Reconstructing Fertility Trends in Sub-Saharan Africa by Combining Multiple Surveys Affected by Data Quality Problems

Bruno Schoumaker[1], Université catholique de Louvain

Draft (preliminary results) – April 9, 2010

1. Objectives

Fertility trends in sub-Saharan Africa have generated interesting debates. Since the early 2000s, situations of fertility stalls or reversals of fertility declines have been described and analyzed in several African countries. Kenya and Ghana were the first countries to be identified as experiencing a stall in fertility decline (Bongaarts, 2005), and Bongaarts' (2008) recent study on the progress of fertility transition in developing countries concluded that as many as 12 sub-Saharan African countries had recently experienced a stall. As a result, the overall pace of fertility decline in Africa is thought to have considerably slowed down in the second part of the 1990s and early 2000s (Bongaarts, 2008).

The debates about the speed of fertility transition in sub-Saharan Africa have important policy implications. Stalls in fertility transition have been linked to the slowing down of investments in family planning programs in several settings. Westoff and Cross (2006) suggested that the Kenyan stall may partly result from shortages of contraceptive supplies; Recently, Steven Sinding (2008), during an online interview organized by the Population Reference Bureau, stated about fertility stalls in sub-Saharan Africa "I don't think they are spurious and I think the cause is very clear: the redirecting of resources away from family planning and toward other (usually health-related) programs, most especially HIV/AIDS".

The measurement of the speed of fertility transition also has theoretical implications. For instance, the apparent plateauing of the fertility decline in some countries, despite economic progress, is at odds with the demographic transition theory. In contrast, some fertility declines appear to have been more rapid than expected from changes in

[1] Bruno.schoumaker@uclouvain.be

socioeconomic determinants. Such rapid fertility changes could be interpreted as evidence for diffusion effects.

Recent work (Ortega, 2008; Schoumaker, 2008; Machiyama and Slogett, 2009) suggest that the standard approach used to measure fertility trends – namely comparing published recent fertility estimates (over the three years preceding the survey) in several consecutive surveys – may be flawed. Omissions and displacements of births may lead to large underestimates in fertility levels, often by 15 %. Variations across surveys in the degree of underestimation of fertility may lead to identify spurious stalls, or to conclude that fertility declines when it is in fact stable. This data quality problem seems particularly pronounced in sub-Saharan Africa.

The objective of this paper is to estimate fertility levels and trends in sub-Saharan Africa using full birth histories from DHS, taking into account data quality problems. Pooling surveys together, fertility can be reconstructed over periods as long as 25 to 30 years in many African countries. The method relies on the organisation of the birth histories as a person period data file, which can be analyzed with Poisson regression (log rates models). Using age and calendar years as time-varying explanatory variables in the model, it is possible to estimate levels and trends of TFRs in a flexible way. The fertility trend is then smoothed using restricted cubic splines (Harell, 2001). Dummy variables are included in the models to obtain fertility trends corrected for omissions and displacement of births. Two approaches are used to estimate fertility trends in the last years – when estimates from only one survey are available. The methods are used for all the 23 sub-Saharan African countries with at least 2 available DHS surveys.

2. DATA

The analyses rely on data from the Demographic and Health Surveys conducted in Sub-Saharan Africa since the mid 1980s. We retain all the countries of Sub-Saharan Africa where at least two comparable surveys have been conducted, with data published on the STATcompiler website, and for which data files are available[2]. Overall, 23 countries (73 surveys) are included in this study.

[2] Only TFRs from 'Standard DHS surveys' are published on the STATcompiler website. Fertility rates from surveys such as AIS (eg. 2006 AIDS Indicator Survey in Côte d'Ivoire) are not published on the STATcompiler website and were not used in this paper. Although two DHS were conducted in Eritrea, they were not used because the individual data files are not available from Macro

Both published data and individual data files are used. Published data are taken from the STATcompiler website (www.measuredhs.com), and individual data files were obtained from Macro international. Published data are used to identify fertility levels and trends as they are most commonly used. The TFRs we use (published on STATcompiler) are measured in the three years preceding the survey. Individual data files are used to reconstruct fertility trends using birth histories. All the analyses used the sampling weights provided in the DHS data files.

3. METHOD

The approach we use to reconstruct fertility trends relies on the following methods/steps:

1. Poisson regression to compute Total fertility rates
2. Pooling data from several surveys together
3. Use of restricted cubic splines to smooth fertility trends
4. Correction of omissions and displacement of births with dummy variables
5. Estimating omissions in the last survey with two approaches

We first use data from the 4 DHS conducted Zimbabwe to illustrate the general approach to reconstructing fertility trends (from point 1 to 3), as the DHS in Zimbabwe are little affected by data quality problems. We illustrate the fourth and fifth point with data from Cameroon and Senegal. We next apply the method to the 23 countries.

3.1. Computation of TFRs with Poisson regression

In this section, we present the general approach used to compute trends in Total Fertility Rates using Poisson regression. We illustrate it with data from one survey (Zimbabwe 1999), to compute annual variations in TFRs. In the next sections, the same approach is used with pooled data, with restricted cubic splines to smooth fertility trends, and with additional independent variables to correct for omissions and displacements.

The first step to compute TFRs with Poisson regression is to transform the birth history into a person-period period data set (Schoumaker, 2004). The history of a woman is split into segments. Every time the age of the woman or year changes, a new segment is created. As

---

International. We also did not include Liberia, because the time interval between the two surveys (1986 and 2007) was too long to be meaningful in this study.

a result, segments have varying lengths (lower than 12 months). The length of the segments (exposure) is controlled in the models through an offset. In each segment, a woman either gives birth to one or several children, or does not give birth. The number of births in the segment is the dependant variable. The age and year in which the segment is located are used to create dummy variables, indicating age group and time period. These variables are used as explanatory variables.

The second step is to select observations above age 15, and (for instance) from 15 years before the survey. The next step is to estimate a Poisson regression model using the person period data. The model is of the following form:

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

μ_i is the expected number of children born in each time segment, t_i is the length of the time segment (exposure), f(age) is a function of age, and g(time) is a function of the calendar time. This approach assumes that there is no interaction between the age effect and the period effect (time); i.e. that the shape of the age-specific fertility rates is constant over time. Although the assumption does not strictly hold, simulations indicate that violating this assumption does not have a strong influence on fertility trends (Schoumaker, 2006).

The following model illustrates this method for reconstructing fertility over the 15 years preceding the 1999 DHS in Zimbabwe. Age is included as a set of dummy variables for five-year age groups (Table 1). The function of calendar time is first measured by a set of dummy variables to model annual variations in fertility. The regression coefficients of age groups are exponentiated, summed and multiplied by 5 to obtain the TFR for the reference year (1984, first year of the 15-year period). 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. Standard errors for the TFR are computed using the delta method[3]. Figure 1 shows the annual variations of the TFRs (same values as in Table 1), with 95% confidence intervals.

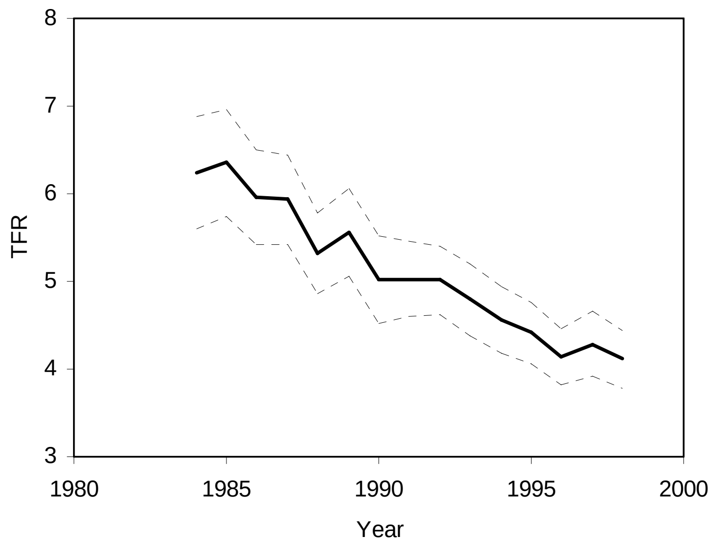
---

[3] This approach is implemented in Stata, with the *predictnl* command. That command allows computing predicted values and standard errors for any non linear combination of the regression coefficients. Standard errors also take account of the sample design (stratification, clustering, weighting), using Taylor linearization (*svy* commands in Stata).

Table 1: Age specific fertility rates and reconstruction of fertility trends over the fifteen years preceding the 1999 DHS survey in Zimbabwe. Poisson regression on person-period data.

Age groups	Regression coefficients (β)	Exp(β)		Year	Regression coefficients (β)	Exp(β)	Estimated TFR
15-19	-1.945	0.143		1984 (ref)	-	-	6.24
20-24	-1.259	0.284		1985	0.0179	1.018	6.35
25-29	-1.314	0.269		1986	-0.0459	0.955	5.96
30-34	-1.432	0.239		…	…	…	…
35-39	-1.674	0.188		1993	-0.2638	0.768	4.79
40-44	-2.270	0.103		…	…	…	…
45-49	-3.792	0.022		1998	-0.4154	0.660	4.12
1984 TFR		6.24					

Figure 1: Reconstruction of TFRs (15-49) and 95% CI over the 15 years preceding the survey (1984-1998), Zimbabwe 1999 DHS.



The main advantages of the Poisson regression approach are that:

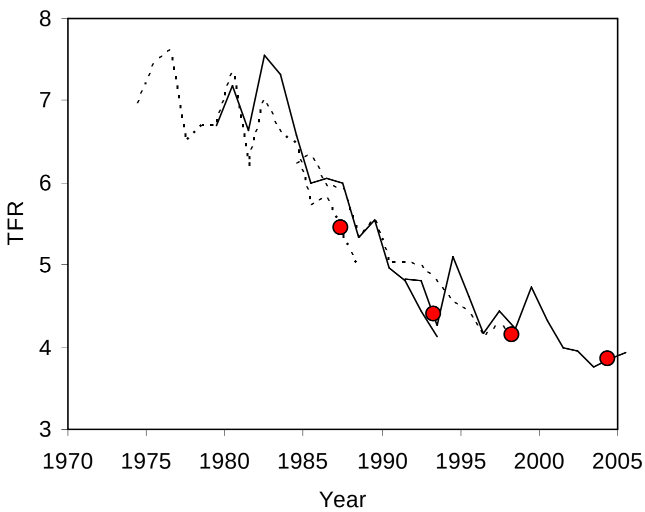
1. It allows computing TFRs between 15 and 49 over the 15 years preceding the survey. Since only women aged 15-49 were interviewed, fertility data is incomplete in the past among older women. For example, no data is available 15 years before the survey among women aged 34 and over. Using Poisson regression, and making the assumption that the shape of the age-specific fertility rates is constant, fertility rates can be predicted by the model at all the ages and for all the years.

2. Independent variables are easily included in the regression model. As explained later, years can be replaced by spline functions; dummy variables can be introduced to correct for omissions and displacements. In fact, any relevant independent variable can be included in the model.

3.2. Pooling several surveys together

Our approach to reconstructing fertility trends relies on pooling several surveys together[4]. This is based on the observation that - except for the periods affected by data quality problems (see section 3.4) – retrospective fertility estimates from consecutive surveys usually match quite well. Figure 2 shows the annual TFRS estimated from the four DHS surveys in Zimbabwe (1988, 1994, 1999, 2004), illustrating the fact that the estimates from consecutive are quite consistent (in this case). Figure 3 shows the TFRS and confidence intervals estimated with Poisson regression after pooling the four surveys together. A clear downward trend is apparent from the early 1980s.

Figure 2: Comparisons across four surveys of retrospective fertility trends (by single years) in Zimbabwe (Poisson regression on person-period data)

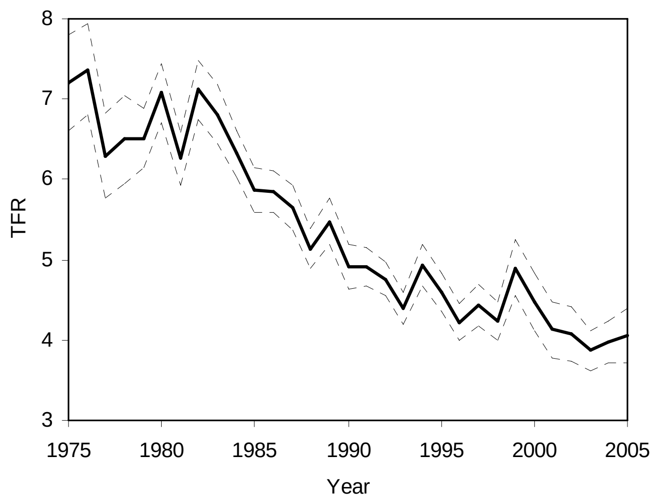


Solid lines and dotted lines are alternated to represent fertility trends from the four DHS. Large dots represent published values of TFRS.

---

[4] Original sampling weights are retained for each survey.

Figure 3: Annual variations of TFRs and 95% confidence intervals in Zimbabwe obtained by pooling data of four DHS (Poisson regression on person-period data)



The Solid line indicates the values of TFR; the dotted lines indicate the 95% confidence interval.

3.3. Use of restricted cubic splines to smooth fertility trends

Restricted cubic splines are used to reconstruct smooth (possibly non-linear) fertility trends. Generally speaking, “regression splines are piecewise polynomial functions that are constrained to join at points along the range of *x* called knots” (Andersen, 2009, p.70). Cubic splines are very flexible and allow fitting a large variety of shapes with relatively few parameters. Restricted cubic splines have the additional property of constraining the smoothing function to be linear in the tails, i.e. before the first knot and after the last knot (Harrell, 2001). This approach is useful when little or unreliable data is available in the tails. We will use this property to constrain the trend of fertility to be linear after the last knot, and to estimate recent fertility (see next section).

To fit restricted cubic splines with K knots, K-1 variables (functions of time periods) need to be created. The construction of these variables depends on the number and the location of knots[5]. The new variables are then introduced as explanatory variables in the Poisson regression model where the dependant variable is the number of births in the

---

[5] The *mkspline* command in Stata creates automatically these variables after the number and the location of knots have been defined (StataCorp, 2007).

period, controlling for age and exposure[6]. Predicted values of total fertility rates can then be 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 created to fit the restricted cubic splines.

The number and location of knots have to be defined before adjusting the restricted cubic splines. Several authors have shown that the smoothing functions are not very sensitive to the location of the knots (Harrell, 2001; Andersen, 2009), and are more sensitive to the number of knots. We have chosen to locate knots every five years, for two reasons. Locating knots every five years means that, in most countries, 4 or 5 knots will be used, and it has been shown that 4 to 5 knots are usually sufficient to reach a good compromise between flexibility and rigidity (Harell, 2001). Secondly, locating knots every five years allows identifying stalls in fertility transitions[7].

The location of knots every five years is done backward, starting from the last knot. As explained earlier, restricted cubic splines constrain the smoothing function to be linear after the last knot. The location of the last knot thus defines the year after which the trend in the logarithm of fertility is considered to be linear[8]. In the next section, the last knot will be located on the year just before the cut-off year of the health module in the latest survey. This means that the last portion of the restricted cubic spline is constrained to be linear after that point.

The figure below illustrates, for Zimbabwe, the trend obtained using restricted cubic splines with 6 knots (1979, 1984, 1989, 1994, 1999 and 2004)[9], as well as 95% confidence intervals. We also show (Figure annex 1) three curves with knots located every five years, but with different locations of knots. As is clear from these figures, the shape of the fertility trends is very well captured with restricted cubic splines located every five years. Figure annex 1 further indicates that the location of the knots matters very little.

---

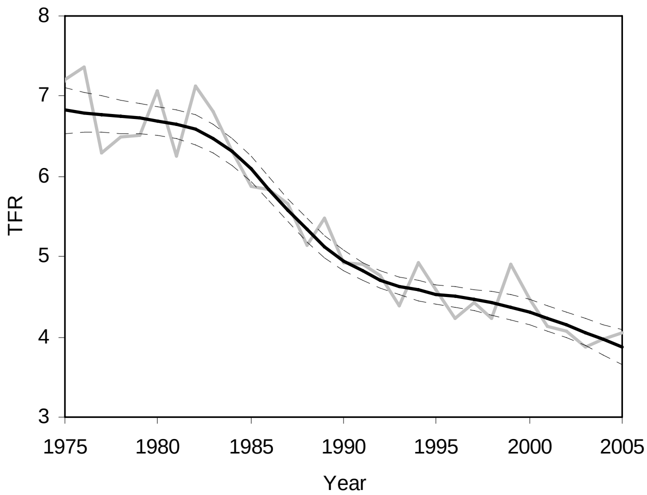
[6] We used a linear function of age and of the logarithm of age to control age. An offset is included in the model to control for differential exposure.

[7] Several authors suggest that a stall should be considered as such if it lasts for at least five years (Machiyama and Sloggett, 2009; Moultrie et al., 2008)

[8] The Poisson regression models the logarithm of the rates as a function of explanatory variable. A linear trend in the logarithm of rates means the trend of the TFR is exponential.

[9] In the next examples, the 2004 knot will be dropped

Figure 4: Fertility trends in Zimbabwe smoothed by restricted cubic splines with 6 knots located every five years, and 95% confidence intervals. Estimates obtained by pooling four DHS (Poisson regression on person-period data)



The grey solid line represents annual values of TFRs obtained with pooled data (Poisson regression with dummy variables for each year). The solid black line is the smoothed TFR using restricted cubic splines. Dotted lines represent limits of the 95 confidence interval.

3.4. Correcting fertility trends for omissions and displacement of births

Although in countries like Zimbabwe, TFRs from consecutive surveys match very well (Figure 2), most surveys in sub-Saharan African countries are affected by serious data quality problems (Schoumaker, 2008; Pullum, 2006). This is best illustrated with data from Mozambique and Cameroon. Figures 5a et 5b compare annual fertility trends estimated from two surveys in Mozambique and in Cameroon (for the sake of clarity, only two surveys are shown on these figures). Vertical lines are drawn to indicate the cut-off years for the health module in the surveys, and the large dark dots indicate the values of TFRs published on the statcompiler. As these examples show, the cut-off year of the health module correspond to a sharp decrease in fertility in three of the four surveys, which clearly suggests omissions and displacements of births[10]. This type of problems has been long known in DHS surveys (Institute for Resource Development, 1990). The displacement of births is notably linked to the fact that some interviewers can change the birth dates of certain children to avoid having to administer the lengthy health module in the DHS[11].

[10] The same pattern is observed in the large majority of surveys, where the cutoff-year corresponds to a steep fertility decline.

[11] The cut-off date for the health module is often January five years before the survey.

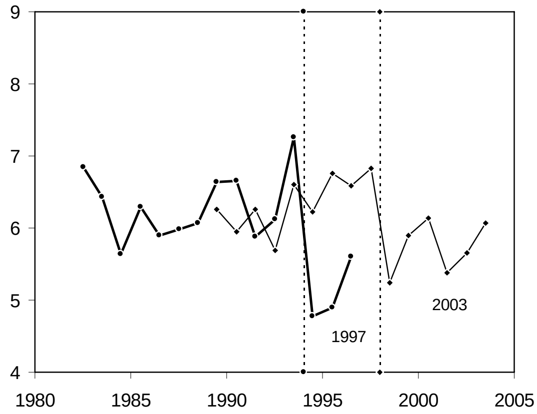
Displacements of births lead to underestimating recent fertility, and to overestimate past levels of fertility. Pullum (2006) has recently showed that this was still a serious problem in DHS in sub-Saharan Africa. *Omission* of recent births is another possible consequence of the desire to avoid the health module, but other factors may also explain omissions (e.g. if respondents are not willing to mention deceased children). The impact of omissions of recent births is to underestimate levels of fertility a few years before the survey.

Mozambique is a particularly striking example. Using the 1997 survey, the TFR was estimated at 6.2 children per woman in 1993 and at 4.7 in 1994. The sharp decrease corresponds with the cut-off date for the health module (January 1994). The same type of problem is found in the 2003 survey: The TFR was estimated at 7.1 children per woman in 1997, and at 5.5 in 1998 (the cut-off date for the health module was January 1998). The fertility increase one year before the cut-off year suggests there were birth displacements, but it seems from this figure that birth omission is the major issue. The Cameroon example shows that that extent of omissions and displacements may vary across surveys. In the 1998 survey, the sharp drop in fertility after the cut-off year, and the fact that fertility is much lower than the TFR estimated from the 2004 survey mean that there are probably large omissions (and maybe displacements). In contrast, omissions and displacements in the 2004 Cameroon survey seem much less pronounced. Although no other survey is available to estimate the extent of omissions precisely, there is no sudden break in the fertility trend at the cut-off year.

This graphical approach suggests that omissions and displacements are major data quality problems. Periods after the cut-off year for the health module are likely to be affected by massive omissions and birth displacements, leading to serious underestimates in fertility. However, these figures also suggest that - except for the periods affected by omissions and/or displacements – retrospective fertility estimates from consecutive surveys usually match quite well. This makes it realistic to pool surveys together to reconstruct fertility trends – provided omissions and displacements of births are taken care of.

Figure 5: Comparisons across surveys of retrospective fertility trends (by single years) in Mozambique and Cameroon (Poisson regression on person-period data)

(a) Mozambique



(b) Cameroon

3.4.1. Creating dummy variables to correct for omissions and displacements of births

For each survey included in the pooled data set, three dummy variables are computed. We illustrate this with three surveys (subscripts indicate the survey number, see Table 2). The first variable (O_1), is a variable capturing omissions after the cut-off year of the health module in the first survey. It is equal to zero, except for the years starting from the cut-off year of the health module until the last year covered by the survey, where it is equal to 1. The coefficient of this variable is expected to be negative, as it measures the ratio of fertility in periods and surveys affected by omissions compared to fertility levels in periods and surveys without omissions. The second variable (DB_1) is a dummy variable capturing displacements of births to the year before the cut-off year. It is equal to zero, except for the year just before the cut-off year of the health module, where it is equal to 1. Its coefficient is expected to be positive, since displacement of births will increase fertility just before the cut off year. The third variable (DA_1) is a dummy variable capturing displacements of births from the cut-off year. It is equal to zero, except for the cut-off year of the health module, where it is equal to 1. Its coefficient is expected to be negative, since displacement of births will decrease fertility on the cut off year. For the three surveys, nine dummy variables will then be created (O_1, O_2, O_3, DB_1, DB_2, DB_3, DA_1, DA_2, DA_3). The following box and table illustrate the computation of the dummy variables.

Box 1: Computation of dummy variables to correct for omissions and displacement of births

$O_{1t} = 1$ if $t \geq COY_1$ and $survey = 1$; otherwise $O_{1t} = 0$
$O_{2t} = 1$ if $t \geq COY_2$ and $survey = 2$; otherwise $O_{2t} = 0$
...
$DA_{1t} = 1$ if $t = COY_1$ and $survey = 1$; otherwise $DA_{1t} = 0$
$DA_{2t} = 1$ if $t = COY_1$ and $survey = 2$; otherwise $DA_{1t} = 0$
...
$DB_{1t} = 1$ if $t = COY_1$ - 1and $survey = 1$; otherwise $DB_{1t} = 0$
$DB_{2t} = 1$ if $t = COY_1$ - 1and $survey = 2$; otherwise $DB_{1t} = 0$
...

COY=Cut-off year of the health module
T=year

The approach consists in including dummy variables in the regression model to capture birth displacements and/or omissions. [Eq. 2] describes the model that is adjusted.

$$\log(\mu_i) = \log(t_i) + f(age) + g(time) + \beta_1.O_1 + \beta_2.O_2 + \beta_3.O_3 + \beta_4.DB_1 + ... \qquad \text{[Eq. 2]}$$

TFRs are then predicted using the coefficients estimated with this regression and setting the values of the dummy variables to zero. In other words, TFRs are predicted only as a function of age and time.

The major issue in reconstructing fertility trends lies in the estimation of the level of fertility in the last survey for the period where data is unreliable (for instance, from 1998 in Mozambique – Figure 5a). Since no other survey is available for that period, the coefficients of the omission variable does not measure omissions compared to fertility in a previous survey, but compared to an estimated fertility level, based on past trends and fertility trend in the period affected by omissions. This approach relies on the assumption that omissions are constant over the period starting from one year after the cut-off year. This means that the overall fertility trend is influenced by the fertility trend during that period.

Table 2: Illustration of dummy variable coding for the measurement of omissions and displacement of births.

	First survey Cut-off year : 1990			Second survey Cut-off year : 1996			Third survey Cut-off year : 2001		
Year	O1	DB1	DA1	O2	DB2	DA2	O3	DB2	DA2
1982	0	0	0	0	0	0	0	0	0
1983	0	0	0	0	0	0	0	0	0
1984	0	0	0	0	0	0	0	0	0
1985	0	0	0	0	0	0	0	0	0
1986	0	0	0	0	0	0	0	0	0
1987	0	0	0	0	0	0	0	0	0
1988	0	0	0	0	0	0	0	0	0
1989	0	1	0	0	0	0	0	0	0
1990	1	0	1	0	0	0	0	0	0
1991	1	0	0	0	0	0	0	0	0
1992	1	0	0	0	0	0	0	0	0
1993	1	0	0	0	0	0	0	0	0
1994	1	0	0	0	0	0	0	0	0
1995	1	0	0	0	1	0	0	0	0
1996	0	0	0	1	0	1	0	0	0
1997	0	0	0	1	0	0	0	0	0
1998	0	0	0	1	0	0	0	0	0
1999	0	0	0	1	0	0	0	0	0
2000	0	0	0	1	0	0	0	1	0
2001	0	0	0	0	0	0	1	0	1
2002	0	0	0	0	0	0	1	0	0
2003	0	0	0	0	0	0	1	0	0
2004	0	0	0	0	0	0	1	0	0
2005	0	0	0	0	0	0	1	0	0
2006	0	0	0	0	0	0	1	0	0

We test two approaches to reconstruct fertility trends. In the first approach –Poisson regression models with dummy variables are estimated, without any constraint. The second approach relies on constraining the omission variable in the last survey. Instead of estimating its coefficient as part of the regression model, we constrain it to be equal a certain value. To take account of the uncertainty of the constraint, we actually sample constraints from a normal distribution and combine models estimated with different constraints.

3.4.2. Unconstrained estimates

The first series of model provide unconstrained estimates for each of the parameters. We illustrate this approach with two countries: Cameroon and Senegal.

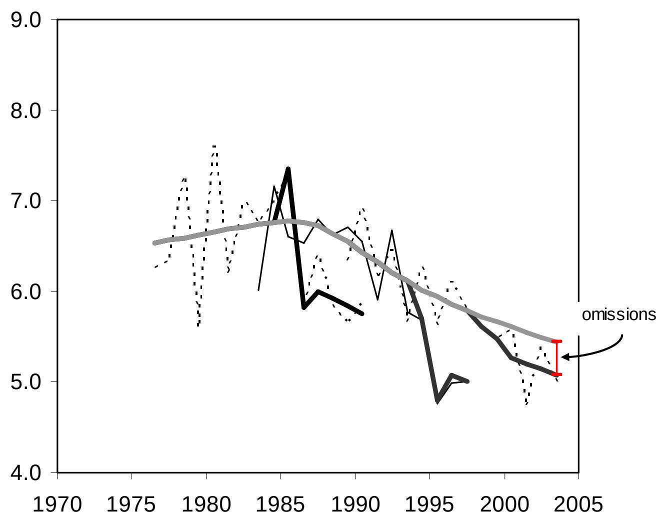
Regression coefficients of the dummy variables for Cameroon are presented in Table 3. As expected, the coefficients of the Omissions variables (O1, O2, O3) are all negative, indicating omissions. The third coefficient is closer to zero (and not significant) and suggests that omissions are less pronounced in the third survey than in previous ones. Most of the “displacement” dummy variables are not statistically significant. For the first survey, DB1 is significant and as expected positive. The coefficients of the other variables are not always of the expected sign, but none of them are significant. In short, there is no clear evidence of transfers of births before the cut-off year in the second and third survey.

Table 3: Regression coefficients (and SE, p-values, and exponentials) of the omission and displacement dummy variables, Cameroon DHS (3 surveys: 1991, 1998, 2004)

Year	Beta	SE	p-value	Exp(B)
O1	-0.114	0.020	0.000	0.892
O2	-0.144	0.032	0.000	0.866
O3	-0.056	0.036	0.123	0.946
DB1	0.082	0.036	0.025	1.085
DB2	-0.053	0.042	0.208	0.948
DB3	-0.018	0.034	0.599	0.982
DA1	-0.035	0.046	0.451	0.966
DA2	-0.068	0.044	0.120	0.934
DA3	0.026	0.033	0.424	1.026

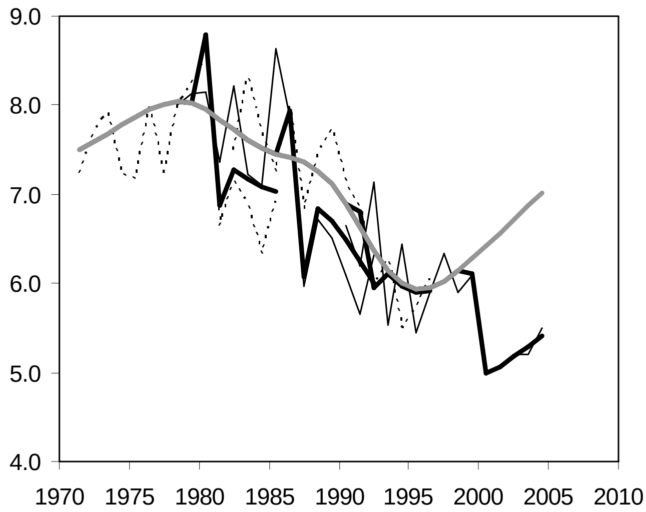
Figure 6 illustrates graphically the results of this approach in Cameroon. On this figure, we represent (1) predicted values of Total fertility rates from the regression model with all the variables included in the prediction, and (2) predicted values when the variables for omissions and displacements are set to zero. The solid grey line shows predicted values with variables for omissions and displacements set to zero. This is the corrected fertility trend. The dark lines indicate predicted TFRs for each survey when dummy variables are taken into account. The ratio between the dark line and the grey line indicate the extent of omissions (equal to the exponential of beta coefficients of dummy variables).

Figure 6: Comparisons of corrected fertility trend and predicted fertility trends with omissions and displacements in Cameroon (Poisson regression on person-period data)



Figures annex 2a to 2w (left column) show the reconstructed fertility trends using this method for the 23 sub-Saharan African countries. The large majority of the trends look plausible. However, for a few countries, the reconstructed trends look implausible.

Figure 7: Comparisons of corrected fertility trend and predicted fertility trends with omissions and displacements in Senegal (Poisson regression on person-period data)



Senegal is such a country (Figure 7). Reconstructed trends suggest that fertility started to increase in the mid 1990s, and that omissions in the last survey are massive (25 %). This surprising result stems from the upward trend after the cut-off year (2000) and the slight upward trend before the cut-off year. Controlling for omissions, the overall trend is a mixture of the trends before and after the cut-off year. Although this upward trend is not impossible, it looks highly implausible. The second approach we test tries to limit this type of problems.

3.4.3. Constrained estimates

The second approach is very similar to the first approach. The major difference lies in the way the omission parameter is estimated for the last survey. In the first approach, it is estimated as part of the regression model (regression coefficient of the last dummy variable measuring omissions). In the second approach, the omission parameter (regression coefficient) is constrained to lie in a plausible interval.

This approach relies on the following principle.

1) First, likely values for the omission parameter in the last survey are predicted. For each country, we consider the omission parameter in the last survey to follow a normal distribution, from which random values are selected.

2) 500 random values of the omissions parameter are selected from the normal distribution, and 500 models are estimated for one country[12]. In each model, the omission parameter in the last survey is constrained to be equal to a value randomly selected from the normal distribution.

3) Results of the 500 models are combined to obtain one series of regression coefficients and one variance-covariance matrix[13]. The coefficients of regression are computed as the average of the coefficients of the 500 models. The variance-covariance matrix is computed as the sum of two components: (1) the average of the variance-covariance matrices of the 500 models, and (2) the variance-covariance matrix of the coefficients of regression across the 500 models. Predicted total fertility rates and standard errors are computed from the new vector of coefficients and the new variance-covariance matrix.

Prediction of omission parameters

The major challenge rests in the prediction of reasonable values of the omission parameter for the last survey, i.e. with small bias and with reasonably small variance. The prediction of the omission parameter can be based on information from the last survey (survey-specific information), and on country-specific information.

In this paper, values of the omission parameters were predicted using an OLS regression model.

1) The dependant variable of the regression model is the series of omission parameters (we call them *O*) in all the surveys except the last one (50 surveys out of 73). These omission parameters are computed with models similar to those in section 3.4.2 (restricted cubic splines with dummy variables), but with data limited to the years before the cut-off year of the last survey. These omission parameters are considered reliable, as data from two surveys are available to estimate them.

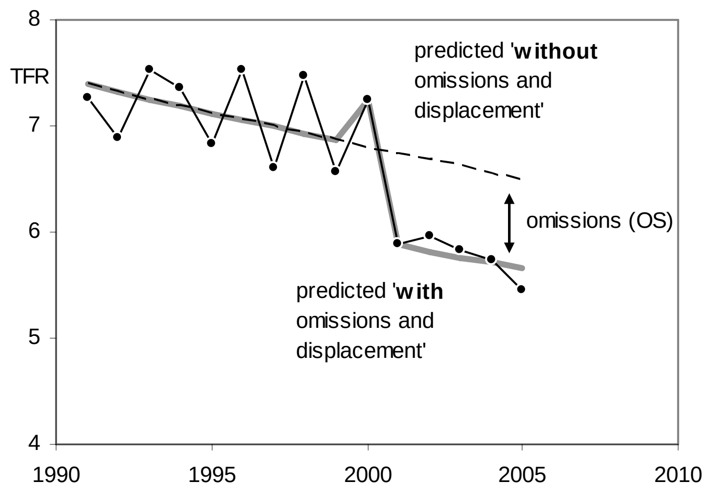
2) For each of the 73 surveys separately, we estimate a parameter that can be obtained from a single survey and is likely to be a good predictor of omissions. The parameter is obtained by fitting a Poisson regression model, with the same dummy variables as explained in section 3.4.1, and a linear trend. The omission variable measures the

[12] The number of models was limited to 500 at this stage because of time constraints. A larger number of models can be estimated, leading to more precise estimates.

[13] This done in the same way as in multiple imputation (Allison, 2001).

sudden drop in fertility after the cut-off year of the health module (it is expressed as a relative drop). Figure 8 shows the fertility trends obtained using this method for the 2001 survey in Benin. The omission variable obtained in this way is used as to predict omissions in the last survey (we call this variable OS).

Figure 8: Predicted fertility trends in Benin (2001 DHS survey) with and without omission and displacements of births, and measurement of OS.



3) Next, a regression model is estimated to predict the value of the 50 omission parameters (*O*) from the pooled data set, using the omission parameters computed from the single surveys (OS). The regression coefficients can then be used to predict values of omissions parameters (*O*) in the 23 last surveys, based on the value of OS in those surveys.

As shown on figure 9, there is a positive relationship between the omissions parameter estimated from several surveys and the omission parameter estimated from a single survey. The best fitting regression model has two independent variables: the mean of the omission parameters (OS) in a country and the deviation from the country mean of the omission parameter (OS) in a specific survey. The first variable (highly-significant) captures a country-specific factor. It indicates a strong positive relationship between the mean level of omissions measured in separate surveys (mean of OS), and the value of the omission in one survey measured from the pooled data set (O). The other variable measures the deviation from the mean of OS. It is as expected positively related to O (but not statistically significant). It means an increase in the deviation from the mean of the OS variable is associated with an increase in O.

Figure 8: Relationship between omissions parameters estimated from single surveys (OS, X-axis) and omission parameters estimated from pooled surveys (O, y-axis), 50 DHS from 23 sub-Saharan African countries.

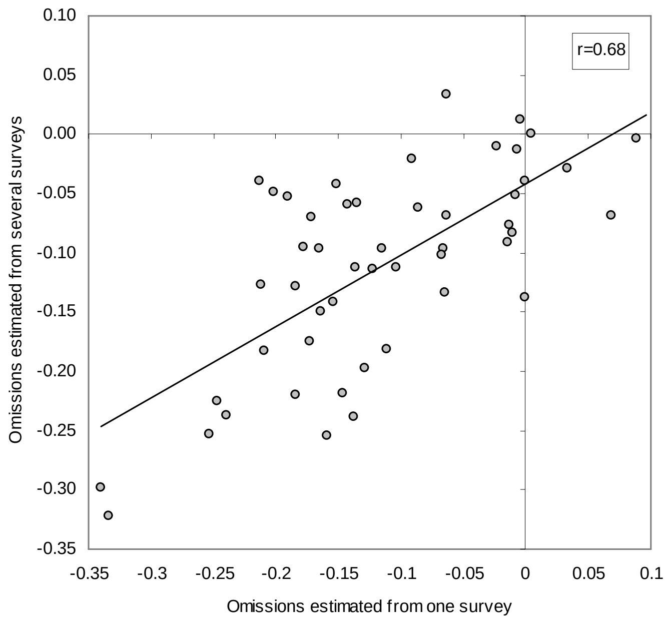


Table 8: Regression of omission parameters (O) on the mean of omissions parameters estimated from single surveys (Mean of OS) and the deviation from the mean in each survey, 50 DHS from 23 sub-Saharan African countries.

Variable	Beta	SE	p-value
Mean of OS	0.905	0.101	0.000
OS-Mean(OS)	0.195	0.117	0.103
R^2 = 0.635; N=50			

Using the regression coefficients in Table 8, values of the *O* parameter for the most recent surveys in the 23 countries are predicted, as well as confidence intervals for the individual forecast (Table annex 1). For each country, 500 values of the omission parameter are selected from a normal distribution with the mean equal to the predicted score and the standard deviation of the individual forecast. The 500 models are fitted with different values for the omission parameter, and combined to predict TFRs and confidence intervals.

4. RESULTS

Figures annex 2a to 2w (left column) and 3a to 3w (right column) show the reconstructed fertility trends using the first and the second approach for 23 sub-Saharan African countries. Each figure contains several pieces of information.

- The annual TFRs computed from each survey separately. These are shown on the graphs to compare annual (uncorrected) TFRs to the reconstructed trends, and to illustrate data quality problems (such as a steep drop in total fertility rates after the cut-off year reflecting omissions and displacements)[14].

- The smooth central curve indicates the reconstructed fertility trend. The statistical significance of the fertility decline is illustrated by a different colour. Years where fertility decreases significantly (marginal effect significantly lower than zero) are represented in black, while portions of the curves that are in lighter grey (orange on colour printers) indicate that fertility is either stable or increasing. The 90% confidence interval for the TFRs is also shown on these figures (black smooth lines above and below the reconstructed trends).
- Finally, values of the TFRs published on the STATcompiler website are also shown on these figures (dark dots). This allows comparing reconstructed fertility trends and fertility levels and trends inferred from published data.

For many countries, reconstructed trends from both approaches are quite similar, although confidence intervals are wider for the second method. In a few countries, results are very different. We start by discussing results that are similar with both approaches, and discuss differences between the approaches in the next section.

Conclusions from both approaches

- Published **fertility levels** are largely underestimated in most surveys. Differences between published TFRs and estimated TFRs are often above 1 child (e.g. in Chad, Côte d'Ivoire, Ethiopia, Guinea, Mali…).
- Reconstructed **fertility trends** may be quite different from trends inferred from published data. For instance, published values of TFRs indicate a stall in fertility in Ghana between 1998 and 2003; reconstructed fertility trends indicate no such stall. In contrast, our models suggest fertility may have stalled in Namibia and in Senegal, while published TFRs suggest a continuous fertility decline.
- Fertility levels and trends are affected by a **large uncertainty**. Confidence intervals for recent fertility levels are often very large (between 0.5 and 1 child).
- Fertility declines seem to have started in the 1970s in a few countries (Ghana, Kenya, Madagascar, Namibia, Tanzania, Togo, Zambia), in the 1980s in a few more (Cameroon, Nigeria, Senegal, Zimbabwe). Half of the countries either started their fertility

[14] To improve readability, the trends from consecutive surveys are represented alternatively by solid and dotted lines.

transition in the 1990s (Uganda, Burkina Faso…) or do not show clear signs of fertility declines (Benin, Chad, Guinea, Mali, Mozambique, Niger)

- Fertility declines are not linear processes. Although some non-linearities may be due to data quality problems, reconstructed trends show fluctuations (Niger), slowing downs (Zimbabwe, Kenya in the 1990s), stalls (Nigeria), accelerations (Kenya in the 1980s)…

Differences between the two approaches

- Results from the second approach look in general more plausible (mean TFRs), especially in countries were fertility was increasing implausibly (e.g. Senegal, Malawi).
- However, using the second approach (500 models with randomly selected values of omissions) leads to larger confidence intervals for the levels of TFRs. This is due to the relatively large uncertainty in the prediction of the omission parameters, which leads to large variations of the coefficients of regression across models.
- Because of the greater confidence intervals with the second approach, many recent fertility declines are not statistically significant. In Tanzania for instance, fertility trends is not significantly negative according to the second approach, while the first approach indicates fertility is decreasing in a significant way (similar cases are Burkina Faso, Cameroon, Côte d'Ivoire, Rwanda).

5. CONCLUSIONS

The proposed method allows reconstructing fertility trends over periods of 20 to 30 years by pooling several surveys together. Results from Zimbabwe show that, when data are little affected by data quality problems, smooth and reliable trends can be obtained.

However, most surveys in sub-Saharan Africa are affected by serious data quality problems (omissions and birth displacements). In such cases, corrections for birth omissions and displacements are necessary. When two surveys overlap, omissions and displacements can be accounted for in a relatively straightforward way by including appropriate dummy variables. The major challenge lies in the estimation of omissions (and hence fertility levels) in the most recent surveys. Two approaches were tested to correct for omissions and displacements in the most recent survey, and to reconstruct fertility trends. They lead to similar results for past trends; but recent trends differ in some cases, and are also affected by large uncertainty.

Although our estimates are not foolproof, they indicate that published fertility levels and trends from DHS should be interpreted very cautiously. Published fertility seems underestimated in the large majority of DHS in sub-Saharan Africa (especially in countries like Mozambique, Benin, Mali, and Niger…), and published trends differ from reconstructed trends in many countries. DHS estimates may also give a false impression of precision, because only sampling errors are reported; while measurement errors are a very important issue.

Further research is necessary to improve reconstructed fertility trends. One way would be to improve the prediction of the omission parameters used in the second approach. This could be done by using information on the proximate determinants (comparing observed fertility and predicted fertility based on proximate determinants in one survey) in the regression to predict omissions. Other information reflecting data quality (e.g. heaping on age in household survey) or likely to affect data quality (e.g. education) could be included in the model to improve the prediction of omissions.

The reconstruction of fertility trends could also be further refined (and validated) by using micro-simulations. Micro-simulation can be used to generate birth histories affected by data quality problems, and the methods can be applied to the distorted birth histories. This would allow identifying more clearly the impact of the assumptions of the model on the trends.

Finally, additional surveys may also prove useful. In this paper, only standard DHS were used. Other surveys such as the World Fertility Surveys, Interim DHS and AIS could be combined to the DHS to improve the reconstructed trends.

6. References

Allison, P., 2001, *Missing Data*, Quantitative applications in the Social Sciences, Sage, Thousand Oaks.

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

Bongaarts J., 2005, “The causes of stalling transitions”, *Studies in Family Planning*, vol. 37, n°1, p. 1-16.

Bongaarts J., 2008, “Fertility Transitions in Developing Countries: Progress or Stagnation?”, *Studies in Family Planning*, vol. 39, n°2, p. 105-110.

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

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

Institute for Resource Development, 1990, *An assessment of DHS-I data quality*, IRD / Macro Systems Inc., Calverton, 143 p.

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.

Moultrie T., V. Hosegood, N. McGrath, C. Hill, K. Herbst, M.-L. Newell, 2008, "Refining the Criteria for Stalled Fertility Declines: An Application to Rural KwaZulu-Natal, South Africa, 1990-2005", *Studies in Family Planning*, vol. 39, n°1, p. 39-48.

Pullum, T., 2006, *An assessment of age and date reporting in the DHS Surveys, 1985-2003*, DHS Methodological Reports, n°5, Macro International, Calverton, 93 p.

Ortega, J.A., 2009, "The fertility transition in sub-Saharan Africa: Early transition rise and (slow) decline", *Meeting of the Population Association of America*, Detroit, May 2009.

Schoumaker B., 2004, "A person-period approach to analyzing birth histories", *Population-E*, vol. 59, n°5, p. 689-702.

Schoumaker B., 2006, "The reconstruction of fertility trends with DHS birth histories. Application of Poisson regression to person-period data", *Conference of the European Association for Population Studies*, Liverpool.

Schoumaker B., 2008, "Stalls and reversals in fertility transitions in Sub-Saharan Africa. Real or Spurious ?", IUSSP Seminar on Human Fertility in Africa, Cape Coast, Ghana, September 2008.

Shapiro D., T. Gebreselassie, 2007, "Fertility Transition in Sub-Saharan Africa: Falling and Stalling", *Meeting of the Population Association of America*, New York, March 2007.

Sinding, S., 2008, "Will Africa's fertility decline ?", *Transcripts of online interview (10 December 2008)*, Population Reference Bureau, Washington DC (http://discuss.prb.org/ content/interview/detail/3027, consulted on 8 January 2009).

StataCorp, 2007, *Stata Statistical Software: Release 10*. StataCorp LP, College Station.

Westoff C., A. Cross, 2006, *The Stall in the Fertility Transition in Kenya*, Macro International, Calverton, 46 p.

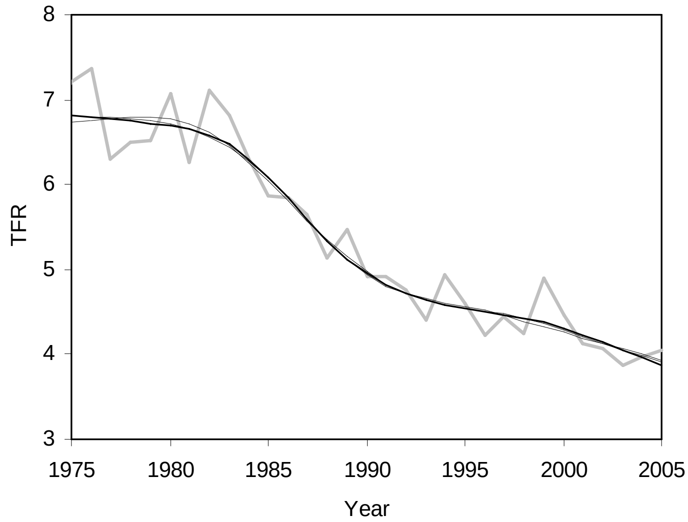
Table annex 1: Predicted values of omission parameters and 95% confidence intervals of individual forecasts

Year	Country	Predicted score for O	lower boundary (95% CI)	Upper boundary (95% CI)
1996	Benin	-0.177	-0.281	-0.072
2001	Benin	-0.168	-0.273	-0.063
2006	Benin	-0.189	-0.294	-0.083
1993	Burkina	-0.136	-0.239	-0.032
1998	Burkina	-0.132	-0.236	-0.028
2003	Burkina	-0.135	-0.239	-0.032
1991	Cameroon	-0.134	-0.238	-0.030
1998	Cameroon	-0.136	-0.240	-0.032
2004	Cameroon	-0.107	-0.213	-0.002
1996	Chad	-0.142	-0.247	-0.038
2004	Chad	-0.165	-0.269	-0.060
2000	Ethiopia	-0.156	-0.264	-0.047
2005	Ethiopia	-0.205	-0.313	-0.096
1988	Ghana	-0.052	-0.157	0.053
1993	Ghana	-0.064	-0.168	0.040
1998	Ghana	-0.078	-0.183	0.028
2003	Ghana	-0.054	-0.158	0.051
2008	Ghana	-0.068	-0.172	0.036
1999	Guinea	-0.256	-0.365	-0.148
2005	Guinea	-0.271	-0.380	-0.161
1989	Kenya	-0.004	-0.110	0.103
1993	Kenya	-0.032	-0.144	0.080
1998	Kenya	0.015	-0.093	0.123
2003	Kenya	0.016	-0.092	0.125
1992	Madagascar	-0.073	-0.178	0.031
1997	Madagascar	-0.081	-0.185	0.022
2003	Madagascar	-0.102	-0.208	0.003
1992	Malawi	-0.121	-0.225	-0.016
2000	Malawi	-0.101	-0.205	0.003
2004	Malawi	-0.110	-0.213	-0.006
1987	Mali	-0.248	-0.356	-0.139
1995	Mali	-0.232	-0.339	-0.125
2001	Mali	-0.229	-0.336	-0.123
2006	Mali	-0.214	-0.322	-0.106
1997	Mozambique	-0.153	-0.256	-0.049
2003	Mozambique	-0.154	-0.257	-0.050
1992	Namibia	-0.065	-0.170	0.040
2000	Namibia	-0.104	-0.212	0.003
2006	Namibia	-0.073	-0.177	0.031
1992	Niger	-0.192	-0.297	-0.087
1998	Niger	-0.172	-0.279	-0.066
2003	Niger	-0.206	-0.312	-0.099
1990	Nigeria	-0.165	-0.270	-0.061
1999	Nigeria	-0.165	-0.269	-0.060
2003	Nigeria	-0.148	-0.252	-0.043
2008	Nigeria	-0.146	-0.250	-0.041
1994	RCI	-0.264	-0.374	-0.155
1998	RCI	-0.236	-0.345	-0.127

Table annex 1: Predicted values of omission parameters and 95% prediction intervals

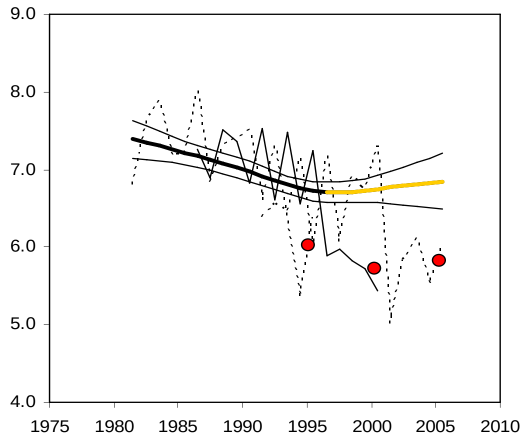
Year	Country	Predicted score for O	lower boundary (95% CI)	Upper boundary (95% CI)
1992	Rwanda	-0.098	-0.205	0.009
2000	Rwanda	-0.055	-0.162	0.052
2005	Rwanda	-0.075	-0.179	0.029
1986	Senegal	-0.147	-0.251	-0.044
1992	Senegal	-0.145	-0.249	-0.041
1997	Senegal	-0.146	-0.250	-0.042
2005	Senegal	-0.129	-0.234	-0.024
1991	Tanzania	-0.086	-0.189	0.018
1996	Tanzania	-0.095	-0.200	0.009
1999	Tanzania	-0.090	-0.194	0.014
2004	Tanzania	-0.066	-0.172	0.040
1988	Togo	-0.107	-0.211	-0.003
1998	Togo	-0.119	-0.223	-0.016
1988	Uganda	-0.030	-0.135	0.075
1995	Uganda	-0.032	-0.137	0.072
2000	Uganda	-0.032	-0.136	0.073
2006	Uganda	-0.049	-0.154	0.057
1992	Zambia	-0.055	-0.159	0.050
1996	Zambia	-0.029	-0.137	0.079
2001	Zambia	-0.068	-0.174	0.038
2007	Zambia	-0.055	-0.159	0.049
1988	Zimbabwe	-0.038	-0.143	0.067
1994	Zimbabwe	-0.069	-0.176	0.038
1999	Zimbabwe	-0.041	-0.145	0.064
2004	Zimbabwe	-0.043	-0.148	0.061

Figure Annex 1: Adjustment of restricted cubic splines to fertility trends in Zimbabwe with different knot locations (6 knots located every five years; last knot located in 2000, 2002 or 2004)

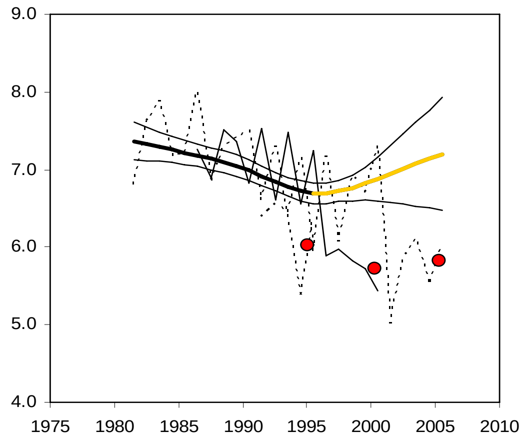


Figures Annex 2a to 2w and 3a to 3w: Adjustment of restricted cubic splines to fertility trends in 23 sub-Saharan African countries. Comparisons of results from the first approach (unconstrained, 2a to 2w) and second approach (combination of 500 models, 3a to 3w).

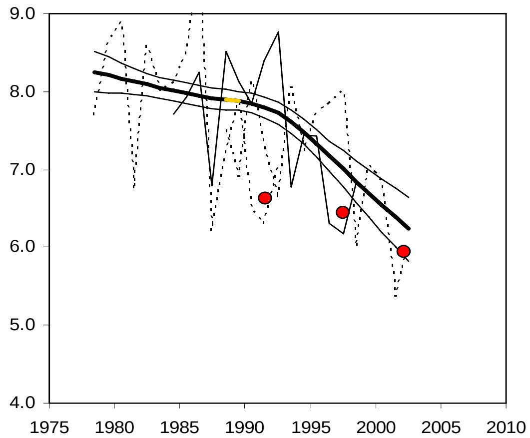
2a - Benin - unconstrained



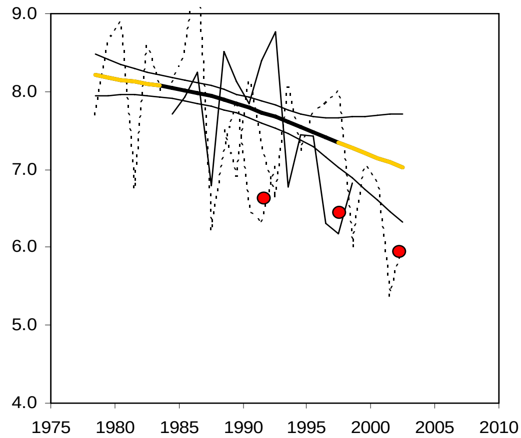
3a - Benin - 500 models



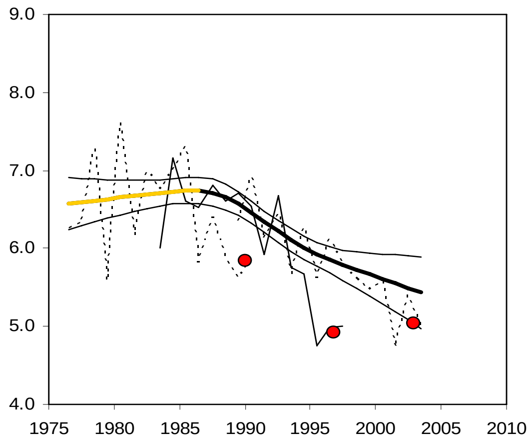
2b - Burkina Faso - unconstrained



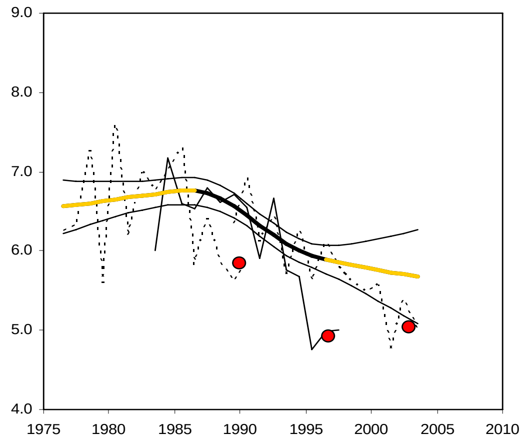
3b - Burkina Faso - 500 models



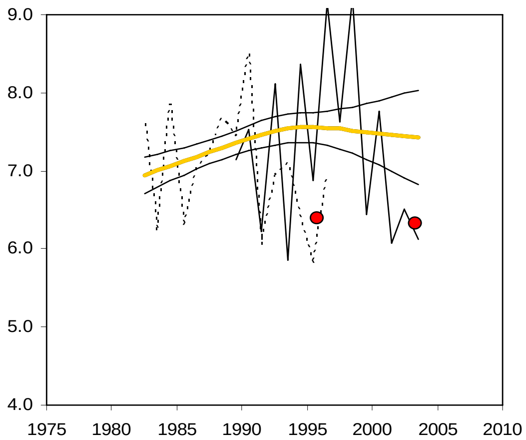
2c - Cameroon - unconstrained



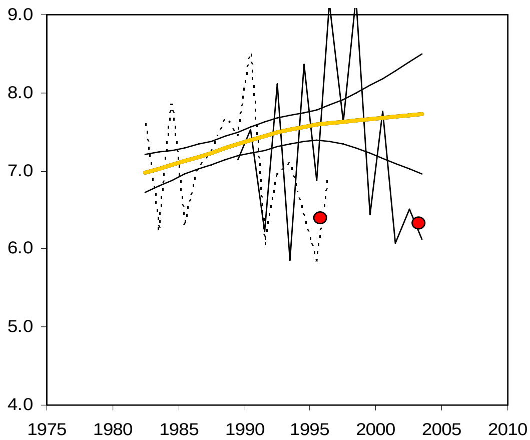
3c - Cameroon – 500 models



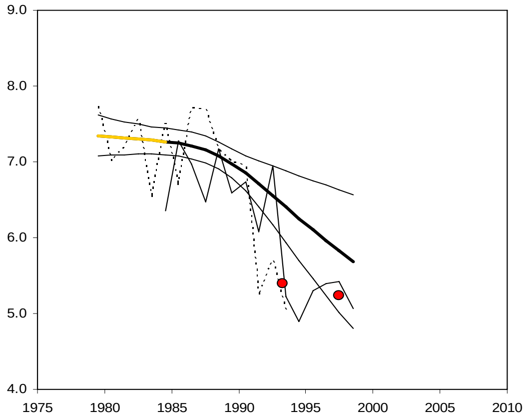
2d - Chad - unconstrained



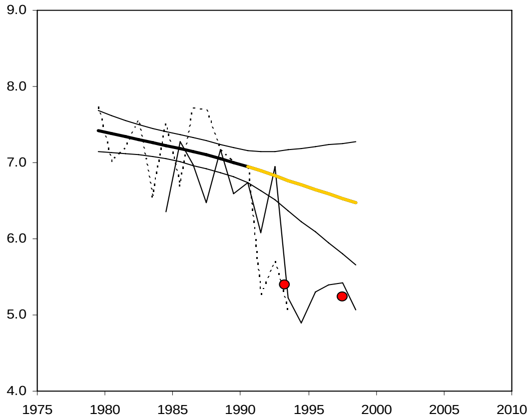
3d - Chad – 500 models



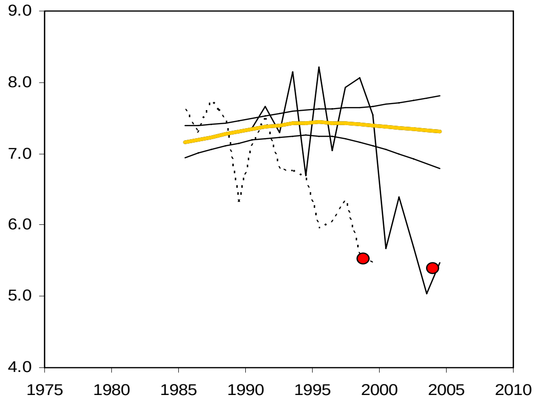
2e - Côte d'Ivoire - unconstrained



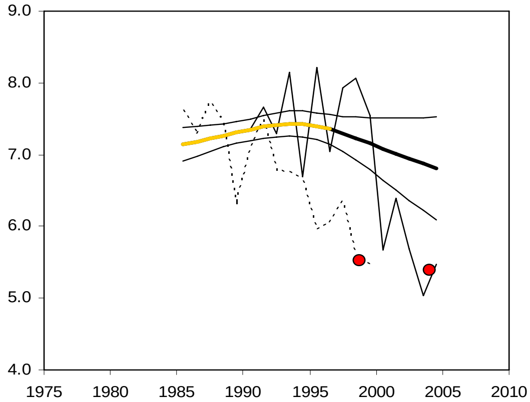
3e - Côte d'Ivoire – 500 models



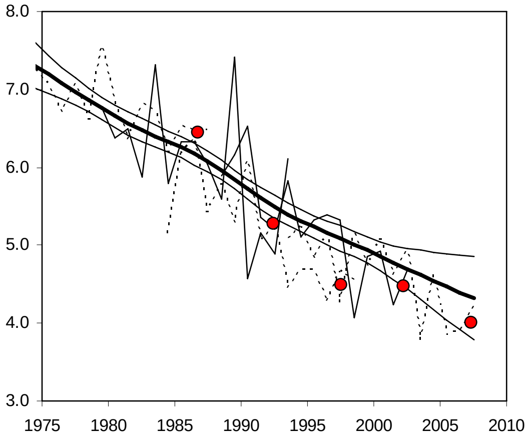
2f - Ethiopia - unconstrained



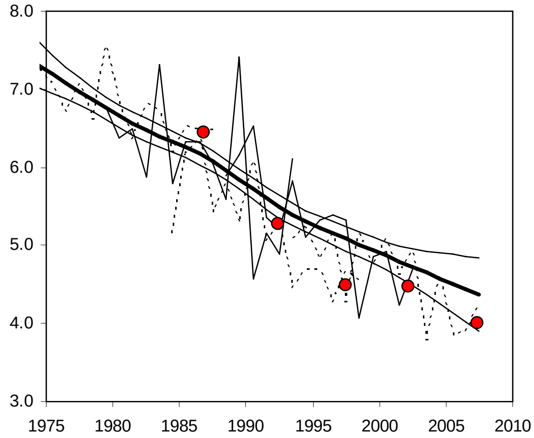
3f - Ethiopia – 500 models



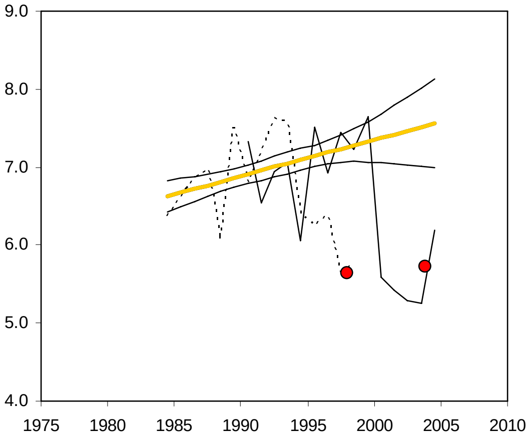
2g - Ghana - unconstrained



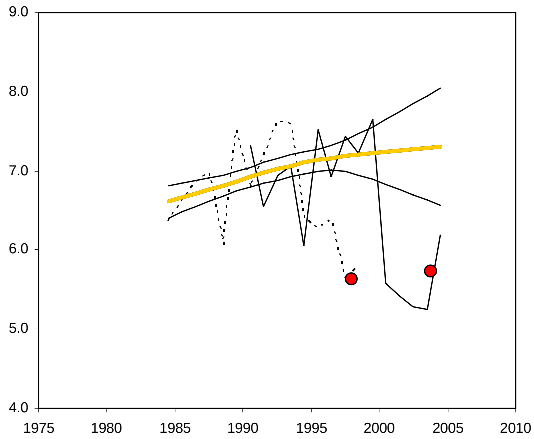
3g - Ghana – 500 models



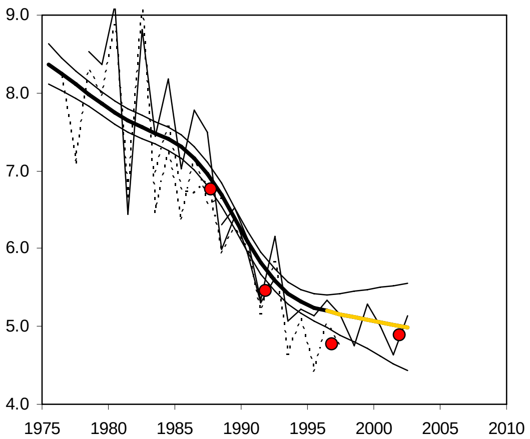
2h- Guinea - unconstrained



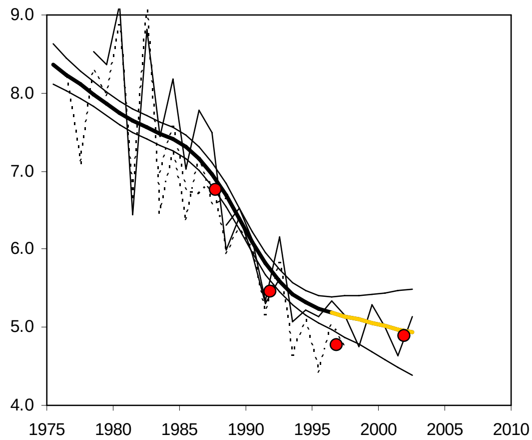
3h - Guinea – 500 models



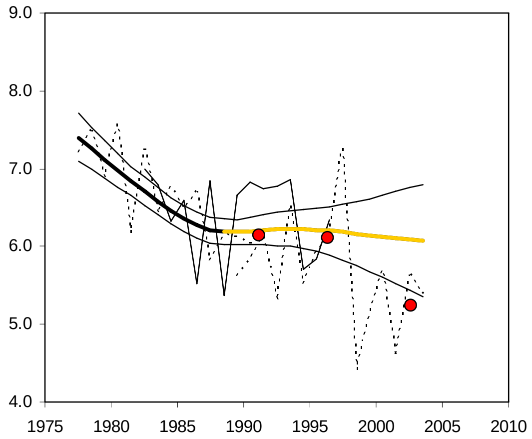
2i - Kenya – unconstrained



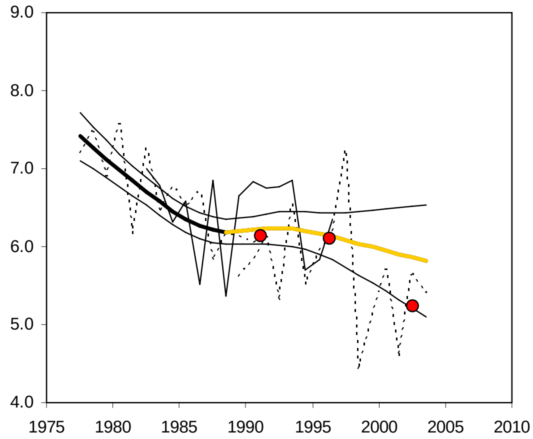
3i - Kenya – 500 models



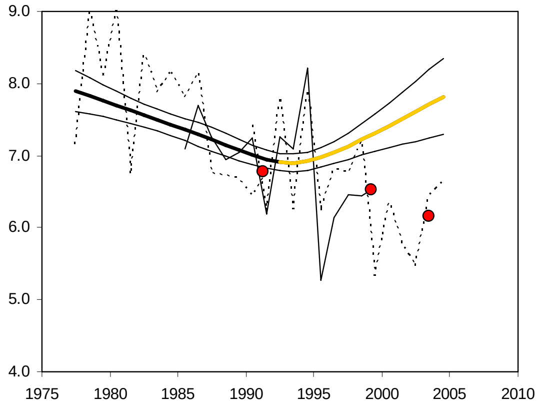
2j – Madagascar - unconstrained



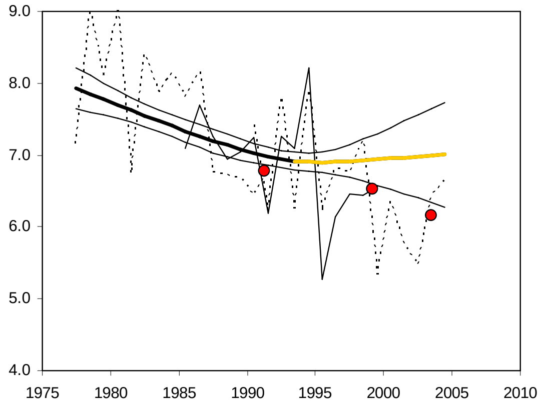
3j - Madagascar – 500 models



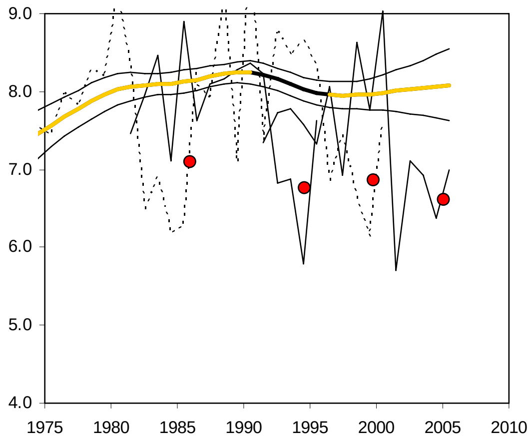
2k – Malawi - unconstrained



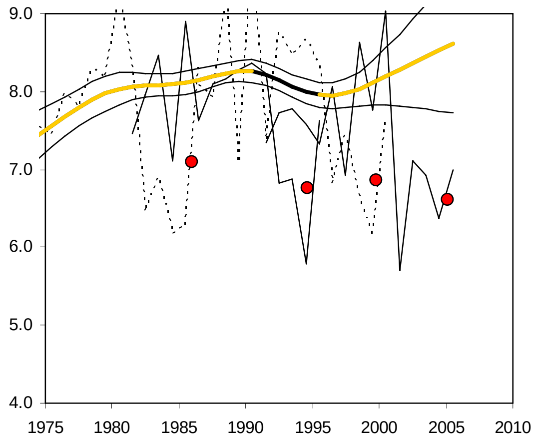
3k - Malawi – 500 models



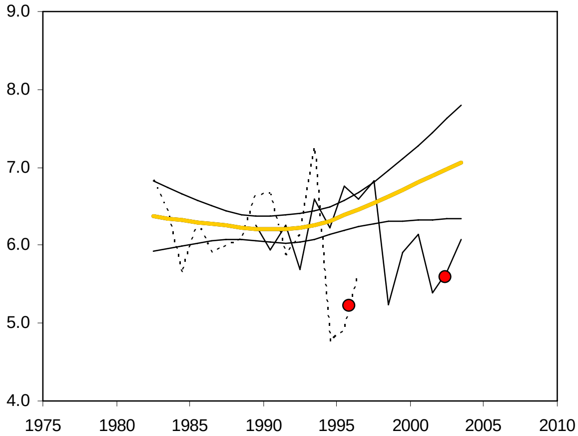
2l – Mali - unconstrained



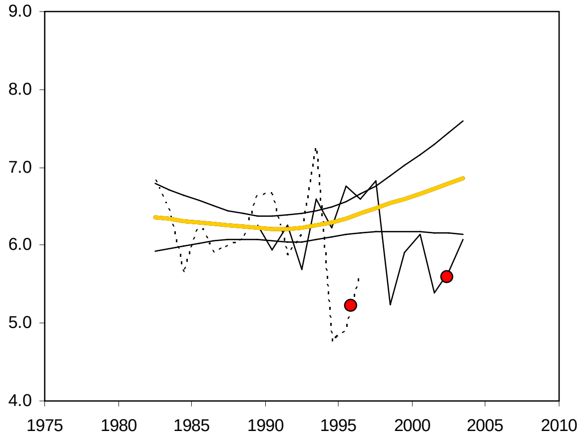
3l - Mali – 500 models



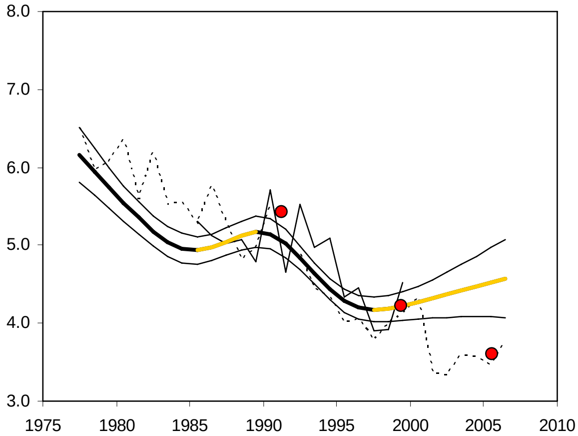
2m - Mozambique - unconstrained



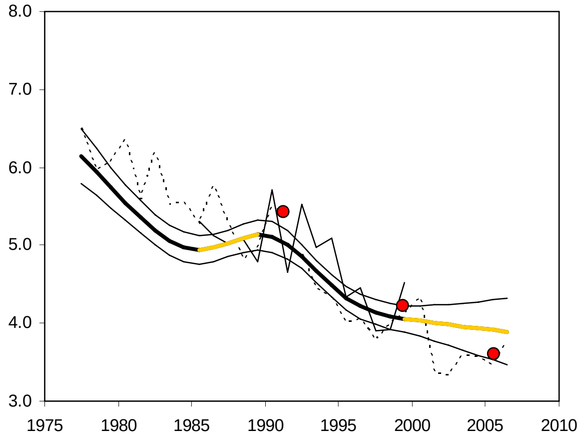
3m - Mozambique – 500 models



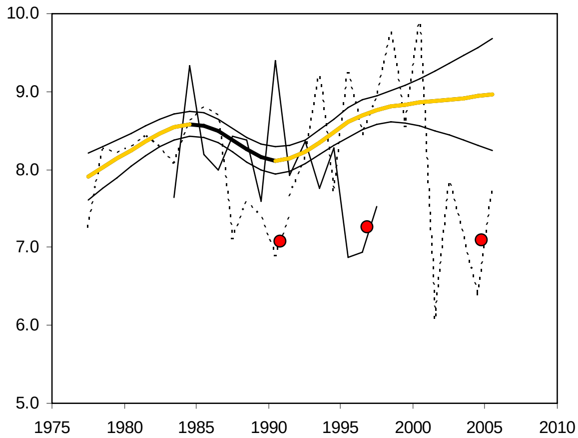
2n – Namibia - unconstrained



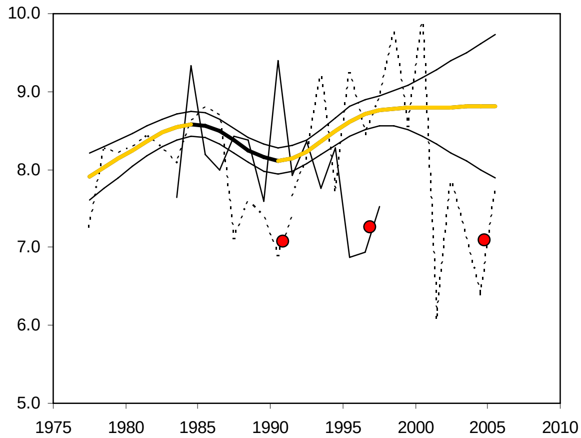
3n - Namibia – 500 models



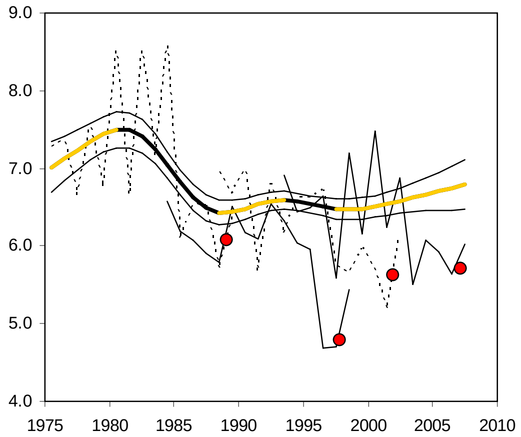
2o – Niger - unconstrained



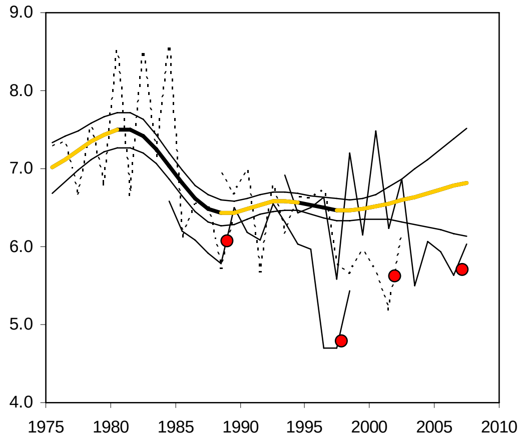
3o - Niger – 500 models



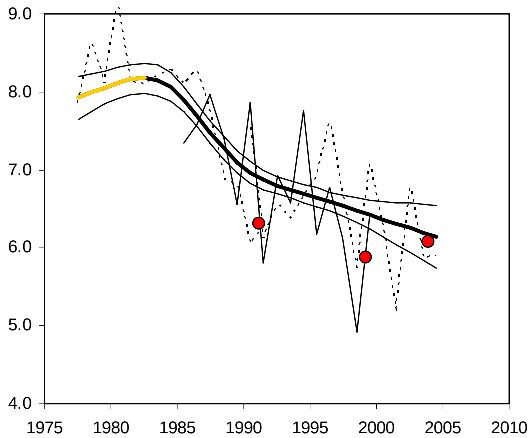
2p – Nigeria - unconstrained



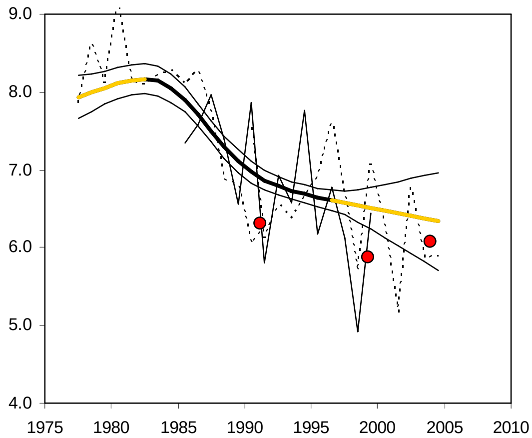
3p - Nigeria – 500 models



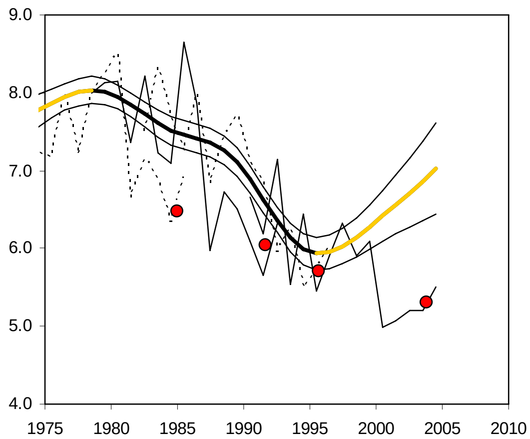
2q – Rwanda - unconstrained



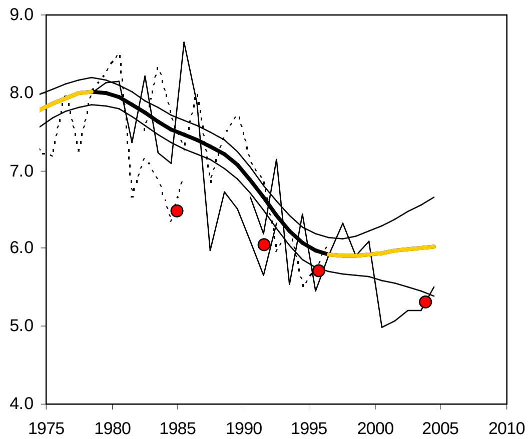
3q - Rwanda – 500 models



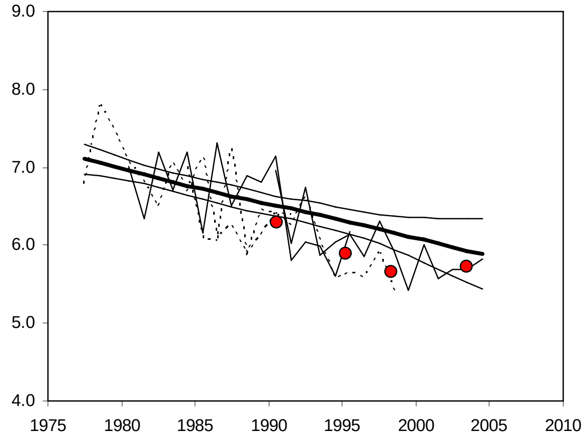
2r – Senegal - unconstrained



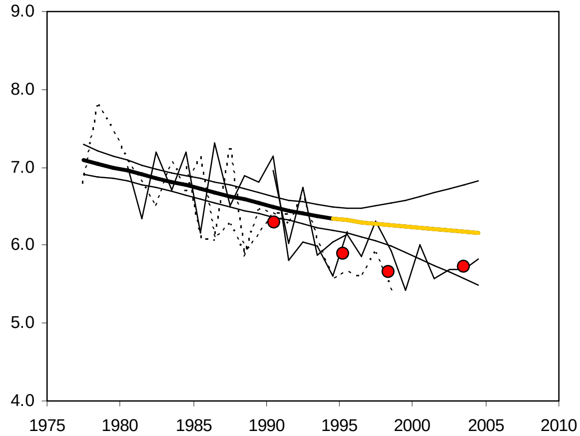
3r - Senegal – 500 models



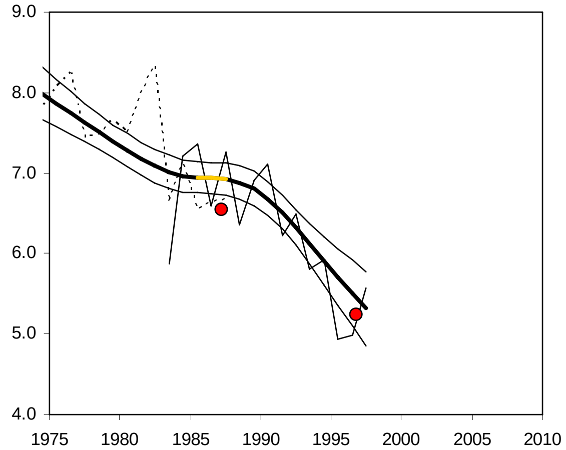
2s - Tanzania - unconstrained



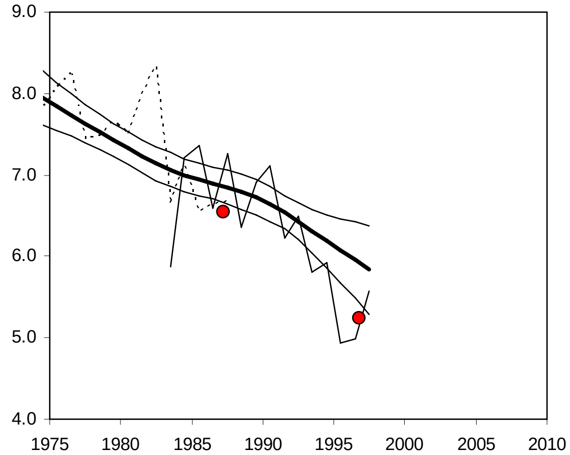
3s - Tanzania – 500 models



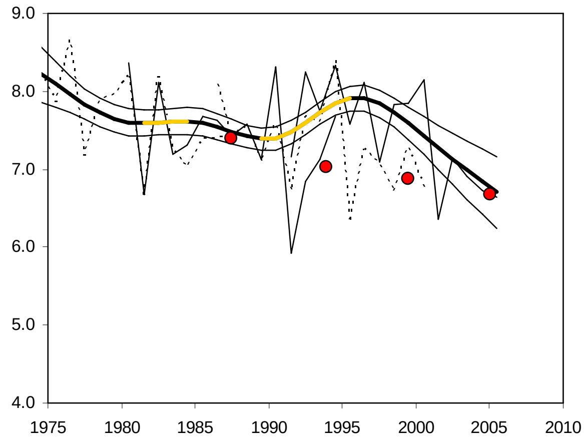
2t - Togo - unconstrained



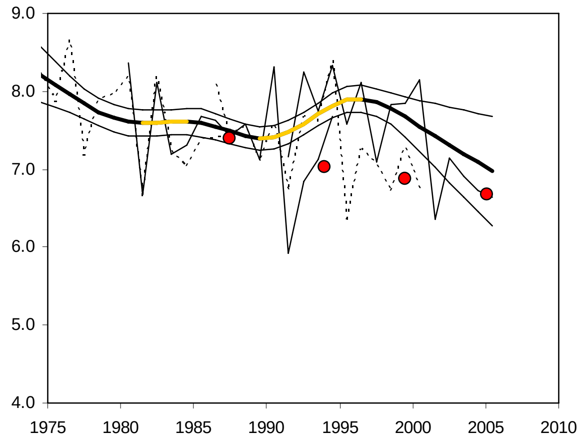
3t - Togo – 500 models



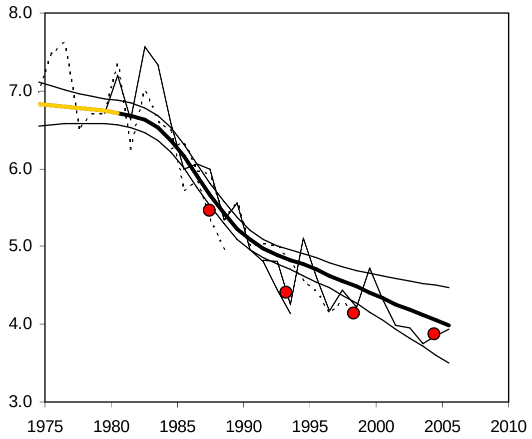
2u – Uganda - unconstrained



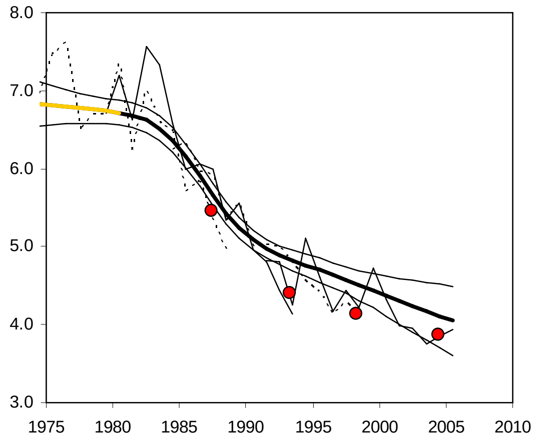
3u - Uganda – 500 models



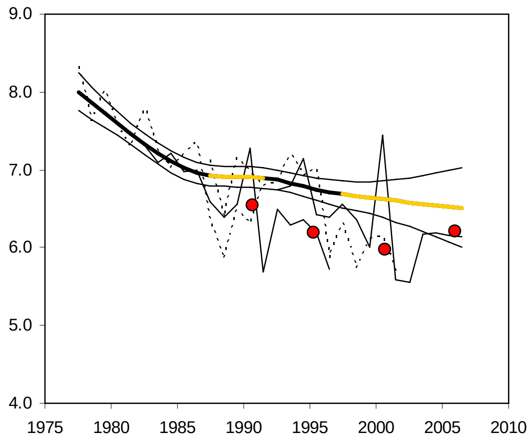
2v - Zimbabwe - unconstrained



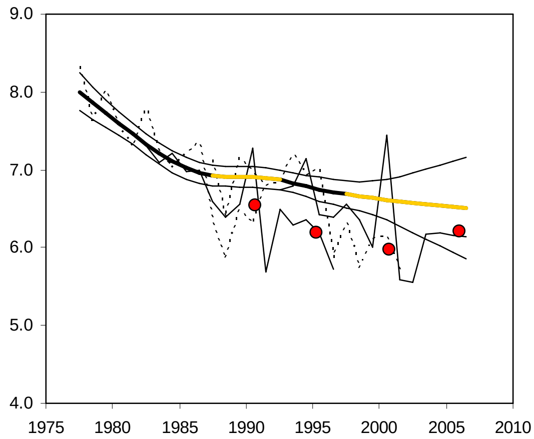
3v - Zimbabwe – 500 models



2w – Zambia - unconstrained



3w - Zambia – 500 models



Red (dark) dots represent published TFRs (on STATcompiler)
Solid smooth lines represent predicted TFRs and limits of 90% confidence intervals
Black portions of the trend indicate a significant decline. Orange portions indicate stable or increasing fertility.