seen in the profile of area **B** of the same figure. Area **C** of this figure delimits years 1876 until 1971, where both cohort and matched offspring sizes are directly observed, and where fluctuations would appear to covary quite strongly. In the first instance for reasons of aesthetic continuity, and in the second instance for the sake of a more sensible count graduation, we have opted to adjust the counts in area **B** to carry the pattern of fluctuation observed over cohort size from 1775 to 1890.

This adjustment works by extracting the fluctuation pattern from **A** and transfering it to **B**. We do this by first smoothing the annual time series of total cohort size $B(t)$ according to some smoothness parameter, λ.[5] The ratio of $B(t)$ to the smoothed birth series $B(t)^s$ defines the multiplicative adjustment factor, $adj(t) = B(t)/B(t)^s$. Total offspring size $B(c)$ is then adjusted as $B(c)' = adj(t) * B(c)$, for $c = t$.

[3] The `pclm` package has since been extensively modified, and it is now called `ungroup` (Pascariu et al. 2017)

[4] One subtracts 1 because data are in period-cohort bins.

[5] For the present case we've used a loess smoother, using the `R` function `loess()` with smoothing parameter $\lambda =$ `span`. It would be straightforward to swap this smoothing method out with a different one.

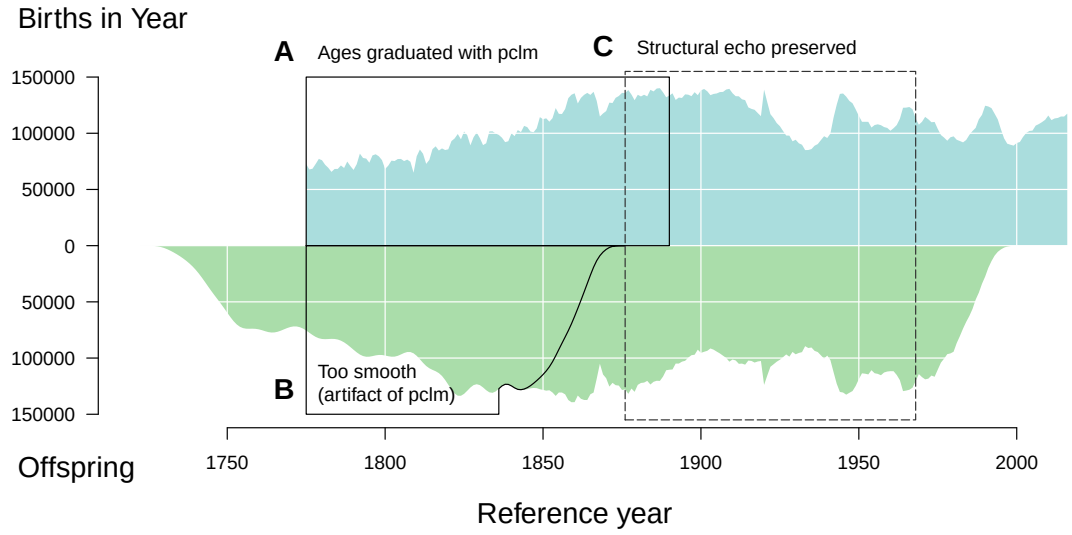


Figure 10: In reference years $\geq$ 1891 both births by year and cohort offspring are directy observed in single year bins, which means that the structural echo between total birth cohort and offspring size is preserved for reference years $\geq$ 1876 (**C**). Total per annum births in years $\leq$ 1890 (**A**) are presumed accurate, and so first differences of these are observed. Offspring from cohorts born in years $\leq$ 1876 (**B**) were partially (1836–1876) or entirely ($<$ 1836) born in years $\leq$ 1890, implying a smooth redistribution over single years of mother cohorts. We wish to adjust the births in **B** to recuperate the kind of structural echo in **C**.

Counts in single ages are then rescaled to sum to the original totals in 5-year age groups, and counts for years $>$ 1890 are unaffected. The smoothing parameter is selected such that the linear relationship in fractional first differences $rd(B(t)) = \frac{B(t+1)-B(t)}{B(t)}$ between the annual birth series and adjusted offspring series $rd(B(c)')$ for years 1775-1890 matches that for the reference years 1877-1971 as closely as possible. Specifically, we select λ so as to minimize the sum of the difference in the slope and residual standard deviation for the periods before and after 1891. Further clarifications about this adjustment, and code for diagnostic plots can be found in the annotated code repository. The end effect is to adjust the series to look like Fig. 11.

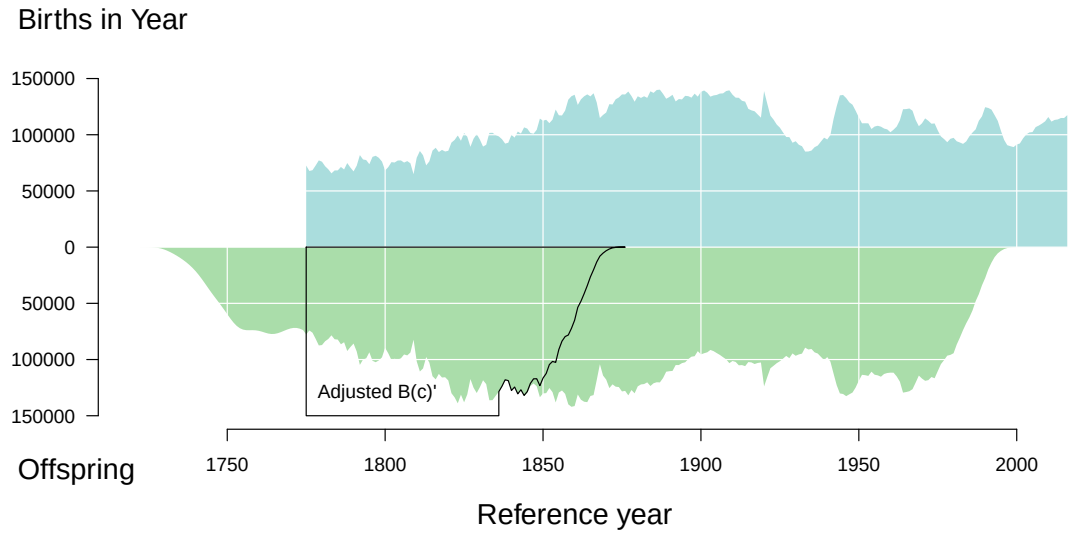


Figure 11: The adjusted birth series. Annual total births $B(t)$ on top axis and annual total offspring $B(c)$ on bottom axis, with adjusted offspring counts $B(c)'$ outlined.

We adjusted in this way for the sake of a more nuanced time series of total offspring, but this approach may be used to good effect in graduating age-structured counts (births, deaths, populations) whenever time series are long enough to permit information on birth cohort size to propagate through the Lexis surface. These aspects are visible to some degree in the shaded polygons of Fig. 6 in years < 1891.

A.2 Projected birth counts

[TODO: complete when exercise done] Offspring counts by year of ocurrence, $B(c,t)$ are only fully observed for years ≤ 1971. To complete the reflection, we have opted to project birth counts for cohorts whose fertility careers are incomplete. This is done by combining a projection of cohort fertility rates using the method proposed by de Beer (1985) with a standard projection of population denominators (mortality projection too? Could also just take the pop projection from Statistics Sweden.). Light documentation to follow here, as well as an update of Fig. 11.

A.3 Meandering baseline

A peculiar feature of Fig. 6 is the meandering baseline, which replaces the standard straight-line x-axis. The baseline is derived from the crude cohort replacement rate $\mathbb{R}(c)$, defined as $\mathbb{R}(c) = B(c = r)/B(t = r)$. This measure is not a replacement for the classic measure of net reproduction R_0, which differs in a few key ways: i) crude replacement is not sex-specific (our birth series is composed of boy and girl births combined), whereas R_0 is typically defined for females only. ii) while births arise from fertility rates over the life course, the number of potential mothers over the life course is not a mere function of mortality, but of migration as well, and the Swedish birth series will have been affected by heavy out-migration from 1850 until the Second World War (cite SCB), and some in-migration in more recent decades. Cohort R_0 is purged of population structure such as this (except to the extent that subgroups have differential vital rates), whereas $\mathbb{R}(c)$ is not, and for this reason we call it *crude*.

The series of $\mathbb{R}(c)$ is rather smooth without further treatment, save for 11 periodic breaks between 1970 and 1840, a period of rupture between 1865 and 1880, and another set of at least four breaks since the great depression in the 1930s. Rather than preserve these ruptures, we opt to smooth them out and instead capture long term trends in $\mathbb{R}(c)$ in the baseline. Fig. **??**. Keeping the baseline meander smooth minimizes the visual penalty in assessing the variation in $B(c)$ or $B(t)$ separately, but it enhances our ability to see the long term pattern.

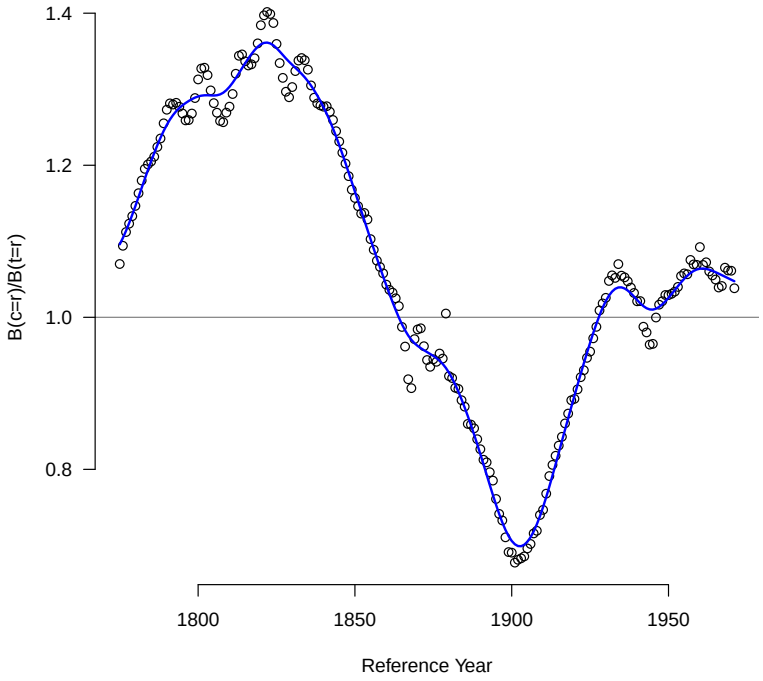


Figure 12: The time series of crude cohort replacement, $\mathbb{R}(c)$, and its smooth pattern (blue line) on which the Fig. 6 meandering baseline is based.