

# Boom, echo, pulse, flow

(Open version)

Tim Riffe<sup>\*1</sup>, Kieron Barclay<sup>1</sup>, Christina Bohk-Ewald<sup>1</sup>, and Sebastian Klüesener<sup>2</sup>

<sup>1</sup>Max Planck Institute for Demographic Research

<sup>2</sup>Bundesinstitut für Bevölkerungsforschung

October 8, 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.

**Keywords:** Fertility, Population structure, Population momentum, Population renewal, Data visualization

## 1 Introduction

Usually demographers think of fertility 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. Since births can happen at any time throughout the year, and since demography usually deals in large numbers, it is common to imagine the birth flow as a continuous stream. This is so not only as a pragmatic assumption to allow for calculus, but it also gives us a heuristic understanding of fertility as a smoother of population structure (Arthur 1982). In this treatment, we retreat from rates, the material of projections, to the absolute number of babies born, the raw material of population renewal.

We aim to represent a historical view of Sweden’s historical birth series in a single multilayered visualization. The birth series is rendered as a flow, for the sake of beauty, and to invite newcomers and curious minds deeper into the discipline of demography. This image entails investment from the viewer, and this manuscript serves as a protracted legend and caption. Intellectual payoffs include a simultaneous sense of long term patterns of generational mixing and generational replacement, medium term baby booms and echos, and the short term shocks of population momentum. We challenge experienced demographers to relate this image to the Lexis diagram, to imagine how the picture would change if fertility were indexed to fathers’ age, and to reimagine this image of aggregates as immense set of lineages.

We use birth count data from Sweden, covering a total of 241 occurrence years from 1775 to 2016. Data for the years 1891 to 2016 is taken directly from the Human Fertility Database. (2017) without further adjustment. We augment the HFD series in both directions, including newly digitized data for the period 1775 to 1890 (de la France) 1907), which we have graduated and adjusted. We describe these

---

<sup>\*</sup>riffe@demogr.mpg.de

adjustments in Appendix A. To complete our picture, we project the fertility of cohorts whose fertility careers are still incomplete (1972-2016) through age 45. We describe the details of this projection in Appendix A.2. The temporal spread from the earliest mother cohort in our final data set (1721) to the latest offspring cohort (2061) is 341 years.

## 2 Age and cohort-structured birth count distributions

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  $x$ .

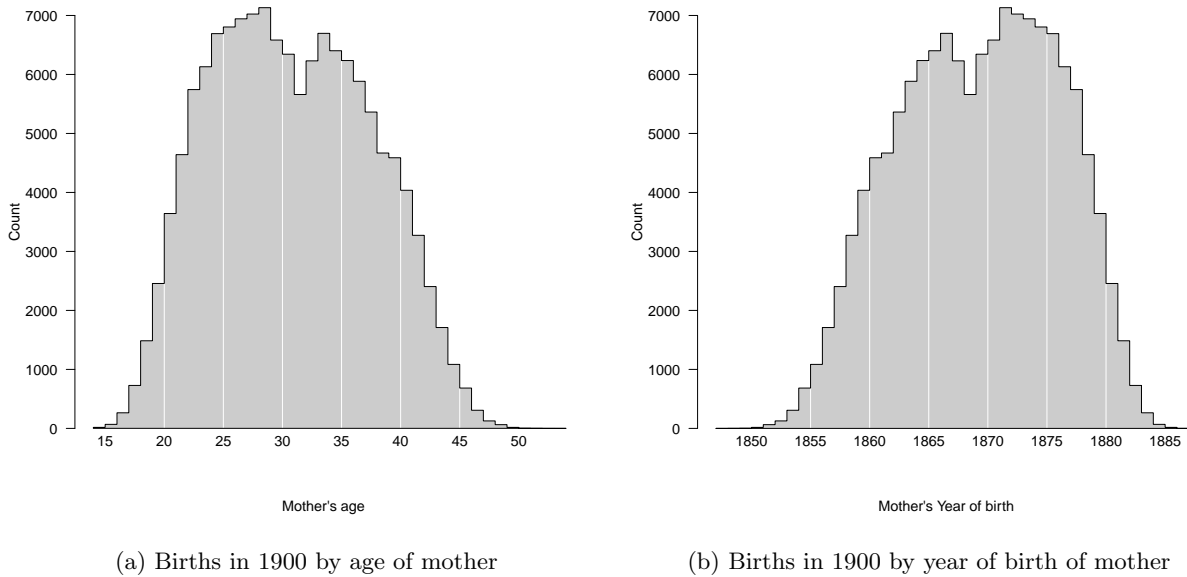


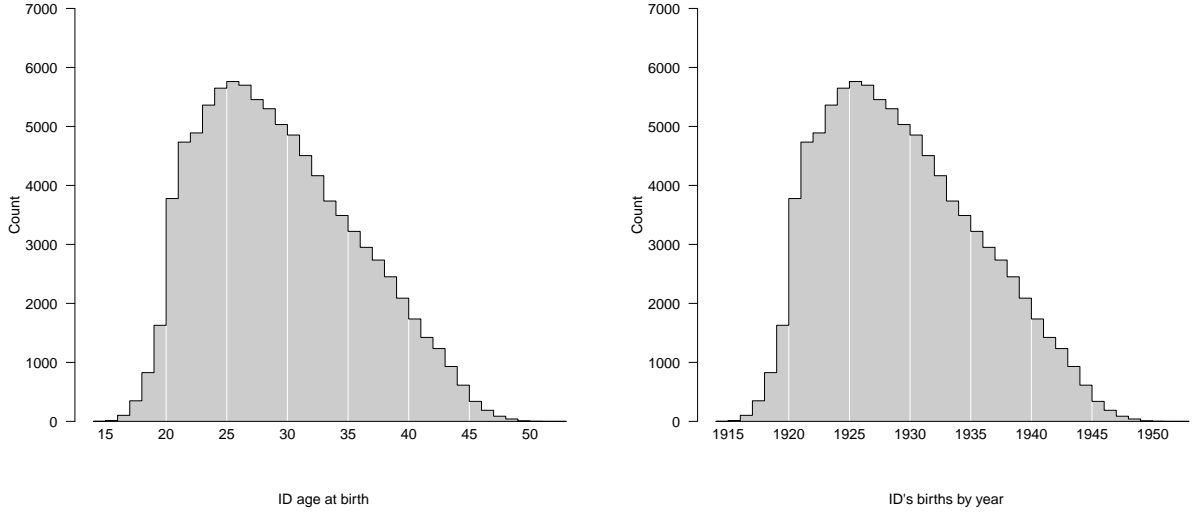
Figure 1: Births in a year structured by mothers’ age versus mothers’ year of birth are a reflection over  $y$  and shift over  $x$ . Count distributions such as this may be jagged, even if the underlying rate distributions are smooth, due to population structure. The deficit around age 31 in 1a is due to a smaller number of potential mothers: the 1871 birth cohort was smaller than the surrounding cohorts.

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, occurrence 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.

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) 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 shift over  $x$ . The births over the life of a cohort more often resemble the smoothness of fertility rate schedules,

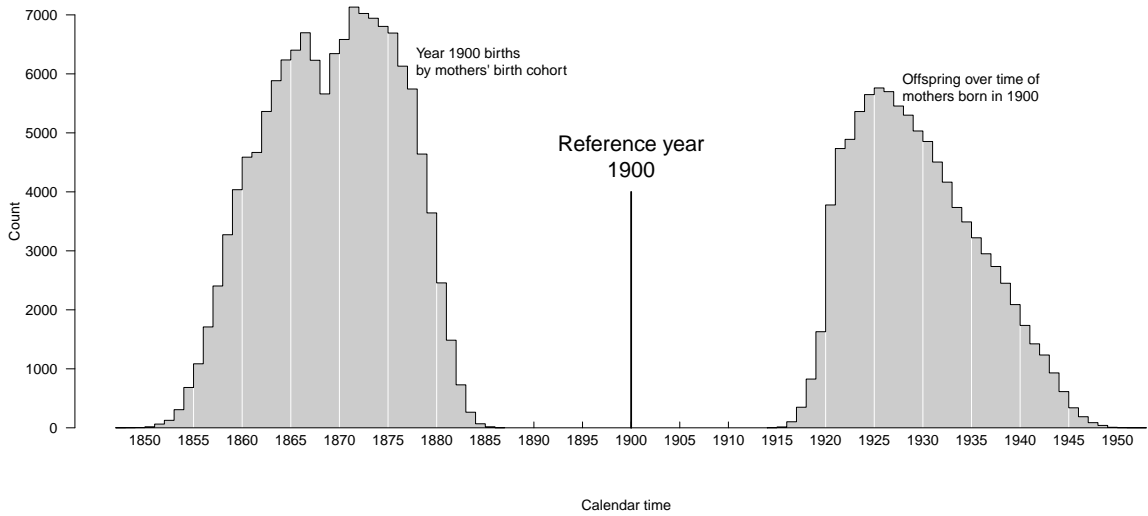


Figure 3: The cohort distribution of mothers who gave birth in 1900 and the births from mothers born in 1900 by year. These two distributions link three generations.

(**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 over the series.

Fig. 4 produces at least two noteworthy artifacts that we may wish to preserve and clarify. 1) First order differences in the top series appear to cascade into the lower series— This derives from a specific kind of population momentum (Keyfitz 1971): larger cohorts have more offspring than smaller

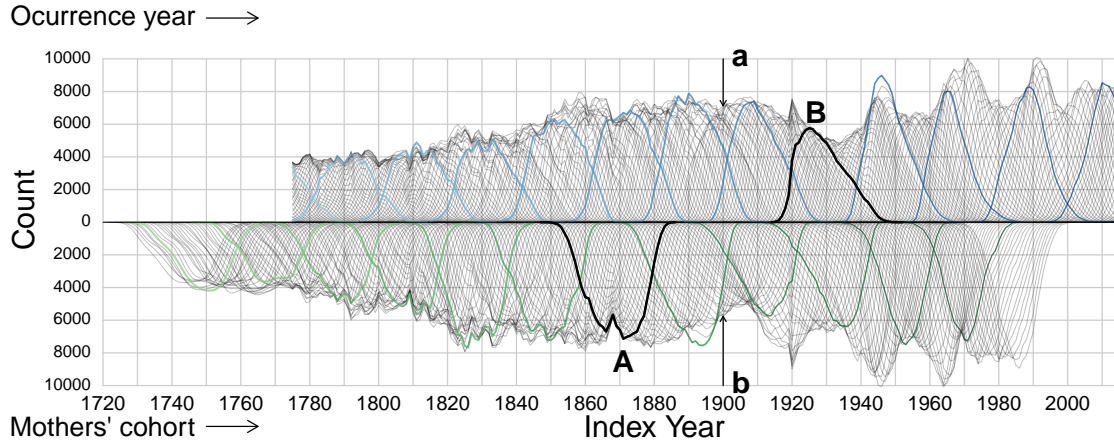


Figure 4: Two time series of birth count distributions. The top series is composed of offspring distributions of mother cohorts over time, indexed to occurrence years. The bottom series is composed of the offspring of a year indexed to mothers' birth cohorts. **B** is the offspring of mothers from the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

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**. This observation deserves further elaboration: The curve **A** is composed of all the births in 1900 indexed *back* to mothers' cohorts. Each point on the curve **A** comes from a different top-axis distribution as it crosses the year 1900 (and vice versa for the bottom). The cross-section of curves **a** is therefore a redundant encoding of the single highlighted curve **A**. The cross-section **b** is a redundant encoding of **B** in the same way. While **A** and **B** are disjoint, and difficult to relate, **a** and **b** share a single  $x$  coordinate, and so may lend themselves to comparison. The "problem" with the cross-sections **a** and **b** is that points from the corresponding distributions **A** and **B** are overlapped due to collasation on a single  $x$  coordinate. It is basically impossible to work out what **A** (**B**) might look like if presented only with **a** (**b**) and its surroundings.



[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,<sup>2</sup> 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).

## 2.1 Notes on visual form

This visualization derives from stacked area charts in general and riverplots (theme river) and stream-graphs in particular (Byron and Wattenberg 2008), but we wish to point out a few notable differences. Our birth flow visualization is composed of two separate stacked area graphs, where polygons appear in chronological order from left to right. If the top and bottom graph sections were vertically centered independently of one another, then these would comprise two “river” graphs. Instead, the two series are squeezed together to share a  $y$  coordinate at 0, and vertical centering is approximated on average by smooth baseline shifts. Different kinds of visual analytic tasks are probably penalized by this choice of form. For example, using the metrics proposed by Thudt et al. (2016), we hypothesize that our visualization would perform moderately well in terms of “individual discrimination” (presumably conditioned on non-zero birth counts!), *very* well in terms of “stream comparison”, since the rendering of each birth distribution matches to  $x$ , and *very* well in terms of aggregate discrimination of top vs bottom (because the meandering baseline gives this), and moderately well in terms of aggregate discrimination of top plus bottom, because river and stream plots also perform well on this metric and our visualization resembles these. On the other hand, certain visual tasks are augmented due to the nature of the data: visual discrimination of polygons is all but guaranteed. Chronological order is clear to the viewer. Color darkness and saturation may make it easier to compare the areas of non-adjacent polygons. Even so, we accept high losses of value lookup ability, for the sake of an aesthetic welcome mat to those who wish to learn more about the fundamentals of demography in general and the Swedish birth flow in particular. Few large questions can be answered with this graphic, but many may be inspired.

On the other hand, this visualization is not an augmented table, but is rather intended to covertly introduce the active viewer into some fundamental features of demography in general and Sweden in particular. It is supposed to infect with curiosity rather than diagnose data. We accept this tradeoff, but note that in a small trial

---

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

### 3 Discussion

[TODO: To be continued...This section and the following analytical perspectives section are only temporary. We probably won't want to introduce new] 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.

### 4 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  $\mathbf{B}(c, t)$ .

#### 4.1 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 occurrence distribution. Since both distributions are derived from the same table of counts by year of occurrence 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 occurrence  $t$  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

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

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.

#### 4.2 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.

#### 4.3 Analytical vignette 3

[TODO: KB suggests highlighting changes in the distribution of maternal age.] This is an interesting take: indeed the outer profiles would end up being the same, but the filled polygons would look different. If we did both at once then it'd make a plaid feel. Worth thought. Age could map to some other color property

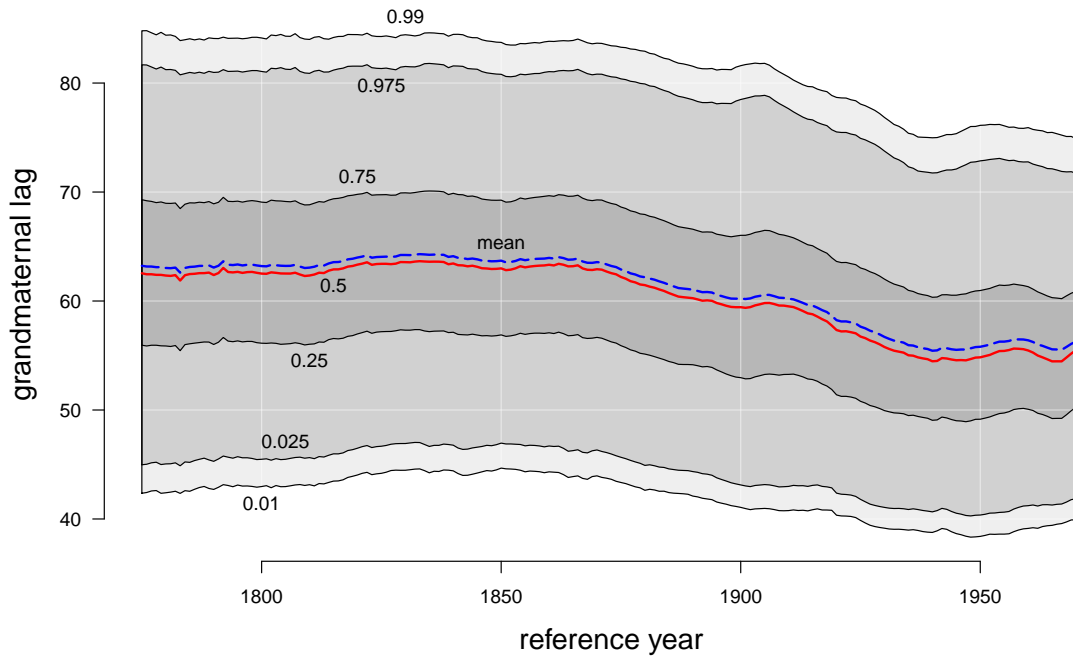


Figure 7: Time lag from mothers' cohort to next generation offspring year of birth, the *grandmaternal lag*, referenced to central (ego) cohorts. Decimals indicate birth distribution quantiles, where the red line indicates the median, and the blue dashed line the mean. The lag decreased broadly by all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

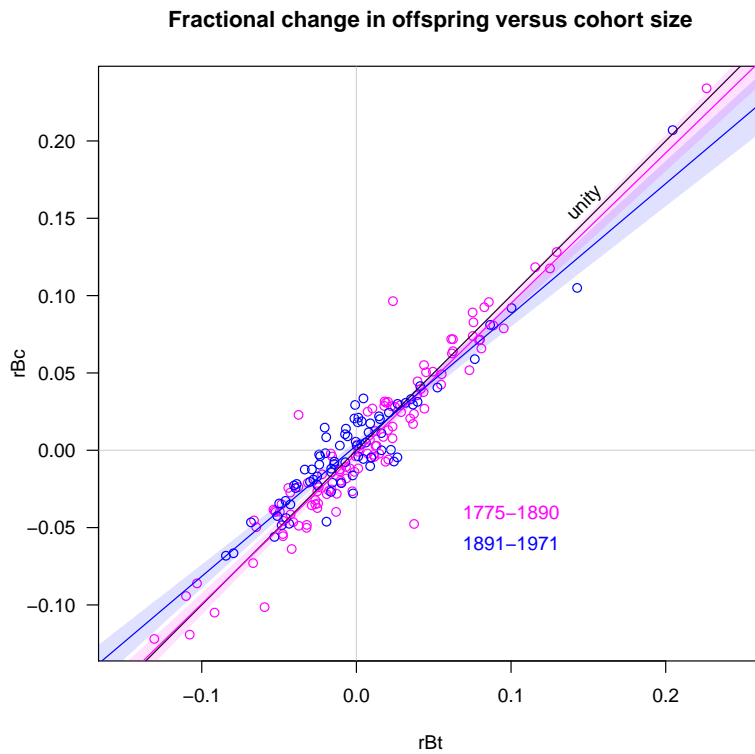


Figure 8: A roughly 1:1 does-response relationship in relative size of structural echo.



within polygons? Note: the younger-older gradient isn't reflected over  $x$ , as chronology is enforced. That is, earlier years of birth of offspring mean young (outside on bottom), but earlier years of birth of mother means old (outside on top). So earlier on the outside moves to later on the inside of the flow, but the younger-older gradient rather follows a down-up pattern within both the top and bottom. So adding age to the picture would require further care. But that's not to say there couldn't be an entirely separate plot of the same data indexed to age. Of the plaid remix of APC, with 'younger-to-older' or 'older-to-younger' from the baseline outward.

## References

- W Brian Arthur. The ergodic theorems of demography: a simple proof. *Demography*, 19(4):439–445, 1982. doi: 10.2307/2061011.
- Lee Byron and Martin Wattenberg. Stacked graphs—geometry & aesthetics. *IEEE transactions on visualization and computer graphics*, 14(6), 2008. doi: 10.1109/TVCG.2008.166.
- Joop de Beer. A time series model for cohort data. *Journal of the American Statistical Association*, 80(391):525–530, 1985. doi: 10.1080/01621459.1985.10478149.
- SGF (Statistique Gnrale de la France). *Statistique internationale du mouvement de la population daprs les registres dtat civil: Rsum rtrospectif depuis lorigine des statistiques de ltat civil jusquen 1905*. Imprimerie national, 1907.
- Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). online, 2017. Available at [www.humanfertility.org](http://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-Institute for Demographic Research and the Vienna Institute for Demography, 2015. URL <https://www.humanfertility.org/Docs/methods.pdf>.
- Nathan Keyfitz. On the momentum of population growth. *Demography*, 8(1):71–80, 1971. doi: 10.2307/2060339.
- 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. doi: 10.1093/aje/kwv020.
- Alice Thudt, Jagoda Walny, Charles Perin, Fateme Rajabiyazdi, Lindsay MacDonald, Diane Vardeleon, Saul Greenberg, and Sheelagh Carpendale. Assessing the readability of stacked graphs. In *Proceedings of Graphics Interface Conference (GI)*, 2016. doi: 10.20380/GI2016.21.

## 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 occurrence 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`<sup>3</sup>. 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 diagrammed in Fig. 9.

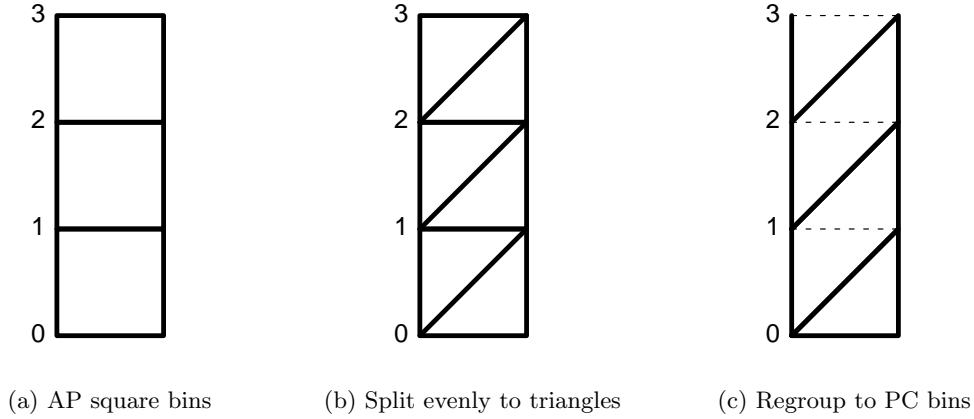


Figure 9: The count regrouping procedure for years 1776 to 1890, step three of data adjustment. Data are graduated to single ages (Fig. 9a), then split in half (Fig. 9b) and regrouped to period cohort (PC) bins (Fig. 9c).

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  $C = P - A - 1$ .<sup>4</sup> 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 transferring it to **B**. We do this by first smoothing the annual time series of total cohort size  $B(t)$  according to some smoothness parameter,  $\lambda$ .<sup>5</sup> 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  $c = t$ .

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

<sup>4</sup>One subtracts 1 because data are in period-cohort bins.

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

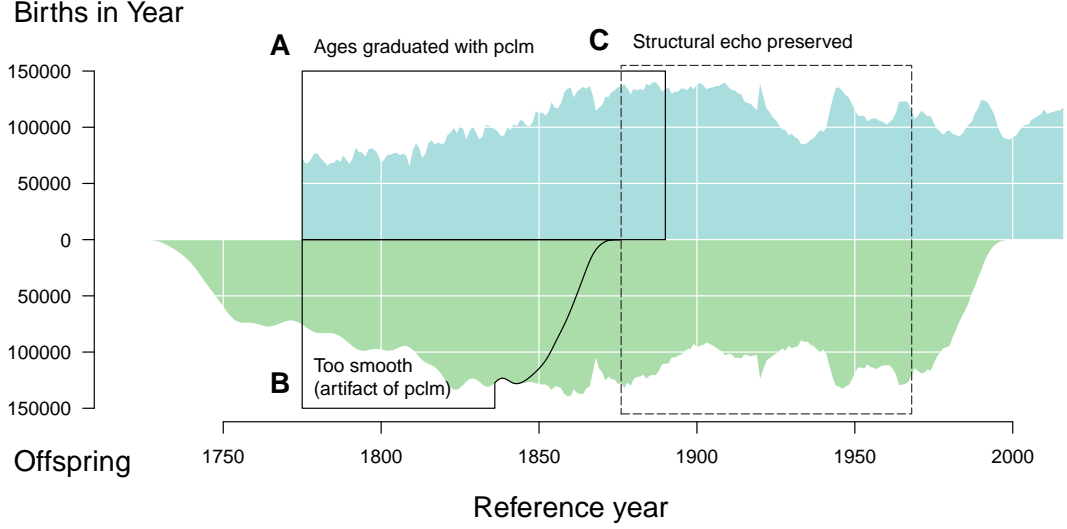


Figure 10: In reference years  $\geq 1891$  both births by year and cohort offspring are directly observed in single year bins, which means that the structural echo between total birth cohort and offspring size is preserved for reference years  $\geq 1876$  (C). Total per annum births in years  $\leq 1890$  (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  $\lambda$  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.

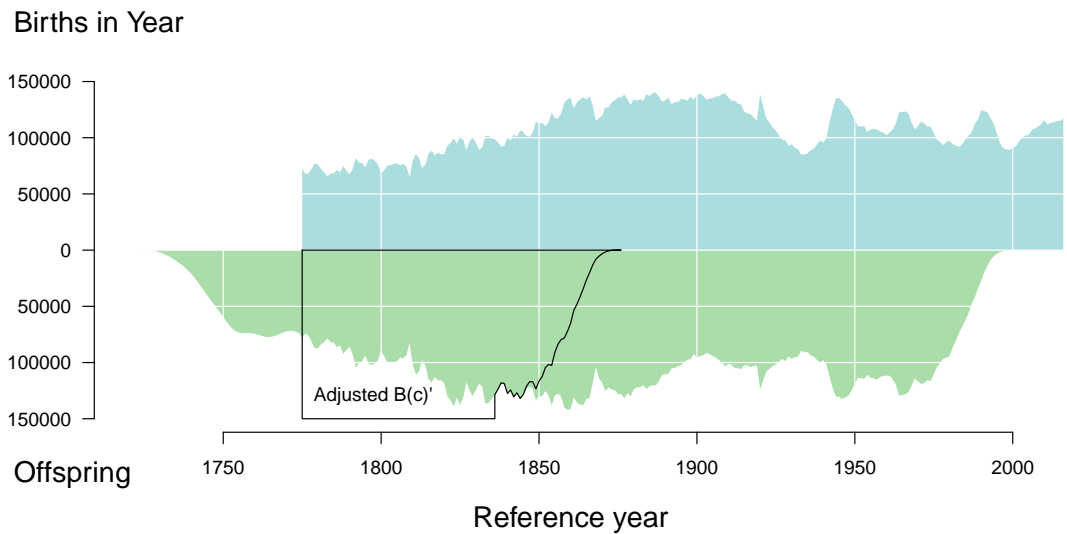


Figure 11: The adjusted birth series. Annual total births  $B(t)$  on top axis and annual total offspring  $B(c)$  on bottom axis, with adjusted offspring counts  $B(c)'$  outlined.

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 occurrence,  $B(c, t)$  are only fully observed for years  $\leq 1971$ . 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. 12. 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.

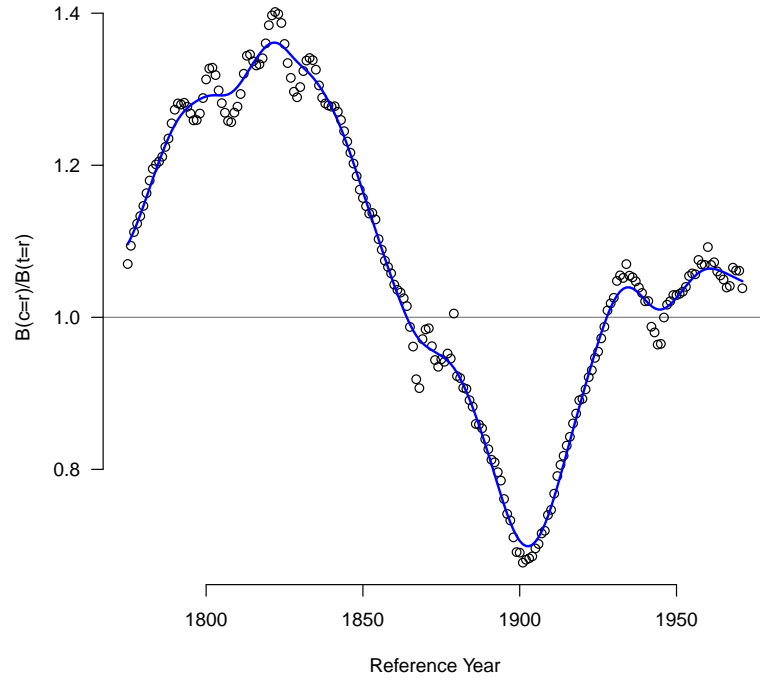


Figure 12: The time series of crude cohort replacement,  $\mathbb{R}(c)$ , and its smooth pattern (blue line) on which the Fig. 6 meandering baseline is based.