Reconstructing long term fertility trends with pooled birth histories

Bruno SCHOUMAKER[1]

PAA Meeting, New Orleans, April 2013

DRAFT

1. Introduction

The accurate measurement of past fertility trends has important practical, theoretical and policy implications. Assumptions about future fertility in population projections are largely influenced by past fertility trends. The evaluation of population policies is also to some extent based on the reliable measurement of fertility trends. The measurement of the timing and of the speed of fertility decline is also central to theories of fertility changes.

In developing countries - where civil registration systems are deficient - fertility trends are often obtained with a few estimates from censuses and surveys (United Nations, 2011). Such an approach tends to mask changes that occurred between two data points, especially when there are only a few estimates. Fertility trends computed in this way may also be largely influenced by varying data quality between surveys (Schoumaker, 2010).

In this paper, I present a method for reconstructing and smoothing long term fertility trends by combining birth histories from multiple surveys, using Poisson regression and restricted cubic splines. The data used in the paper come from World fertility surveys and Demographic and Health Surveys[2]. In the first part, I present the method and I illustrate its application by combining several fertility surveys in Colombia. Next, I use simulated birth histories to test the method in controlled situations. Finally, I apply the method to several countries, from various parts of the world, with varying numbers of surveys and with different data quality issues. The reconstructed trends are compared with published trends (United Nations Population Division and DHS).

[1] Bruno.schoumaker@uclouvain.be, Centre de recherche en démographie et sociétés, Université catholique de Louvain (Belgium).

[2] Data from any other survey with birth histories (such as some MICS) can also be used.

2. Reconstructing fertility trends from birth histories – a brief review of approaches

DHS reports typically publish age-specific fertility rates by 5-year periods before the survey that allow reconstructing partial TFRs (sometimes computed in reports). Garenne and Joseph (2002) used DHS birth histories to compute births and exposure by age groups and calendar years on pooled data from several surveys, and estimated total fertility rates over long periods in several countries. They smooth TFRs by fitting polynomials on the estimated TFRs. More recently, Machiyama (2010) computed partial TFRs from birth histories in several sub-Saharan African countries, and used Loess techniques to smooth fertility trends. In addition, she corrected displacements of recent births by transferring some births from one year to the preceding year before computing rates[3].

The approach presented in this paper is a generalization of an approach used by the author in previous research (Schoumaker, 2004; Schoumaker 2010; Schoumaker, 2013). It uses Poisson regression to reconstruct fertility trends from several surveys pooled together. The major difference with Garenne and Joseph 2002), and Machiyama's work (2010) is to use a regression framework to reconstruct fertility trends. This allows estimating TFRs between 15 and 49 with truncated data in a straightforward way (with a few assumptions), whereas other authors estimate TFRs between 15-34. Our method also allows smoothing fertility trends by including appropriate variables in the regression (e.g. splines), and allows testing for changes in speed of fertility decline. Data quality problems can also to some extent be corrected including appropriate variables in the regression model. Finally, our approach is tested with simulated birth histories.

3. Method

The method consists in reconstructing trends in total fertility rates (15-49) with Poisson regression, and smoothing trends using restricted cublic splines. It is first described with data from a single survey; next, it is presented with several surveys that are pooled together.

3.1. Fertility data from a birth history

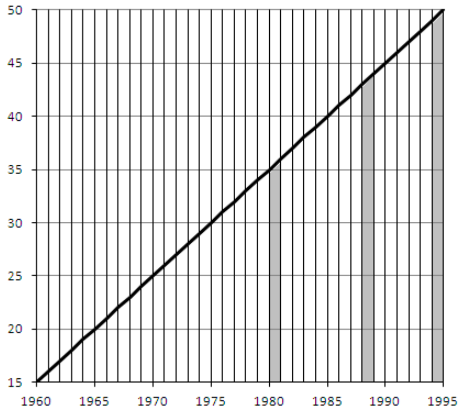
The basic principle can be best illustrated with a Lexis diagram showing the data that are obtained from a birth history (figure 1). A typical fertility survey collects birth histories among women aged 15-49 at the time of the survey. In this example, the survey is conducted on the 31st of December 1994. The corresponding data

[3] Other recent research on the reconstruction of fertility trends includes Alkiema et al. (2012).

needed to compute fertility rates is shown in Table 1. Data come from the 1995 DHS in Colombia.

Using data from the full birth history, it is possible to compute the number of births and the total exposure (number of women-years) for each year and each age-group below the black diagonal[4]. Computing fertility rates thus necessitates producing a table containing for each year and each age group, the number of births and the total exposure. Each cell in the following table corresponds to a rectangle in the Lexis diagram. As is clear from this table and the Lexis diagram, some data are missing because of the truncation.

Figure 1: Lexis diagram illustrating birth history data.



For the year preceding the survey (1st year, or year 15), births and exposure are available for all age groups. In year 9 (7th year before the survey), no data is available for the age group 45-49, and the age group 40-44 is only partially covered. In the 15th year before the survey (1980), births and exposure are complete for the first four age groups (15-19, 20-24, 25-29, 30-34), and a very small part of the 35-39 age group is observed (9 births, 78 years of exposure). No data is available for age groups 40-44 and 45-49.

Using classical demographic methods, the total fertility rates are obtained by computing age-specific fertility rates in each of the seven 5-year age groups, summing these age-specific fertility rates[5] and multiplying them multiplied by 5. For the last year, the TFR can be computed between 15 and 49. As one goes back in time, the age range becomes more limited. For instance, the TFR for 1988 can only be obtained for women 15-44, and the rate for this last age group will be slightly

[4] A stata command (tfr2) was produced by the author

[5] The computation of age-specific fertility rates consist in dividing the number of births in a rectangle of the Lexis diagram by the exposure in that rectangle.

biased, since no information is available for women aged 44 in 1988. In 1979, the TFR can only be computed among women aged 15-34.

Table 1: Table of births and exposure from a retrospective birth history, Data from 1995 Colombia DHS

Period	Year before survey	Age-group	Births	Exposure (years)
15	1st	15-19	181	2099
15	1st	20-24	326	1937
15	1st	25-29	286	1822
15	1st	30-34	151	1578
15	1st	35-39	61	1385
15	1st	40-44	26	1192
15	1st	45-49	2	848
…				
9	7th	15-19	189	1937
9	7th	20-24	325	1822
9	7th	25-29	222	1578
9	7th	30-34	165	1385
9	7th	35-39	75	1192
9	7th	40-44	21	848
…				
0	15th	15	162	1652
0	15th	20	261	1417
0	15th	25	285	1247
0	15th	30	150	985
0	15th	35	9	78

When reconstructing TFRs from birth histories, the usual approach consists in computing truncated TFRs. For instance, the TFR for the 10 years preceding the survey would be computed over the age range 15-39.

3.2 Reconstruction of TFRs from a single survey with Poisson regression

The method consists in estimating TFRs by using Poisson regression with age groups and years as independent variables (Schoumaker, 2004; Schoumaker, 2013). Poisson regression can be used with individual data or grouped data, leading to strictly identical results (Powers and Xie, 2000). Because it is computationally less intensive, using grouped data allows fitting the models more quickly. For this reason, we use grouped data like in Table 1[6].

The model we use can be summarized using this equation.

$$\log(\mu_{it}) = \log(e_{it}) + f(age) + g(time) \qquad \text{[Eq. 1]}$$

[6] However, using individual data leads to strictly identical results, and may be more flexible in some applications. See Schoumaker (2004) for the presentation of the method with individual data.

μ_{it} is the expected number of births (births column) at each age (i) and each period (t), e_{it} is the total exposure at each age and each period (exposure column), f(age) is a function of age, and g(time) is a function of the calendar time. The term $\log(e_{it})$ is the *offset,* and has a fixed coefficient equal to one.

This equation can be reorganized by dividing the number of births by the exposure.

$$\log\left(\frac{\mu_{it}}{e_{it}}\right) = \log(\lambda_{it}) = f(age) + g(time) \qquad \text{[Eq. 2]}$$

$$\lambda_{it} = \exp[f(age)] * \exp[g(time)] \qquad \text{[Eq. 3]}$$

λ_{it} is the fertility rate at age i and period t, and is equal to the product of an age effect and a time effect. This log-rate model is estimated with Poisson regression (Powers and Xie, 2000).

An assumption made here is that the shape of the age-specific fertility rates is constant over time, i.e. that there is no interaction between age effects and time effects. In other words, time has a multiplicative effect on fertility rates that is similar at all ages. Comparisons with trends in partial TFRs between 15-34 and simulations indicate that this assumption does not have a strong influence on fertility trends over a relatively short period of time (10-15 years). As discussed later, the assumption of a constant age fertility schedule can be relaxed by using changing age schedules in the offset. This is what is done when several surveys are pooled together.

The following model illustrates this method for reconstructing fertility over the 15 years preceding the 1995 DHS in Colombia[7]. In this example, age is included as a set of dummy variables for five-year age groups (Table 2). Calendar time is measured by dummy variables to model annual variations in fertility. The model is written in the following way.

$$\log(\mu_{it}) = \log(e_{it}) + \alpha + \sum_{k=20-24}^{45-49} \beta_k . A_{kit} + \sum_{h=2}^{15} \delta_h . T_{hit} \qquad \text{[Eq. 4]}$$

Alternatively, the rate can be expressed with this equation.

[7] Although the method can be used to reconstruct fertility trends from a single survey over a longer period than 15 years, possible omissions of births in the past and smaller sample sizes, as well as the departure from the assumption of constant age fertility schedule lead us to limit the period to the last 15 years.

$$\lambda_{it} = \exp\left[\alpha + \sum_{k=20-24}^{45-49} \beta_k . A_{kit}\right] * \exp\left[\sum_{h=2}^{15} \delta_h . T_{hit}\right] \qquad \text{[Eq. 5]}$$

α is the constant term, A_{kit} are dummy variables for the 6 age groups from 20-24 to 45-49, and T_{hit} are 14 dummy variables for the years after the reference year.

Predicting the fertility rate for a single age group (e.g. age group 25-29) for a specific year (e.g. year 5) is straightforward. The dummy variables are equal to 1 for the specific age group and year (and 0 for the other age groups and years), and the rate is just a function of the constant, the regression coefficient for the 25-29 age group, and the regression coefficient for the 5th year dummy variable.

$$\lambda_{it} = \exp[\alpha + \beta_{25-29}] * \exp[\delta_5] \qquad \text{[Eq. 6]}$$

The total fertility rate (15-49) for year t is equal to 5 times the sum of age-specific fertility rates, multiplied by the exponential of the regression coefficient of the dummy variable for year h.

$$TFR_h = 5 * \left(\exp[\alpha] + \sum_{k=20-24}^{45-49} \exp[\alpha + \beta_k] \right) * \exp[\delta_h] \qquad \text{[Eq. 7]}$$

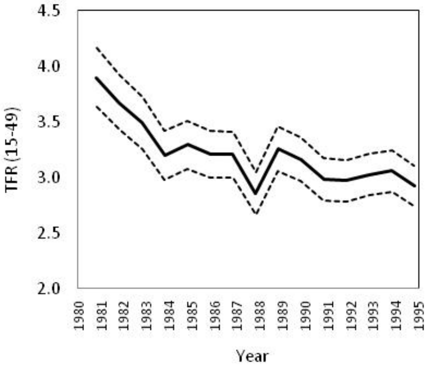
Regression coefficients of the Poisson regression are reported in the second column of Table 1. From these regression coefficients, age-specific fertility rates for the reference year (1) are computed. The fertility rate for a specific age group (column 4) is equal to the exponential of the sum of the constant of the model and of the coefficient of the age group (the sum of the coefficients are log(rates), in column 3). The sum of the age-specific fertility rates for all the age groups multiplied by 5 is equal to the TFR for the reference year (1st year of the 15 year period, equal to 3.90 in Table 1). The TFRs for the following years are obtained by multiplying the TFR of the reference year by the exponentials of regression coefficients of the following years. Figure 2 shows the annual variations of the TFRs (same values as in Table 1), with 95% confidence intervals. Standard errors for the TFR are computed using the delta method[8]. As will be shown later, the trend can be smoothed with restricted cubic splines.

[8] A Stata command (called tfr2) was prepared by the author for computing TFRs and reconstructing fertility trends from a single survey (see Schoumaker, 2013).

Table 2: Age specific fertility rates and reconstruction of fertility trends over the fifteen calendar years preceding the 1995 DHS survey in Colombia with Poisson regression.

Independent variables (1)	Regression coefficients (2)	Log(rates) (α+β) (3)	Rates exp[α +β] (4)	TFR (5)
	A			
Constant	-2.212			
	B			
15-19	(REF)	-2.212	0.110	
20-24	0.669	-1.543	0.214	
25-29	0.559	-1.652	0.192	
30-34	0.248	-1.963	0.140	
35-39	-0.234	-2.451	0.086	
40-44	-1.183	-3.395	0.034	
45-49	-3.116	-5.327	0.005	
	δ			
Year 1	(REF)			3.90
Year 2	-0.060			3.67
Year 3	-0.111			3.49
Year 4	-0.197			3.20
Year 5	-0.169			3.29
Year 6	-0.194			3.21
Year 7	-0.196			3.21
Year 8	-0.311			2.86
Year 9	-0.180			3.26
Year 10	-0.209			3.16
Year 11	-0.268			2.98
Year 12	-0.272			2.97
Year 13	-0.253			3.03
Year 14	-0.243			3.06
Year 15	-0.287			2.93

Figure 2: Reconstruction of TFRs (15-49) and 95% confidence intervals over the 15 years preceding the survey, Colombia 1995 DHS.



Another equivalent way of estimating the model is to control the age pattern of fertility in the offset (controlling both for exposure and age pattern of fertility). The model is written in the following way, where a_{it} measures the age pattern of fertility.

$$\log(\mu_{it}) = \log(e_{it}) + \log(a_{it}) + \gamma + \sum_{h=2}^{15} \delta_h . T_{hit} \qquad \text{[Eq. 8]}$$

Comparing equation [8] with equation [4], we see that

$$\log(a_{it}) = \alpha + \sum_{k=20-24}^{45-49} \beta_k . A_{kit} - \gamma \qquad \text{[Eq. 9]}$$

and

$$a_{it} = \frac{\exp\left[\alpha + \sum_{k=20-24}^{45-49} \beta_k . A_{kit}\right]}{\exp[\gamma]} \qquad \text{[Eq. 10]}$$

The numerator of this expression is the fertility rate for the reference year. The denominator is the constant of the second model. By constraining the sum of a_{it} over the 7 age groups to be equal to 1, a_{it} are proportionate fertility rates. Then

$$\exp[\gamma] = \sum_{k=20-24}^{45-49} \exp[\alpha + \beta_k] \qquad \text{[Eq. 11]}$$

$$a_{it} = \frac{\exp\left[\alpha + \sum_{k=20-24}^{45-49} \beta_k . A_{kit}\right]}{\sum_{k=20-24}^{45-49} \exp[\alpha + \beta_k]} \qquad \text{[Eq. 12]}$$

The total fertility rate is computed as:

$$TFR_h = 5 * \exp[\gamma] * \exp[\delta_h] \qquad \text{[Eq. 13]}$$

In other words, if the age pattern of fertility is controlled in the offset, the TFR for the reference year is the exponential of the constant multiplied by five; the TFR for the other years are obtained by multiplying the TFR of the reference year by the exponential of dummy variables for the year h.

In summary, this alternative approach requires first estimating the age-specific fertility rates for the reference year and dividing these rates by their sum to obtain proportionate age-specific rates. These proportionate rates are then included in the offset[9], and a new model is estimated. Although controlling the age pattern of fertility in the offset has no interest when working with a single survey – and

[9] Mutiliplying exposure by the proportionate rate and taking its logarithm.

involves running two models leading to identical fertility trends – it proves useful when several surveys are pooled together[10]. The age pattern can be estimated for each survey separately, and included in the offset when surveys are pooled together in order to relax the assumption of constant age-schedule.

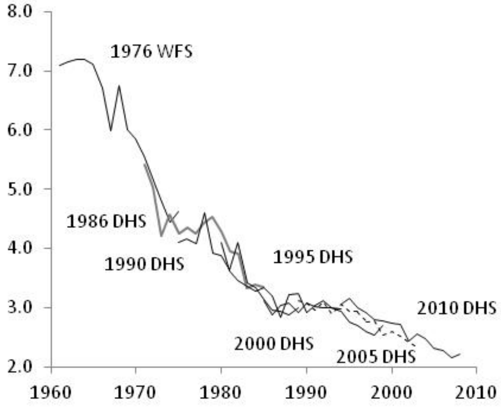
In the next section, the approach is extended to be used with several birth histories, and trends are smoothed with restricted cubic splines.

3.3 Reconstruction of TFRs from several surveys with Poisson regression

The method we just described can be applied to a pooled data set of birth histories from several surveys. We first illustrate it in the case of Colombia with 7 surveys (1976 WFS, 1986 DHS, 1990 DHS, 1995 DHS, 2000 DHS, 2005 DHS, 2010 DHS)[11].

Before pooling the surveys, fertility was reconstructed over the last 15 years for each of the seven surveys separately (Figure 3), in the same way as in section 3.2. The seven surveys match quite well, and show a downward trend, punctuated by slowdowns in the 1970s and 1990s.

Figure 3: Reconstruction of TFRs (15-49) over the 15 years preceding seven surveys, Colombia.



The next step consists in pooling the seven data sets, and reconstructing the fertility trend using Poisson regression. Pooling the data sets consists in preparing a data set similar as in Table 1 for each survey, and appending the

[10] It is also possible to include an age pattern from another source – this is not discussed in this presentation.

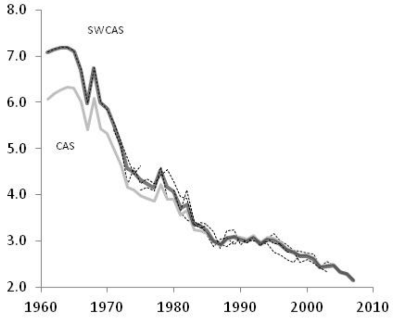
[11] The situation of Colombia is to some extent exceptional, because of the large number of surveys, their relatively good quality, and their large sample sizes.

seven data sets (seven surveys) to for a single data file. In order to compute fertility rates by calendar year, the datasets for each survey are prepared for the fifteen calendar years preceding each survey.

Assuming a constant age pattern of fertility over a 50 year period may lead to biased fertility trends. If fertility decreases more quickly among older women, this would lead to underestimating the TFR in the past[12]. The assumption of constant age schedule can be relaxed by controlling for the age pattern in the offset, and allowing the age pattern to vary over time. In this paper, this is done by considering a survey-specific age schedule. For each survey, the proportionate age-specific fertility rates are computed (see previous section) and included in the offset. The age pattern is considered constant for each survey, but variable across surveys. Although this assumption does not strictly hold, simulations show it performs well[13].

Figure 4 below illustrates the reconstruction of the TFR in Colombia between 1961 and 2008 using Poisson regression with (a) a constant age schedule (CAS), and (2) survey-wise constant age schedules (SWCAS). Dotted lines represent estimates from the 7 separate surveys (as shown on Figure 3). This figure shows that the SWCAS approach fits the separate estimates very well. Before the mid 1980s, the CAS approach leads to an underestimation of the TFR, due to the more rapid fertility decrease at advanced ages. In the following part of this illustration, we use the SWCAS approach.

Figure 4: Reconstruction of TFRs (15-49) between 1961 and 2008 by pooling 7 surveys, Colombia. Comparison of Constant age schedule (CAS) and Survey-wise constant Age schedule (SWCAS)



[12] This is because rates above 35 are reconstructed from the model, and are assumed to have declined at the same pace as the fertility rates at young ages (the opposite would be true if fertility declined more rapidly at younger ages).

[13] Since birth histories from several surveys overlap, the assumption of constant age schedule for each survey does not mean the age schedule is constant over time between for each period covered by the surveys.

3.4. Restricted cubic splines

Smoothing fertility trends is performed with restricted cubic splines, i.e. piecewise polynomial functions constrained to join at predefined years (knots) (Andersen, 2009). Cubic splines are flexible and allow fitting a large variety of shapes with relatively few parameters (Harrell, 2001). To fit restricted cubic splines with K knots, K-1 variables (functions of time periods) are created. The construction of these variables depends on the number and the location of knots[14]. The new variables are introduced as explanatory variables in the Poisson regression model, in place of the dummy variables[15]. The predicted total fertility rate is also obtained for each year using the coefficients of regression. The model is of the same form as in [Eq. 1]. The only difference is that *g(time)* is not modeled as a series of dummy variables, but as a linear function of the K-1 variables (RCS) created to fit the restricted cubic splines.

$$\log(\mu_{it}) = \log(e_{it}) + \log(a_i) + \gamma + \sum_{h=1}^{K-1} \delta_h.RCS_{hit} \qquad \text{[Eq. 14]}$$

For instance, when there are five knots, the model is

$$\log(\mu_{it}) = \log(e_{it}) + \log(a_i) + \gamma + \delta_1.RCS_{1it} + \delta_2.RCS_{2it} + \delta_3.RCS_{3it} + \delta_4.RCS_{4it} \qquad \text{[Eq. 15]}$$

The TFR is estimated as:

$$TFR_t = 5 * \exp[\gamma] * \exp[\delta_1.RCS_{1it} + \delta_2.RCS_{2it} + \delta_3.RCS_{3it} + \delta_4.RCS_{4it}] \qquad \text{[Eq. 16]}$$

The number and location of knots have to be defined before adjusting the restricted cubic splines. The shape of the smoothing function is not very sensitive to the location of the knots (Harrell, 2001; Andersen, 2009; Dupont, 2009), but is more sensitive to the number of knots. Figure 5a to 5c compare reconstructed TFRs using restricted cubic splines with knots located every 10 years, 5 years and 3 years[16]. Spacing knots by 10 years leads to overlooking some changes in the pace in the fertility transition. Locating knots every five years gives a much better fit, and

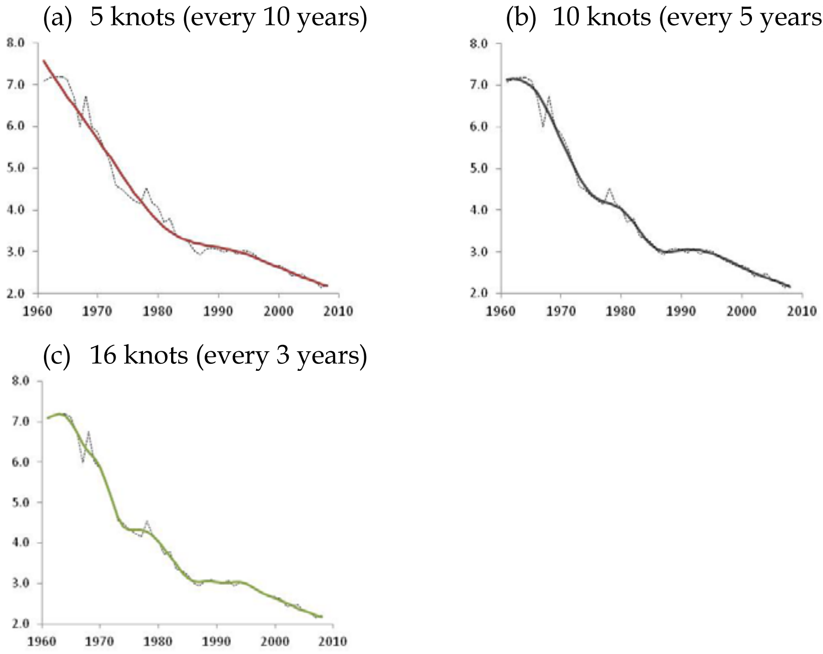
[14] The *mkspline* command in Stata is used to create these variables after defining the number and the location of the knots (StataCorp, 2007).

[15] The rest of the model is similar as in the previous section: the age pattern is controlled in the offset.

[16] The location of knots is done backward, starting from the last knot. The last knot is located (I/2) years before the last year, where I is the width of the interval. For instance, with a 5-year interval, and the last year in 2008 (mid-point 2008.5), the last knot is located on 2006. The next to last knot is located on 2001, and so on. Different locations of knots were tested, and have a very small impact on the shape of the smoothing function.

indicates two slowdowns in the fertility transition. Finally, locating knots every three years fits the TFR only marginally better. In the rest of this paper, knots are located every five years, as a compromise between flexibility and parsimony.

Figure 5: Adjustement of restricted cubic splines to reconstructed TFRs (15-49) between 1961 and 2008, Colombia.



(a) 5 knots (every 10 years)

(b) 10 knots (every 5 years)

(c) 16 knots (every 3 years)

3.5. Testing for stalls

An additional interest of the RCS approach is to allow testing changes in the rhythm of fertility transitions, and notably to identify stalls. With our method, this can be done by testing if the slope of the RCS is significantly negative. This is done by computing marginal effects (the first derivative) of the RCS function, as well as their standard errors (Buis, 2009)[17]. A marginal effect that is not significantly negative is interpreted as a stall[18].

[17] This is done in stata with the user-written command *mfxrcspline* (Buis, 2009).

[18] A one-tailed test is performed by constructing 90% confidence intervals around the marginal effects. If 0 is not included in the 90% confidence interval, the slope is significantly negative with a 95% confidence.

Figure 6: Marginal effect of RCS and 90% confidence interval, TFRs (15-49) between 1961 and 2008, Colombia.

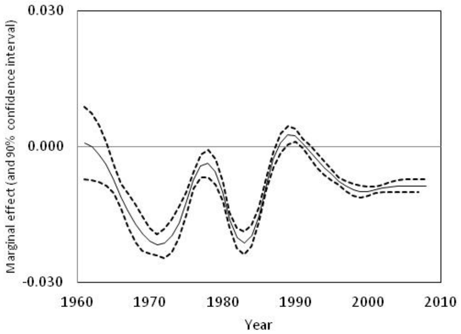


Figure 6 shows the marginal effects and their 90% confidence interval for the RCS with knots located every five years (corresponding to figure 5b). In the first few years, the slope is not significantly negative. The rate of fertility decline reaches -1.5% per year around 1970, and the fertility decline slows down in the late 1970s. A new acceleration of fertility decline occurs in the early 1980s, followed by a deceleration and a stall between 1988 and 1992. After 1992, the decline resumes, and fertility currently declines at a rate of 0.7% per year.

Figure 7: Reconstructed TFR (15-49) in Colombia (1961-2008), 95% confidence intervals, and identification of periods of no significant decline (orange).

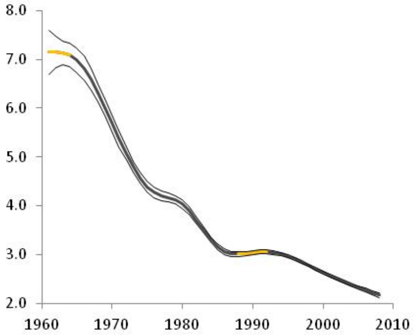


Figure 7 summarizes this information by representing years when fertility decreases significantly in black, and years when fertility is either stable or increasing (stall) in orange (light grey in black and white). The 90% confidence interval for the TFRs is also shown on these figures (black smooth lines above and below the reconstructed trends)[19].

[19] The sample design is taken into account using the jackknife method (correcting for clustering), and standard errors of TFRs are computed from standard errors of the

4. Testing the method with simulations

Two scenarios of fertility transitions are used to generate birth histories through micro-simulation (using SOCSIM)[20]. These scenarios differ by the speed of fertility decline, as well as the changing age pattern of fertility. The first scenario is characterized by a more rapid fertility decline at low ages (as in Morocco). Mean age at childbearing increases, at least in the first phase of the fertility decline. The second scenario represents a situation where fertility decreases more quickly at high ages (as in Colombia). Mean age at childbearing decreases.

5 DHS-like samples are selected from these simulated birth histories at 5 year intervals (the samples are around 4000 women), and fertility is reconstructed over 35 years from these samples using the method described above. 2 periods are covered separately (1945-1980 and 1960-1995) for each scenario. The reconstructed trends can be compared to the trend of the TFR used as the input of the simulation (as well as the TFR computed from the SOCSIM output, i.e. birth histories).

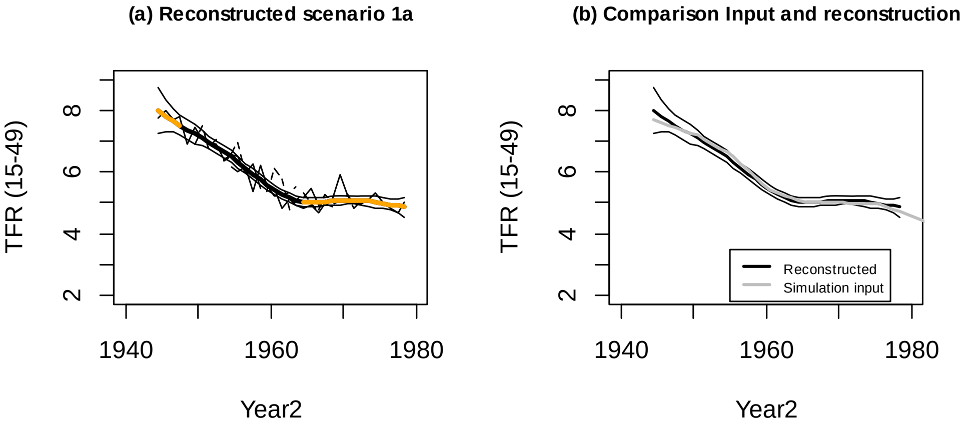
Figure 8 compares smoothed reconstructed fertility (and 95% confidence intervals) with reconstructed fertility in each of the 5 'surveys' (for the 4 scenarios, 1a, 1b, 2a 2b), and compares the smoothed reconstructed fertility with the input of the simulation.. Overall, the black line (reconstructed fertility trends) is close to the grey line (input of simulation), indicating the method based on pooling DHS allows reconstructing the trends with reasonable precision.

coefficients with the delta method. In the case of Colombia, the sample sizes of the various surveys are large, resulting in small standard errors. In most situations, confidence intervals are larger

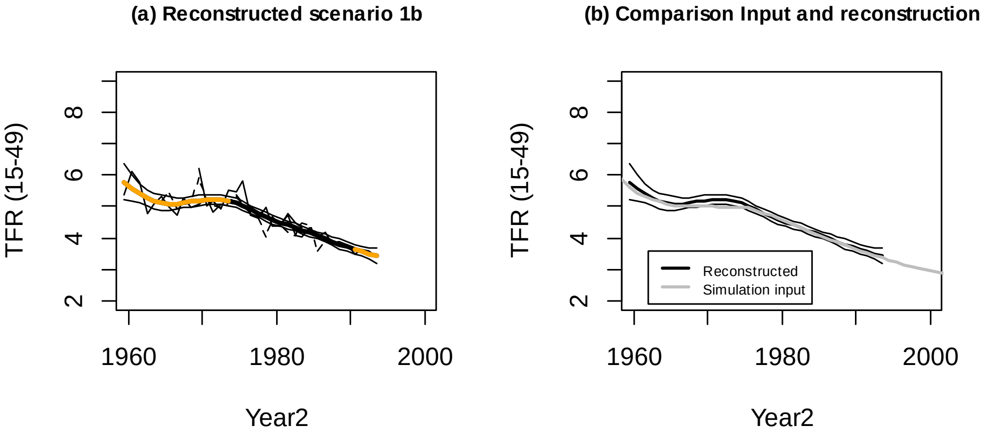
[20] We thank Carl Mason from the University of California for generating simulated birth histories.

Figure 8: Reconstructed fertility trends from 5 simulated DHS-like surveys at 5-year intervals, and comparison with input of simulations

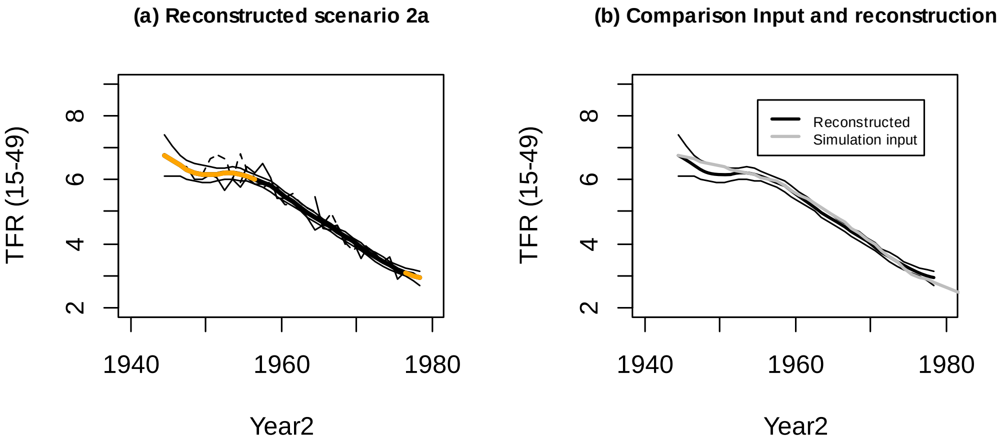
(scenario 1a – decreasing mean age at maternity, 1945-1980)



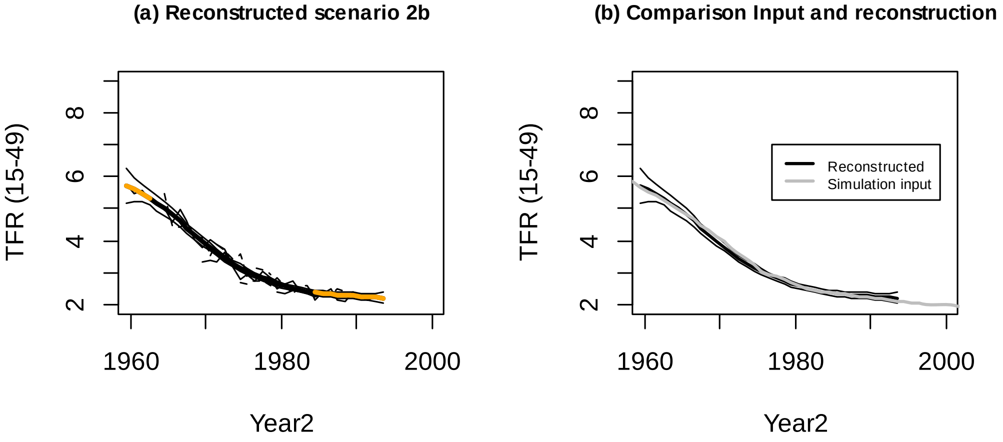
(scenario 1b – decreasing mean age at maternity, 1960-1995)



(scenario 2a – increasing mean age at maternity, 1945-1980)



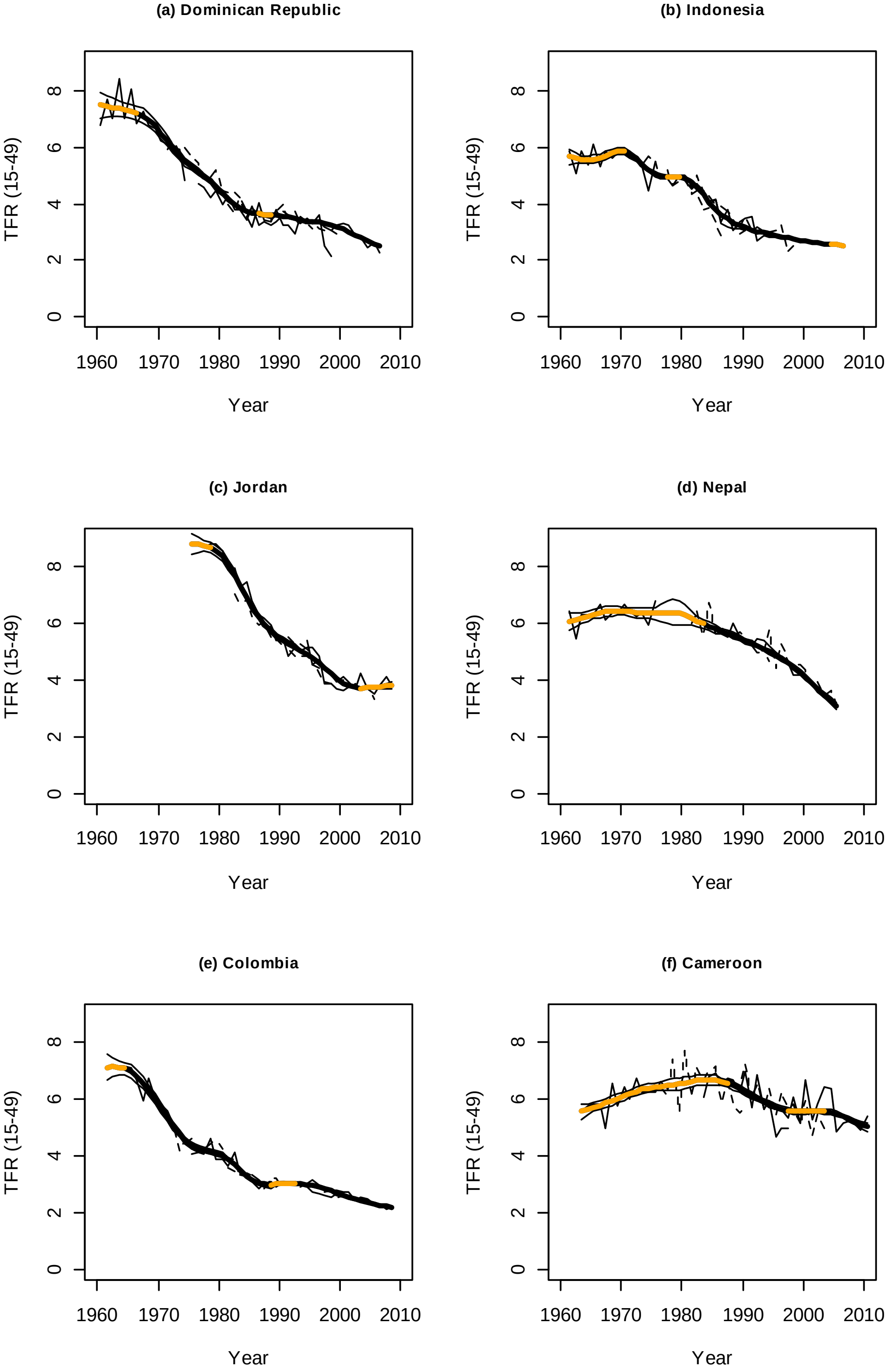
(scenario 2b – increasing mean age at maternity, 1960-1995)



5. Reconstructed fertility trends in selected countries

The method is applied in six countries from various parts of the World, combining WFS and DHS surveys (the Dominican Republic, Indonesia, Nepal Jordan, Colombia and Cameroon) (Figure 9). These six examples show that fertility trends are not linear, and are characterized by accelerations and slow downs. Stalls are visible in five of the six countries (Dominican Republic, Indonesia, Jordan, Colombia, and Cameroon).

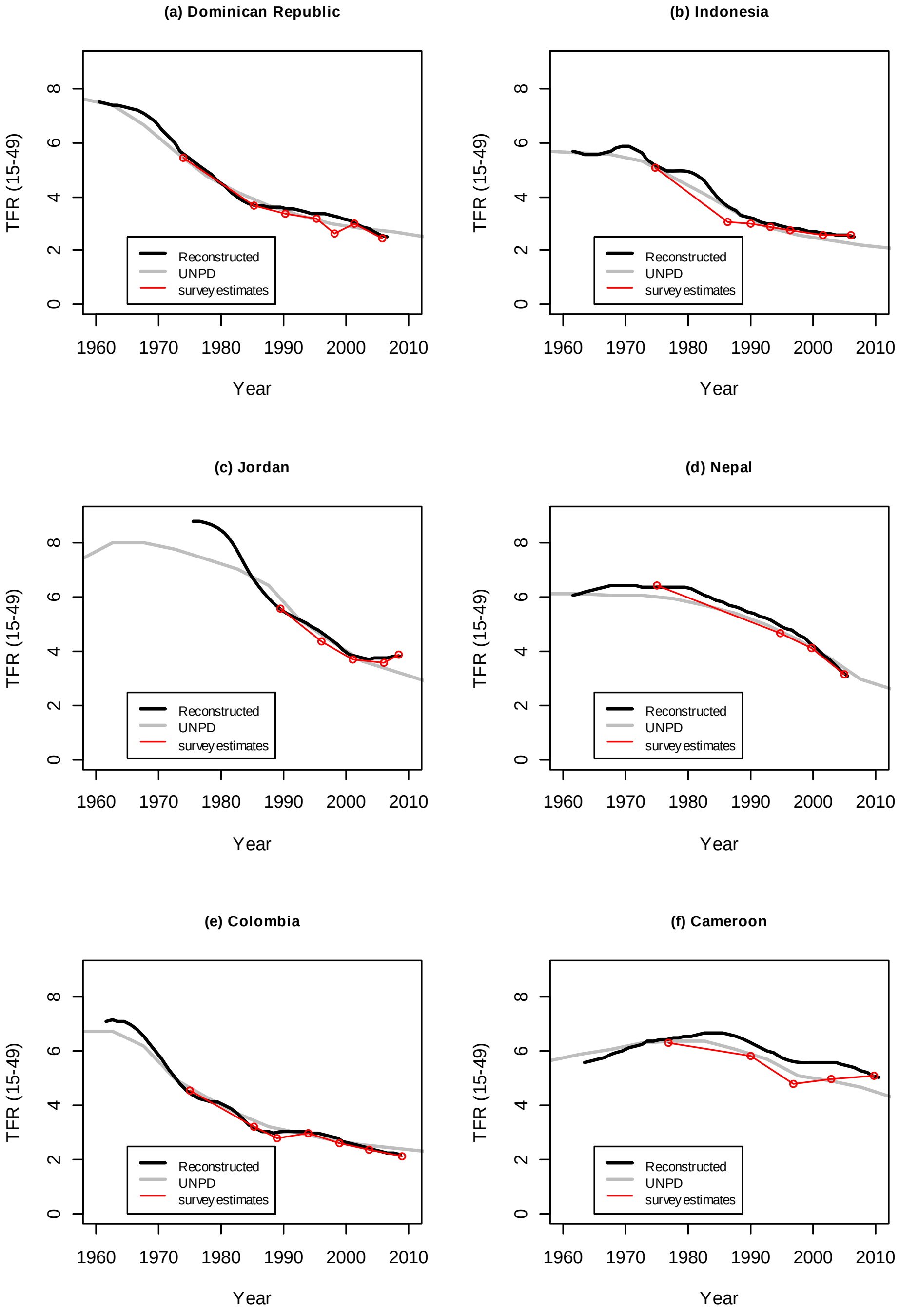
Figure 9: Reconstructed TFR (15-49) in the Dominican Republic, Indonesia, Jordan, Nepal, Colombia and Cameroon, with 95% confidence intervals, and identification of periods of no significant decline (orange).



Comparison with published fertility trends (Figure 10) show that results may differ largely across sources in some countries. Fertility trends inferred from estimates of recent fertility (3 years preceding the survey) cover a shorter period (red line, survey estimates on Figure 10). Moreover, estimates are lower than estimates obtained by pooling birth histories, and sometimes much lower (as in Cameroon). The discrepancy between recent estimates and reconstructed fertility trends reflect data quality issues for recent estimates. Recent fertility tends to be underestimated because of displacements or omissions of births. By pooling birth histories, fertility tends to be higher except for the most recent estimate (which is probably also underestimated in the pooled analysis).

Comparisons of reconstructed trends with the United Nations Population Division estimates also sometimes show serious discrepancies, as in Jordan or Cameroon. Moreover, the trends appear smoother in the UNPD estimates, masking some slowdowns and accelerations that are apparent in the reconstructed estimates. Whether the reconstructed estimates or the UNPD estimates are closer to the true values is difficult to ascertain. The UNPD estimates also include other sources of information (other surveys, censuses, age structures) that may lead to these differences.

Figure 10: Reconstructed TFR (15-49) in the Dominican Republic, Indonesia, Jordan, Nepal, Colombia and Cameroon, and comparison with estimates of the United Nations Population Division (*UNPD*), and trends inferred from successive surveys (*survey estimates*)



6. Correcting for data quality problems

The previous analyses mentioned potential data quality issues, for instance in Cameroon, but no specific treatment was applied. In many countries however, discrepancies across surveys may be large (Schoumaker, 2010), and reconstructing fertility trends from pooled birth histories requires a specific treatment. The discrepancies may be due to several factors, including:

- Differences in sample implementation.
- Displacements of births
- Omissions of births

In this section, I present two examples and discuss possible approaches for reconstructing fertility trends with data quality problems.

6.1 Differences in sample implementation: Morocco

Figure 11a shows reconstructed fertility trends in Morocco with four surveys (the 1980 WFS, the 1987 DHS, the 1992 DHS, and the 2003-2004 DHS). The three DHS match relatively well; in contrast, fertility in the first survey (WFS) seems to be underestimated. The figure suggests a relatively constant difference (about half a child) between the WFS and the DHS, maybe reflecting differences in sample implementation.

A possible approach to reconcile estimates and to reconstruct fertility trends is to consider that fertility is underestimated in the first survey, but that the trend is correct. This could reflect, for instance, the oversampling of women with lower fertility (e.g. educated, urban). Including in the model a dummy variable (SV1) equal to one for the WFS survey and 0 for the other surveys, and predicting the values of the TFR without taking into account the dummy variable provides adjusted estimates.

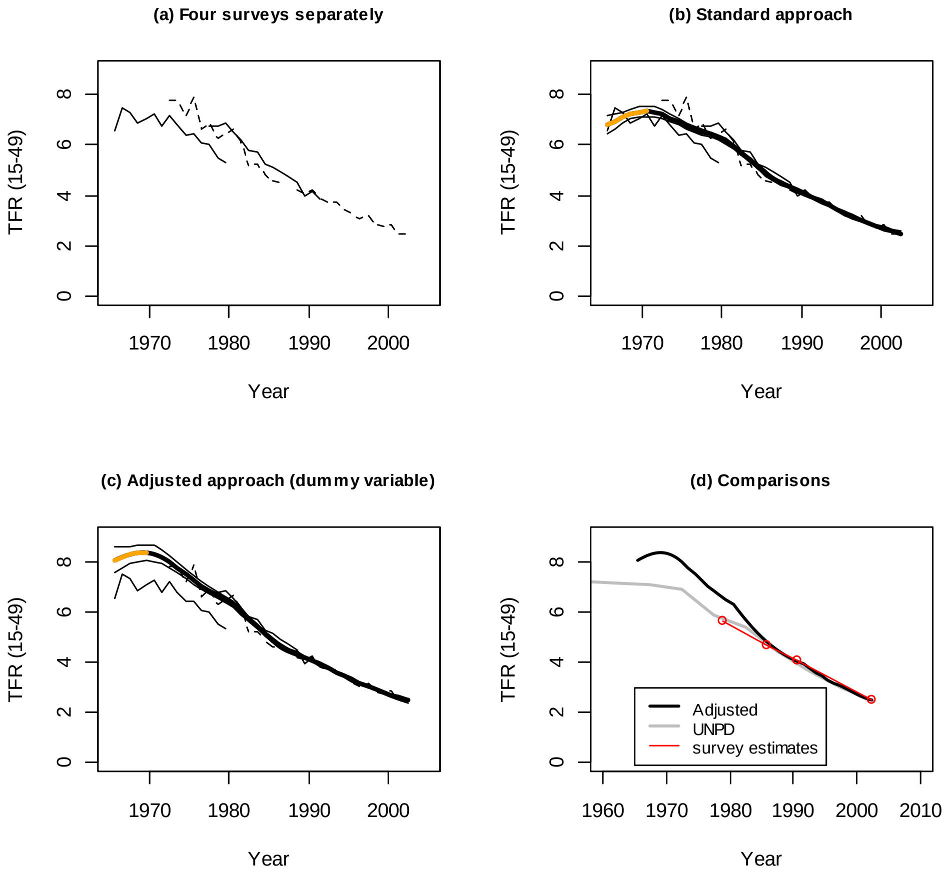
$$\log(\mu_{it}) = \log(e_{it}) + \log(a_i) + \gamma + \sum_{h=1}^{K-1} \delta_h.RCS_{hit} + \beta.SV1 \qquad \text{[Eq. 17]}$$

The TFR for year t is predicted with the following equation.

$$TFR_t = 5 * \exp[\gamma] * \exp\left[\sum_{h=1}^{K-1} \delta_h.RCS_{hit}\right] \qquad \text{[Eq. 18]}$$

Figure 11b shows reconstructed fertility trends by pooling birth histories without adjusting for underestimation in the first survey; Figure 11c shows adjusted fertility trends. According to the adjusted estimates, fertility was above 8 children per woman in Morocco in the late 1960s - early 1970s. This approximately one child higher than the level of fertility measured in the WFS, and the level of fertility reported in the UN Population Division estimates (Figure 10d).

Figure 11: Reconstructed TFR (15-49) in Morocco, with 95% confidence intervals with different approaches, and comparison with UNPD estimates.



6.2 Displacements and omissions of births: Cameroon

As discussed before, fertility trends in Cameroon may be affected by data quality problems. Figure 12a shows reconstructed fertility trends in Cameroon with five surveys (the 1980 WFS, the 1991 DHS, the 1998 DHS, the 2004 DHS and the 2011 DHS). The overall trend is relatively consistent across surveys, but recent fertility seems to be underestimated in several surveys (red circles). A possible source of underestimation of fertility a few years before the survey is the displacement and omissions of births by interviewers, in order to avoid the lengthy health questionnaire. Births will be displaced from the cut-off year of the health module (e.g. 1995) to the year just before (e.g. 1994), or the two years before (1993 and 1994), and some births between the cut-off year (1994) and the date of the survey may be omitted (Schoumaker, 2010). Several options are possible to treat this problem – but currently none is fully satisfying.

One option is to remove the years likely to be affected by displacements or omissions (for the concerned surveys) from the data set. In this example, cut-off years are respectively 1986, 1995, 1999 and 2005 for the 1991, 1998, 2004 and 2011 DHS. Data from 1984 to 1991 are removed from the first survey; data from 1993 to 1998 are removed from the second DHS, etc. (2 years before the cut-off year of the health module until the date of the survey). A drawback is that fertility is not estimated for the most recent period (last 6-7 years). Reconstructed fertility is shown on Figure 11b. It is slightly higher than with the standard approach.

Another option is to include dummy variables for the surveys and years affected by displacements and/or omissions, and to predict TFRs without including the dummy variables (Schoumaker, 2010). The dummy variables are coded in the following way (example for the first DHS, called cmir22fl).

DB1=1 if (year=1984 OR year=1985) AND surv="cmir22fl"; DB1=0 otherwise

DA1=1 if (year =1986 OR year=1987) AND surv=="cmir22fl"; DA1=0 otherwise

OM1=1 if year >=1986 AND surv=="cmir22fl"; OM1=0 otherwise

The model becomes

$$\log(\mu_{it}) = \log(e_{it}) + \log(a_i) + \gamma + \sum_{h=1}^{K-1} \delta_h.RCS_{hit} + \beta_1.DB1 + \beta_2.DB2 + \beta_3.DB3 + \beta_4.DB4 + \beta_5.DB1 ... + \beta_9.OM1 + ...$$

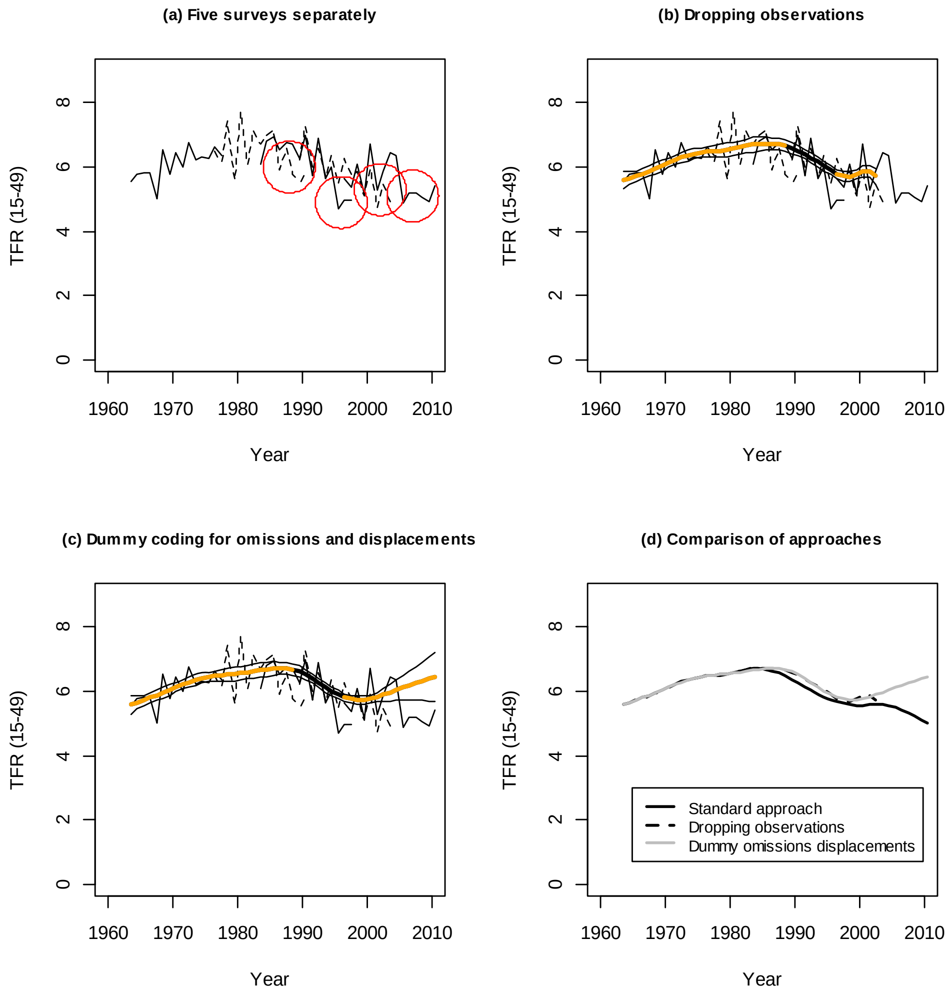
[Eq. 19]

The TFR for year t is, as in eq. 18, predicted without including the dummy variables.

$$TFR_t = 5 * \exp[\gamma] * \exp\left[\sum_{h=1}^{K-1} \delta_h.RCS_{hit}\right] \qquad \text{[Eq. 18]}$$

A drawback of this approach is that the recent estimates relying only on data of the last survey tend to be very unstable and unreliable. Figure 11d shows that recent fertility is much higher with that approach than with the standard approach. The true fertility levels probably lies somewhere between these two approaches. Further work is needed to reach reasonable estimates of fertility in such situations.

Figure 12: Reconstructed TFR (15-49) in Cameroon, with 95% confidence intervals with different approaches.



7. Conclusion

Poisson regression is used with pooled birth histories, controlling for the age pattern of fertility in the offset. The method is tested in countries with data of relatively good quality, and with simulated birth histories. These tests show the method performs well for reconstructing fertility trends over long periods (40-50 years), and provides more details than what is obtained by comparing recent estimates from consecutive surveys. The method also allows testing for stalls in a straightforward way. The method can also be used – to some extent – to correct for data quality problems. Further work is needed on these issues.

8. References

Alkema L., Raftery A., Gerland P., Clark S., Pelletier F. (2012), “Estimating Trends in the Total Fertility Rate with Uncertainty Using Imperfect Data. Examples from West Africa”, *Demographic Research*, 26(15), 331-362.

Andersen, R., 2009, “Nonparametric Methods for Modeling Nonlinearity in Regression Analysis”, *Annual Review of Sociology*, vol. 35, pp. 67-85.

Buis, Maarten L. 2009. "POSTRCSPLINE: Stata module containing post-estimation commands for models using a restricted cubic spline" http://ideas.repec.org/c/boc/bocode/s456928.html

Garenne M. and V. Joseph (2002), “The Timing of the Fertility Transition in Sub-Saharan Africa.” World Development, 30(10), 1835-1843.

Harell, F., 2001, Regression Modeling Strategies, Springer Verlag, Secausus, 568 p.

Machiyama K. (2010), A Re-examination of Recent Fertility Declines in Sub-Saharan Africa, *DHS Working Paper*, 68, ICF Macro, Calverton.

Machiyama K. and A. Sloggett, 2009, “Is Fertility Decline Stalling in Sub-Saharan Africa? Re-Examination of Fertility Trends”, *Meeting of the Population Association of America*, Detroit, May 2009.

Schoumaker B. (2013, forthcoming). “tfr2: A Stata module for computing fertility rates and TFRs from birth histories”, *Demographic Research.*

Schoumaker B. (2010), “Reconstructing Fertility Trends in Sub-Saharan Africa by Combining Multiple Surveys Affected by Data Quality Problems”, *Meeting of the Population Association of America*, Dallas.

United Nations, Department of Economic and Social Affairs, Population Division (2011). *World Fertility Report 2009,* United Nations, New York.