Estimating the Total Fertility Rate from Multiple Imperfect Data Sources and Assessing its Uncertainty

Leontine Alkema, Adrian E. Raftery,
Patrick Gerland, Samuel J. Clark, François Pelletier *

August 2009

Abstract

We develop methodology for estimating and assessing the uncertainty of the total fertility rate over time. The estimates are based on multiple imperfect estimates from different data sources, including surveys and censuses. We take account of measurement error by decomposing it into bias and variance, and estimate both by linear regression on data quality covariates. We estimate the total fertility rate for seven countries in western Africa using a local smoother, and we assess uncertainty using the weighted likelihood bootstrap. We found that taking differences in data quality between observations into account gave better calibrated confidence intervals and reduced bias.

Keywords: Bayesian Inference, Demographic and Health Survey, Local smoother, Retrospective surveys, United Nations, Variable selection, Weighted Likelihood Bootstrap

*Leontine Alkema, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546; Email: alkema@nus.edu.sg. Adrian E. Raftery, Departments of Statistics and Sociology, University of Washington, Seattle, WA 98195-4320; Email: raftery@u.washington.edu. Patrick Gerland, Population Estimates and Projections Section, United Nations Population Division, New York, NY 10017; Email:gerland@un.org. Samuel J. Clark, Department of Sociology, University of Washington, Seattle, WA 98195-3340; Email: samclark@u.washington.edu. François Pelletier, Mortality Section, United Nations Population Division, New York, NY 10017; Email: pelletierf@un.org.

1 Introduction

Estimating demographic indicators is challenging for many developing countries because of limited data and varying data quality. This is illustrated in Figure 1 for Burkina Faso in western Africa. The black and red dots are nationally representative observations of the total fertility rate in Burkina Faso, constructed using age-specific fertility rates. In the period from 1960 until the mid 1970s, there are very few observations for the TFR. After 1970 the number of observations increases, but the observations are very spread out because of issues with data quality, e.g. observations are biased because of the collection process, or measured with large errors.

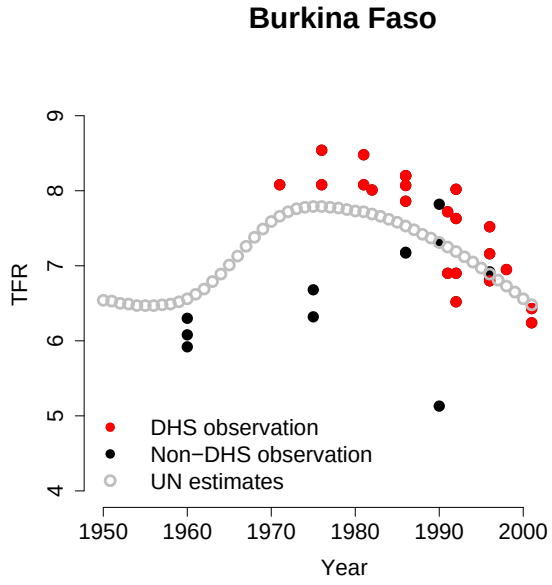

Figure 1: Observations of the total fertility rate in Burkina Faso, and UN estimates.

The United Nations Population Division produces estimates of the total fertility rate from 1950 up to the most recent five-year period for all countries in the world (United Nations, Department of Economic and Social Affairs, Population Division 2007). UN analysts estimate the fertility rates in an iterative fashion: initially age-specific fertility rates are estimated based on all available nationally representative data, combined with expert knowledge of the reliability of the different subsets of observations (e.g. known issues with a particular survey or census, or general knowledge on undercounts or overcounts of certain retrospective estimates of fertility rates). The initial fertility estimates are combined with estimates of mortality and migration to derive estimates of population counts. The population count estimates are then compared to (bias-adjusted) census counts. If estimated and observed population counts differ significantly, the estimates of the three input components of the population counts are reconsidered. Uncertainty in these input components allows for adjustments of the initial input values until population estimates and observations are in agreement. The UN estimates or Burkina Faso are shown in Figure 1.

The UN estimates for the TFR are generally considered to be of fairly good quality and are widely used. The observations from Demographic and Health surveys, shown in red in Figure 1, are also considered to be of good quality and are widely used. Figure 1 shows that when examining the total fertility rate in Burkina Faso, different conclusions about its level and trend can be drawn depending on whether the UN estimates or the DHS estimates are being used in the analysis. What is "the true TFR" in Burkina Faso? The DHS observations are not necessarily equal to the true TFR as these are observed TFR levels in subsets of the population, with measurement errors and possible biases. The UN estimates are based on observed TFR levels, as well as information on other demographic indicators, and are therefore more likely to accurately estimate the TFR. However, the drawbacks of the UN estimates are that the estimates are generated in a non-automated way, and no uncertainty assessment is included in the analysis.

There are no standardized and reproducible methods for estimating fertility rates in developing countries over time based on different data sources that assess the uncertainty of the estimates. Most of the literature on fertility data has focused on the development of indirect estimation methods (Brass 1964; Brass et al. 1968; Trussell 1975; United Nations 1983; Brass 1996; Feeney 1998). These techniques deal with biases that are caused by recall lapse errors in retrospective estimates of fertility rates (Som 1973; Potter 1977; Becker and Mahmud 1984; Pullum and Stokes 1997). The indirect estimation methods correct for reporting biases by reconciling information from recent fertility (in the last year or years) with lifetime fertility. They are typically based on one data source, and underlying assumptions can lead to problems with respect to the accuracy of the indirect estimates (Moultrie and Dorrington 2008). These methods deal with bias only, and not with differences in the variance of the measurement error.

Methods have been developed for estimating child mortality rates for countries with limited data and varying data quality. Murray et al. (2007) used a local regression model to estimate child mortality for all the countries in the world. In their approach data quality was taken into account by excluding extreme outliers from the data set and allowing for biases in observations from vital registration systems. They assessed model uncertainty by varying the smoothing parameter of the local regression. This approach does not allow for biases in observations from other sources, nor does it take differences in measurement errors between observations into account. Varying the smoothing parameter does not provide formal statistical confidence bounds for the estimates (Silverwood and Cousens 2007). Hill et al. (1998) and the Interagency Group for Child Mortality Estimation (UNICEF, WHO, World Bank and UNPD, 2007) fitted piecewise linear splines to log-transformed child mortality rates. In their approach, data quality is taken into account by assigning a weight to each observation, which depends on its data quality covariates (e.g. data collection process) and expert judgement. This approach does not adjust for biases in the different data sources.

In this paper we introduce a new methodology for estimating the total fertility rate (TFR) over time and assessing its uncertainty for countries with limited data from multiple sources with varying data quality. We first estimate average biases and measurement error variance for subsets of TFR estimates with the same set of data quality covariates by linear regression, based on what has been observed in different countries in the same region. We then adjust the TFR estimates by subtracting the estimated biases, and assign weights based

on the estimated measurement error variances. Next we estimate the TFR trajectories by applying a weighted local smoother to the bias-adjusted TFR estimates. Finally, we use the weighted likelihood bootstrap to derive statistical confidence intervals for our estimates.

Assessing the accuracy of an estimation method, including its uncertainty assessment, is important to validate whether modeling assumptions hold, but it is often not done. We describe model calibration criteria to assess the accuracy of our methods.

We apply our methodology to data from seven countries in western Africa. Some of these countries have been experiencing some of the highest fertility rates in the world in recent years. We first describe the data, and then we describe how to analyze bias and measurement error variance for different data sources, estimate the TFR over time, carry out the uncertainty assessment and validate the model. We present the results of the modeling approach for the TFR in the seven countries and compare the results of our method, that takes data quality into account, to the results of a similar method that treats all observations equally.

2 Data

We use a data set consisting of nationally representative observations of the total fertility rate for seven countries in western Africa (Burkina Faso, Gambia, Guinea, Mali, Mauritania, Niger and Senegal). All the observations were collected retrospectively; they were based on asking women about the number of births in some period in the past (e.g. the number of births in the last year before the survey/census) or complete birth histories (birth of their first child, second child, etc.). The observations come from various sources, retrospective periods, and time spans.

The observations are from censuses and surveys, with surveys divided into two groups: Demographic Health Surveys (DHS), including the World Fertility Surveys, and other surveys. The retrospective estimates of the TFR as given by the DHS are based on complete birth histories. The birth history data are tabulated into different periods; the DHS generally uses 0-3 or 0-4 years before the survey as the most recent period. This is considered to give a more robust estimator of age-specific fertility rates than calculating the rates based on the births in the last year before the survey. Censuses and non-DHS surveys generally only collect lifetime fertility and/or recent fertility (in the past year), instead of full or truncated birth histories because of time and cost constraints.

Figure 2 shows the observations in each of the seven countries. For each observation, the horizontal line points from the midpoint of the observation period to the year of data collection to visualize the recall period. The estimates as published by the United Nations in 2007 are shown in grey (United Nations, Department of Economic and Social Affairs, Population Division 2007).

For each observation, four data quality covariates were available: source, period before survey (PBS), direct/indirect estimation method, and time span. Source is either Census, Demographic and Health Survey (DHS), or other survey. Period before survey is the midpoint of the period before the survey to which the retrospective estimate refers. Time span is the

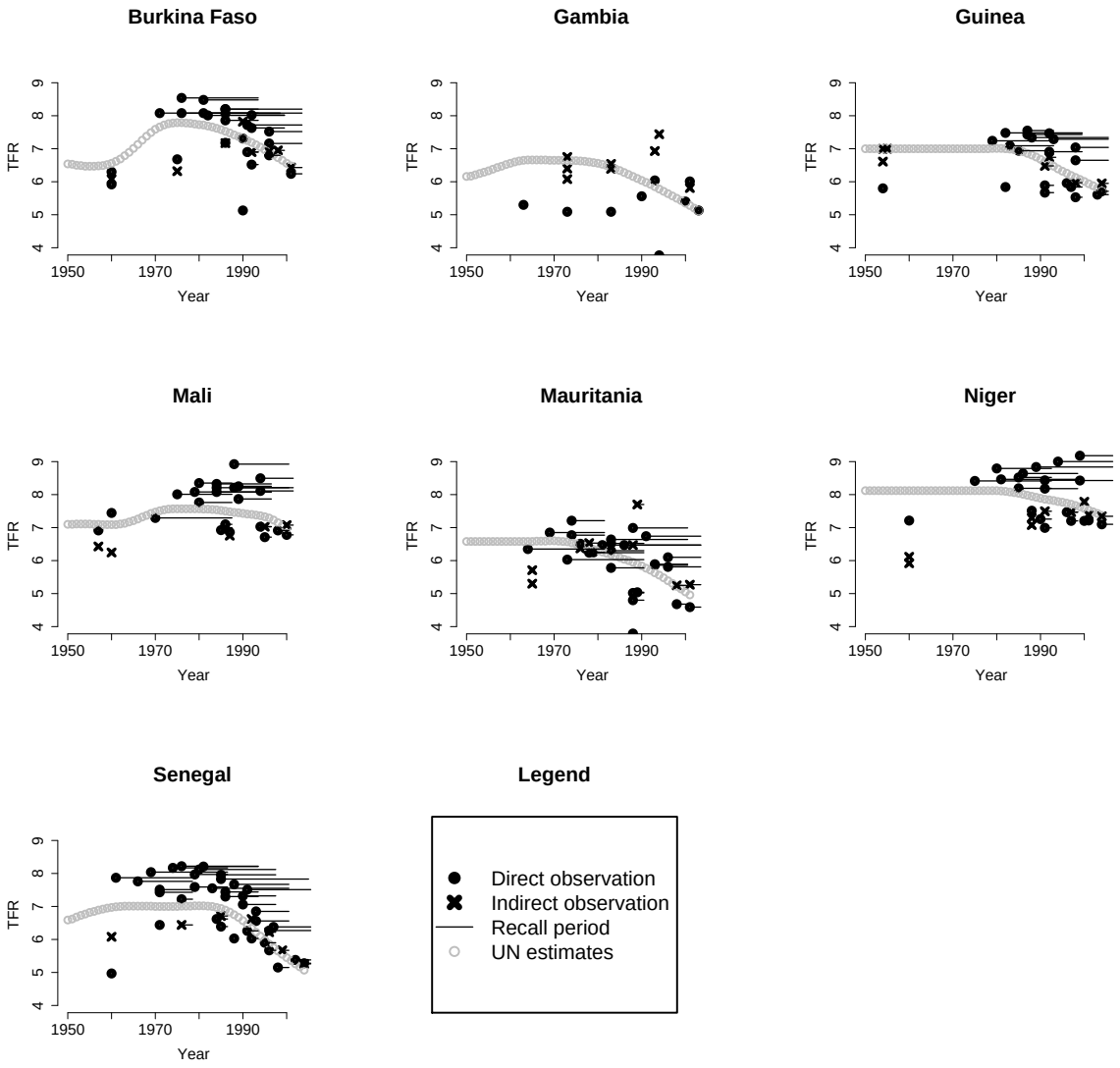

Figure 2: Direct observations (dot) and indirect observations (cross) for different data sources. The black horizontal line extends from the midpoint of the observation period to the year of data collection. The UN estimates are plotted as grey circles.

Table 1: Summary of fertility data set for seven countries in western Africa. The number of observations for each observed combination of the data quality covariates: Source, Period Before Survey (PBS — based on the midpoint of the period before the survey to which the retrospective estimate refers), Direct (specifying whether the observation is a direct or an indirect estimate), and Time Span (the number of years that the observation refers to).

Combination	Source	PBS	Direct	Time span	# Obs.
1	Census	0-1 Year	No	1 Year	18
2	Census	0-1 Year	Yes	1 Year	20
3	Survey	0-1 Year	No	1 Year	12
4	Survey	0-1 Year	Yes	1 Year	12
5	Survey	1-5 Years	No	3 Years	1
6	DHS	1-5 Years	No	3 Years	15
7	DHS	1-5 Years	No	4-5 Years	7
8	DHS	1-5 Years	Yes	3 Years	13
9	DHS	1-5 Years	Yes	4-5 Years	23
10	DHS	1-5 Years	Yes	5+ Years	1
11	DHS	5-10 Years	Yes	4-5 Years	23
12	DHS	5-10 Years	Yes	5+ Years	22
13	DHS	10+ Years	Yes	4-5 Years	50

length of the observation period in years. Table 1 summarizes the 13 combinations of data quality covariates observed in the data set.

The data quality covariate "Direct" in Table 1 divides the data set into *direct* and *indirect* estimates. Direct estimates are observations based on the reported number of births in a certain period. Indirect estimates are constructed using indirect estimation methods, which have been developed to correct for recall lapse biases in retrospective observations of fertility rates, as described in the introduction. The indirect techniques rely on the use of the P/F ratio (Brass 1964, 1968), which compares cumulated cohort fertility to cumulated period fertility. The assumptions underlying this method are that fertility and its age distribution are constant over time, and that the fertility of non-surviving women is equal to the fertility of surviving women (whose number of children is reported). Under these assumptions cohort and period fertility are equal, and deviations from equality are used to adjust the observed fertility rates. Several variations of the P/F ratio are used to relax these assumptions (Trussell 1975; United Nations 1983; Feeney 1998). However, because of the problems with the indirect estimation techniques (Moultrie and Dorrington 2008), indirect estimates can be biased too. Therefore direct estimates are included in the data set as well as indirect ones.

3 Methods

We now describe our methodology, which consists of four steps. First, the bias for each TFR observation is estimated by regression on the data quality covariates, and subtracted from the observation. Then the measurement error variance is estimated, also by regression on data quality covariates. Next, the TFR trajectory for each country is estimated by weighted local smoothing of the bias adjusted observations, the weights being the reciprocals of the estimated measurement error variances. Finally, we assess the uncertainty of our TFR estimates using the weighted likelihood bootstrap.

3.1 Modeling data quality

Some observations of the TFR are better than others, depending on the quality of the underlying data. We decompose data quality into two components: bias and measurement error variance. Bias refers to systematic over- or underestimation of the TFR, due, for example, to the observation being based on an unrepresentative sample of the population because of missing data or selection bias. Measurement errors occur randomly during the data collection process, and include sampling and non-sampling errors. Sampling errors occur if the observation is based on a subset of the population. Non-sampling errors are errors that are made during the data collection and input. Unlike sampling errors, non-sampling errors have many different sources and are often hard to detect and control. For many estimates of fertility rates, non-sampling errors are bigger (United Nations 1982).

Previous work on the quality of demographic estimates has usually not distinguished between bias and variance, and we emphasize the importance of doing so because these are distinct and can point in different directions. For example, some observations can have large biases but small measurement errors, while others are unbiased but less precise. Therefore it is important to account for bias and variance separately. We deal with bias by adjusting the observations, and with variance by weighting them. For example, a biased observation with small measurement error variance is adjusted and then assigned a high weight. An unbiased observation with large measurement error variance is not adjusted but gets assigned a low weight.

Our probability model for observation y_{cts} (in year $t = 1, \ldots, T_c$ for observation $s = 1, \ldots, n_{ct}$) is

$$y_{cts}|f_{ct} \sim N(f_{ct} + \delta_{cts}, \sigma^2_{cts}) \quad (t = 1, \ldots, T_c; s = 1, \ldots, n_{ct}),$$

where y_{cts} is the s-th estimate of the TFR for country c in year t, f_{ct} is the unobserved true TFR in year t for country c, δ_{cts} is the bias of observation y_{cts}, and σ^2_{cts} is the observation-specific error variance. We use data quality covariates to assess the bias and error variance of each TFR estimate, extending the work by Hill et al. (1998) and the Interagency group for child mortality estimation (UNICEF, WHO, World Bank and UNPD, 2007) on differences in error variance in child mortality rates, and the work of Gerland (2007) on examining the associations between data quality covariates and the data quality of mortality and fertility rate estimates. In our approach, we use linear regression to estimate how bias and error

variance depend on the data quality covariates, and we combine the observations from the seven countries in western Africa to estimate both.

Estimating bias

Bias is estimated as a function of the data quality covariates, using linear regression. The advantage of this approach, as compared to indirect estimation methods, is that no assumptions are made about the age structure of the fertility rates or the trends in fertility over time. Multiple data sources are modeled and adjusted simultaneously, and biases are estimated based on what has been observed in the seven countries in western Africa.

To examine the association between bias and the data quality covariates, unbiased estimates of the TFR are needed. By saying an estimate is unbiased, we do not mean that it is correct or even necessarily of high quality; we mean that it does not substantially over- or underestimate the TFR. Here, we assume that the estimates of the TFR published by the United Nations Population Division (United Nations, Department of Economic and Social Affairs, Population Division 2007), are unbiased, so that $E[u_{ct}] = f_{ct}$, where u_{ct} denotes the UN estimate for country c, year t. We take the published UN estimates as the least biased available because they are based on the UN analysts' knowledge of the shortcomings of the multiple data sources used, as well as information on mortality, population counts and migration to ensure internal consistency of intercensal birth cohorts. Note that using the UN estimates as unbiased estimates of the TFR does not imply that the estimates of the TFR that are derived with this approach, will be necessarily equal to the UN estimates; here the UN estimates are used only to determine average biases for observations with the same set of data quality covariates.

With the UN estimates taken as unbiased estimates of the TFR, the bias δ_{cts} of observation y_{cts} is the expected value of the difference $d_{cts} = y_{cts} - u_{ct}$ between the observation and the UN estimate, so that $E[d_{cts}] = \delta_{cts}$. We estimate the biases δ_{cts} by regressing d_{cts} on the data quality covariates using the model

$$E[d_{cts}] \quad = \quad \boldsymbol{x}_{cts}\boldsymbol{\beta},$$

where the row $\boldsymbol{x}_{cts}$ of the design matrix $\mathbf{X}$ contains the data quality covariates. Thus the biases δ_{cts} can be estimated from the relationship $\delta_{cts} = \boldsymbol{x}_{cts}\boldsymbol{\beta}$.

The next question is which predictors to include in the bias regression model. There are four data quality covariates, and each of these is a categorical variable that can take several values. We code these as dummy variables and consider the possibility of including or excluding each of them. Dummy variables for the individual DHS surveys (each of which generates multiple observations), as well as the observation year, year of data collection and the level of the TFR (as given by the UN estimates) are also candidates to put into the model. To select the best model we use Bayesian model selection (Raftery 1995). Specifically we consider all possible subsets of predictors and choose the one with the best value of BIC. This is done using the `bicreg` function from the BMA package in the statistical language `R` (Raftery, Painter, and Volinsky 2005), available at http://cran.r-project.org/web/packages/BMA.

The estimated biases $\hat{\delta}_{cts}$ are then subtracted from the observation to get the bias-adjusted observations

$$\begin{aligned} z_{cts} &= y_{cts} - \hat{\delta}_{cts} \\ &\sim N(f_{ct}, \rho^2_{cts}), \end{aligned} \tag{1}$$

where ρ^2_{cts} is the observation-specific error variance of the bias-adjusted observation.

Estimating measurement error variance

A similar approach is used for estimating the measurement error variances, specifically the values of ρ^2_{cts} in Eq. (1). We assume that the absolute differences between the UN estimates and the bias-adjusted observations, z_{cts}, are proportional to the absolute differences between bias-adjusted observations and the true TFR, so that

$$E|z_{cts} - u_{ct}| \propto E|z_{cts} - f_{ct}|.$$

It follows from Eq. (1) that $E|z_{cts} - f_{ct}| = \sqrt{\frac{2}{\pi}}\rho_{cts}$, and so

$$E|z_{cts} - u_{ct}| \propto \rho_{cts}.$$

We can therefore estimate the association between the data quality covariates and the relative differences in ρ_{cts} between observations using the regression model

$$E|z_{cts} - u_{ct}| = \boldsymbol{w}_{cts}\boldsymbol{\lambda},$$

where the row $\boldsymbol{w}_{cts}$ of the design matrix $\mathbf{W}$ contains the data quality covariates that are associated with error variance and $\boldsymbol{\lambda}$ is the vector of regression coefficients. Thus $\rho_{cts} \propto \boldsymbol{w}_{cts}\boldsymbol{\lambda}$. Variable selection is done in the same way as for the bias regression.

3.2 Estimating TFR Trajectories

We apply a local smoother (Cleveland and Devlin 1988; Cleveland et al. 1992; Loader 1999) to the bias-adjusted estimates, weighted by the reciprocals of their estimated error variances, to estimate the annual country-specific TFR. A quadratic polynomial is fitted to the bias-adjusted observations in a neighborhood of t^* to estimate the TFR in year t^*. The observations within the neighborhood are weighted by the product of the *distance* and *error variance* weights. Smaller distance weights are assigned to observations that are farther away from year t^*. The error variance weight is the reciprocal of the error variance, so that observations with larger error variance are less influential when estimating the TFR in year t^*. When fitting the local smoother, the size of the neighborhoods and the distance weights of the observations depend on a smoothing parameter α. This is estimated by cross-validation based on the data sets for all countries combined; α is chosen to minimize the overall mean squared error when leaving out observations one at the time. We use the `R` function `Locfit` (Loader 1999) to carry out the local smoothing.

We assess uncertainty of the TFR trajectories using the weighted likelihood bootstrap (Newton and Raftery 1994). This is similar to the standard bootstrap (Efron 1979), except that where the standard bootstrap resamples data points, the weighted likelihood bootstrap gives a weight to every data point that is sampled from a Dirichlet distribution. The weighted likelihod bootstrap works better in our case because the fit of the local smoother to resampled bootstrapped data breaks down if no or few data points at the end points of the observation period get resampled. In the weighted likelihood bootstrap no data points get left out, and so this problem does not arise.

The weighted likelihood bootstrap works as follows for our data, to sample B bootstrap replicates. For $b = 1, \ldots, B$ we cycle through the following steps:

1. For each country c, sample bootstrap weights $p_{cts}^{(b)}$ for observation y_{cts} from the distribution
$$(p_1, \ldots, p_m) \sim \text{Dirichlet}_m(1, \ldots, 1),$$
with $m = \sum_{t=1}^{T_c} n_{ct}$, the total number of observations in country c. The $\text{Dirichlet}_m(1, \ldots, 1)$ distribution is uniform in the sense that it gives equal probability to all values of the vector $(p_1, \ldots, p_m)$ such that $p_1 + \ldots + p_m = 1$.

2. Estimate the biases $\delta_{cts}^{(b)}$ using weighted regression, based on the data set of all countries and
$$y_{cts} \sim N\left(f_{ct} + \delta_{cts}, \frac{\sigma^2}{p_{cts}^{(b)}}\right),$$
where σ^2 is the error variance for all observations combined. The bias-adjusted observations are given by $z_{cts}^{(b)} = y_{cts} - \hat{\delta}_{cts}^{(b)}$.

3. Estimate the differences in error variance $\rho_{cts}^{(b)}$ using weighted regression, based on the data set of all countries and
$$z_{cts}^{(b)} \sim N\left(f_{ct}, \frac{\rho_{cts}^{2(b)}}{p_{cts}^{(b)}}\right).$$

4. Estimate the TFR by fitting the local smoother to the bias-adjusted observations $z_{cts}^{(b)}$, taking into account the differences in error variance $\rho_{cts}^{(b)}$ and the bootstrap weights $p_{cts}^{(b)}$. The distance weights and the local neighborhoods in the local smoother vary by bootstrap replicate too, because the smoothing parameter α is re-estimated within each bootstrap replicate.

3.3 Model validation

We validate the method by cross-validation, in which some observations are left out, while the method is applied to the remaining observations (called the "training data set"). We then assess how well the resulting predictive distributions agree with what actually happened (the left-out observations). More precisely, we assess whether the predictive distributions

are calibrated, meaning that the prediction intervals contain the truth the right proportion of the time. We use the following measures of calibration: (a) the proportion of left-out observations that fall outside their confidence intervals, (b) the average bias of the estimated TFR compared to the left-out observation, (c) the standardized absolute prediction error, and (d) the probability integral transform histogram of the left-out observations.

The confidence intervals for the left-out observations y_{cts} are based on

$$y_{cts} \sim N(\tilde{f}_{ct} + \tilde{\delta}_{cts}, \tilde{\nu}^2_{cts}). \tag{2}$$

In Eq. (2), $\tilde{f}_{ct}$ is the median TFR, $\tilde{\delta}_{cts}$ is the estimated bias, and $\tilde{\nu}^2_{cts}$ is the estimated total variance, all estimated from the training data set and the data quality covariates of observation y_{cts}. The predictive variance of the observations is

$$\tilde{\nu}^2_{cts} = \mathrm{Var}(\tilde{f}_{ct}) + \tilde{\rho}^2_{cts},$$

where the variance of the TFR $\mathrm{Var}(\tilde{f}_{ct})$ and the observation-specific error variance $\tilde{\rho}^2_{cts}$ are estimated from the training data set.

The bias in the set of left-out observations, with respect to the estimated TFR, is estimated by the mean of the differences between a bias-adjusted left-out observation and the estimated TFR. The standardized absolute prediction error (SAPE) for observation y_{cts} is defined by:

$$\mathrm{SAPE}_{cts} \quad = \quad \sqrt{\frac{\pi}{2}} \frac{|y_{cts} - \tilde{\delta}_{cts} - \tilde{f}_{ct}|}{\tilde{\nu}_{cts}}.$$

If our modeling assumptions hold, the mean SAPE is around 1, because $E|y_{cts} - \tilde{\delta}_{cts} - \tilde{f}_{ct}| = \sqrt{2/\pi}\tilde{\nu}_{cts}$. A larger value of the mean SAPE indicates that the left-out observations are more spread out than expected, while a smaller value says they are less so.

Our last calibration criterion is the probability integral transform (PIT) histogram. The probability integral transform for the left-out observation y_{cts} is

$$\mathrm{PIT}_{cts} = \Phi\left(\frac{y_{cts} - \tilde{\delta}_{cts} - \tilde{f}_{ct}}{\tilde{\nu}_{cts}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. From Eq. (2) it follows that the PIT values should be approximately uniformly distributed between 0 and 1 if our model is valid. Calibration is compared between models by comparing the histograms of the PIT values of the left-out observations. For a histogram with H bars, each bar with width $1/H$ should contain about a proportion $1/H$ of the PIT values and thus have height 1. The summary criterion for model comparison in terms of PIT values is given by the area of the PIT histogram that is located above one, as this represents the deviation from uniformity of the PIT values (Berrocal et al. 2007). We call this the "PIT area". If the TFR estimates are unbiased, a smaller value of the PIT area means a better calibrated model.

For more complete details of our methodology, see Alkema (2008).

Table 2: Results of bias regression model.

	Coefficient	Std. Error	t-value
Intercept	−0.74	0.11	−6.5
PBS 5 - 10 Years	1.07	0.10	11.0
PBS 10+ Years	1.30	0.10	13.4
Direct	−0.45	0.09	−4.9
Year - 1954	0.02	0.003	6.3

4 Results

4.1 Bias regression

The covariates that were selected into the bias regression model are given by the recall period of the observation, its estimation method and observation year. The results for the complete data set are given in Table 2. For the seven countries in western Africa, retrospective estimates that refer to periods more than 5 years in the past have a positive bias of over one child compared to recall periods of less than 5 years. The regression coefficient is 1.07 children if the midpoint of the period before the survey is between 5 and 10 years (PBS 5−10), and 1.30 children if the period is more than 10 years (PBS 10+). Compared to indirect estimates, a direct observation has a negative bias of almost half a child. Bias is positively associated with the observation year: the longer ago the observation, the more negative its bias. The first observation was in 1954; in that year the bias started with a large negative value of −0.74 for indirect estimates with recall period less than 5 years, and this bias became less negative at the rate of 0.02 children per year.

The estimated biases for each outcome category and different years are given in Table 3. Note that the first observation with a recall period of more than 5 years refers to 1961. The most recent estimate with a period before survey of more than 10 years was in 1994, and in 1999 for a retrospective period between 5 and 10 years. The bias was essentially zero for indirect estimates in 1994 with a recall period that is less than 5 years. Direct estimates tend to underestimate fertility because of recall biases. Explanations for the underestimation of TFR in the past include lower coverage of areas with higher fertility, age-misreporting and memory/recall biases. This bias was large and negative in 1954, and decreased steadily in absolute value, at the rate of 0.02 children per year. Biases also increase with the recall period of the observation, illustrated when reading the table from left to right. Observations with a recall period that is less than 5 years tend to have a slightly negative bias, while observations with longer recall periods have a positive bias.

The positive bias of observations with longer recall periods is not surprising in light of Figure 2, where almost all observations with long recall periods (long horizontal lines) are higher than the UN estimates. Differences in survival rates partly explain the positive bias of retrospective estimates: the estimates are based on the birth histories of the women who survived until the year of the survey. Thus a positive correlation between fertility

Table 3: Estimated biases for different observation years and outcome categories.

Obs. year	Direct	PBS		
		< 5 Years	5 - 10 Years	10+ Years
1954	Yes	-1.19		
	No	-0.74		
1961	Yes	-1.06	0.01	0.24
	No	-0.61		
1970	Yes	-0.89	0.18	0.40
	No	-0.44		
1980	Yes	-0.70	0.37	0.59
	No	-0.25		
1994	Yes	-0.44	0.63	0.85
	No	0.01		
1999	Yes	-0.35	0.73	
	No	0.11		
2004	Yes	-0.25		
	No	0.20		

and female survival results in overestimation of the total fertility rate. The positive biases of retrospective estimates may also be due in part to age misreporting and other data reporting issues (Ewbank 1981; United Nations 1982; Pullum 2006).

Figure 3(a) shows the outcomes of the bias regression model for Burkina Faso. The observations in Burkina Faso are plotted with grey dots and the UN estimates are given by the black circles. The black dots are the bias-adjusted observations, when all the observations in the data set are used in the bias regression model. The bias-adjusted observations show a more coherent trend than the uncorrected observations.

4.2 Error variance regression

The selected error variance regression model includes the following variables: collection year after 1995, recall period up to one year and DHS in Mauritania in 1990. The estimated measurement standard deviations $\hat{\rho}_{cts}$ for the different outcome categories are given in Table 4. The indicator for data collection year was included because residual plots showed that the error variance was higher for observations that were collected before the mid 1990's, indicating that data quality improved after 1995. If an observation was based on one year of data before the survey was collected, its error variance increased. The DHS in Mauritania in 1990 had higher error variance than the other DHS's.

The larger the standard deviation of an observation within a certain category, the less informative that observation is about the TFR and the wider the confidence interval for the TFR based on that observation alone. The confidence intervals for the TFR based on the

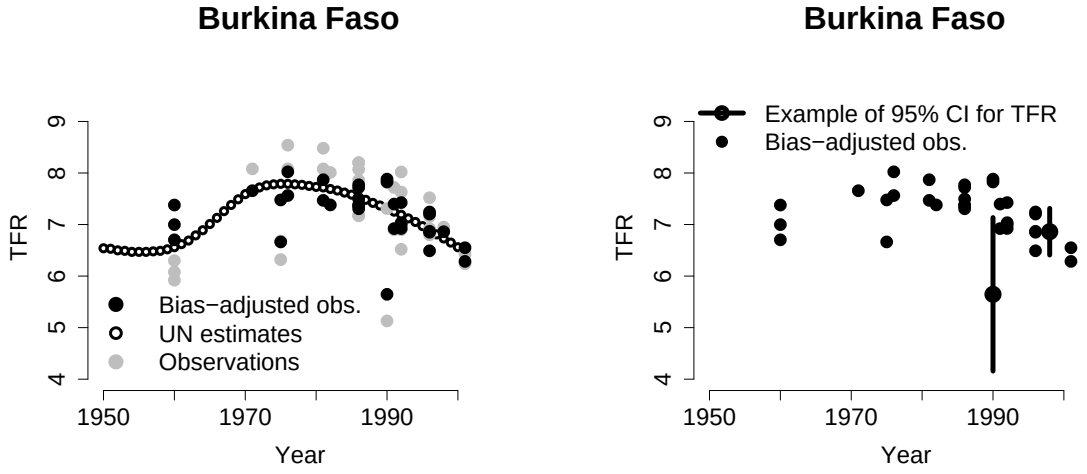

Figure 3: Illustration of bias adjustment and difference in error variance for Burkina Faso: (a) UN estimates (black circles) and observations y_{cts} (grey dots), with the bias-adjusted observations $y_{cts} - \hat{\delta}_{cts}$ (black dots), (b) Bias-adjusted observations (black dots), with the 95% confidence interval for the TFR based on a single observation, $[y_{cts} - \hat{\delta}_{cts} - 2\hat{\rho}_{cts}, y_{cts} - \hat{\delta}_{cts} + 2\hat{\rho}_{cts}]$, shown for two observations (vertical black lines).

Table 4: Results of error variance regression: estimated measurement standard deviations, $\hat{\rho}_{cts}$ for different outcome categories.

	Observations from DHS Mauritania (1990)			
	No		Yes	
Category	No. obs.	$\hat{\rho}_{cts}$	No. obs.	$\hat{\rho}_{cts}$
PBS >1, Before 1995	104	0.42	9	0.85
PBS 0-1, Before 1995	51	0.74	0	-
PBS >1, After 1995	42	0.23	0	-
PBS 0-1, After 1995	11	0.55	0	-

estimated standard deviations are given for two observations in Burkina Faso in Figure 3(b). The comparison illustrates that the bias-adjusted observation that is further away from the general trend had a larger estimated error variance, as expected.

These results underscore the importance of distinguishing between bias and variance. For example, we found that direct estimates made in 1999 on the basis of retrospective data (PBS > 5 years) had large biases but low variances, so our method adds a bias adjustment to these estimates but then gives the adjusted estimates high weights. In contrast, indirect estimates (with short recall periods) made before 1995 had little or no bias, but larger variances. Our method does not adjust these estimates at all, but gives them relatively small weights.

4.3 TFR estimates

The TFR estimates and their uncertainties for each country are shown in Figure 4. The grey lines in the plots are a random sample of TFR trajectories, given by the local smoother fits in the weighted likelihood bootstrap. The solid black line is our TFR estimate and the dashed lines show the annual quantile-based 95% confidence intervals. The UN estimates are plotted in the same figure as a grey solid line with squares. In general, the UN estimates were within or close to our 95% confidence intervals, except for lower UN estimates in Senegal from 1970 through the mid 1990s. We estimated a larger increase in fertility rates in the 1960s than the UN did for Niger and Senegal, and a smaller increase in Burkina Faso. For Gambia, the UN estimates showed a steeper decline during the second half of the observation period than our results.

The confidence intervals for the TFR were wider for the years before the mid 1970s than afterwards, for all the countries. Gambia had the most uncertainty about the past levels of the TFR, because of the scarcity of data sources and in particular of DHS retrospective birth histories, and the spread of the observations. The median width of its annual 95% confidence intervals was 0.74 children. Its confidence intervals were narrowest around 2000 and in the mid 1970s because these were periods with more observations. For the other countries, the median width of the 95% confidence intervals was much smaller, at around 0.35 children.

4.4 Method validation and comparison

Our method takes account of the differences in data quality between data sources, but one could ask whether this actually improved the estimates. We assessed this by comparing our method with a method that is the same except that it does not adjust or weight the observations for data quality. We will refer to the method that takes into account bias and difference in measurement errors as the *corrected method*, and to the method that treats all observations equally as the *uncorrected method*.

Figure 5 shows the confidence intervals for the TFR for both methods in the seven countries in western Africa. The two methods differed most at the start of the observation period, with the uncorrected method giving lower estimates than the corrected method. For most countries the uncorrected method peaked in the mid 1980s, at a TFR that was

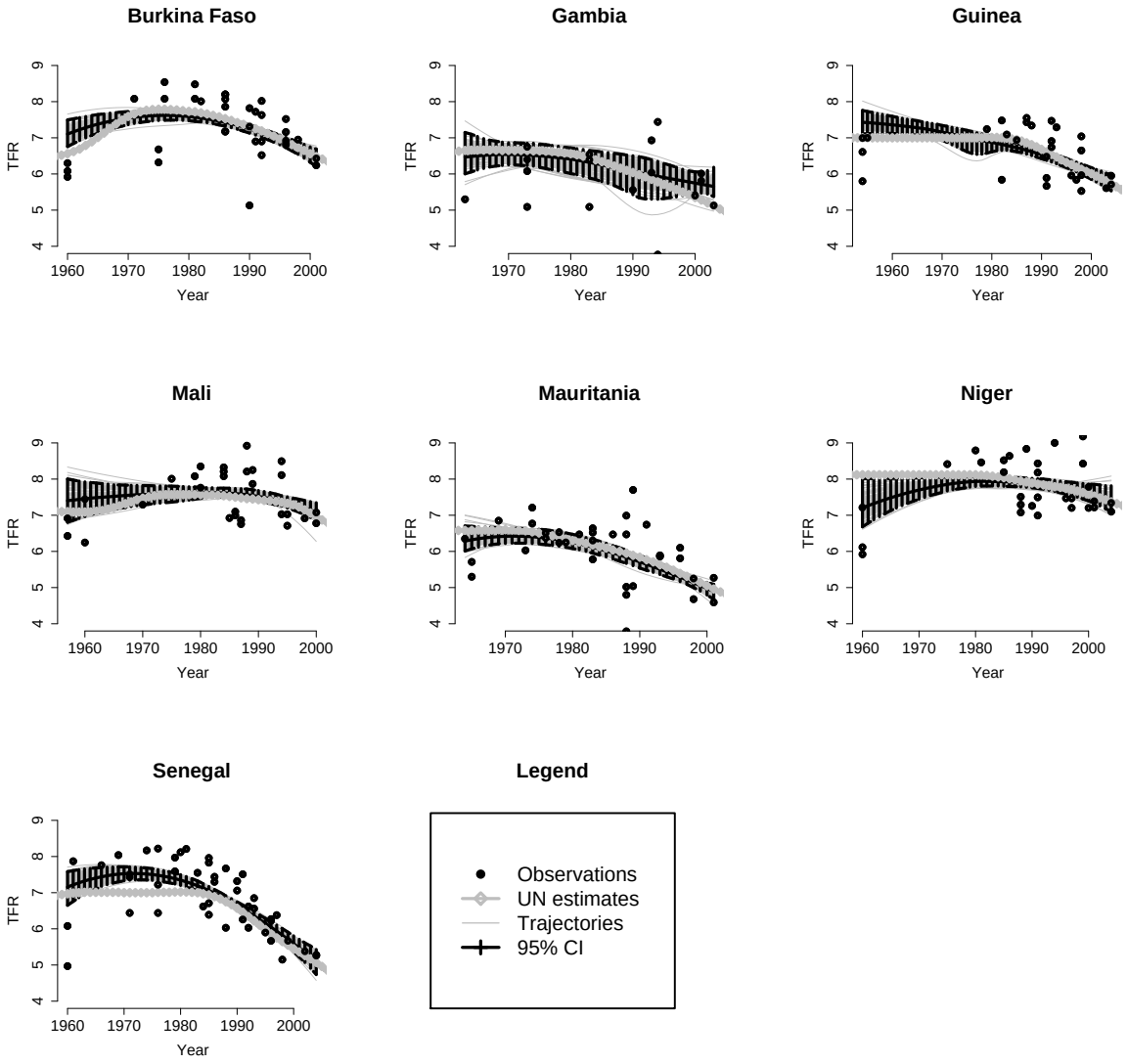

Figure 4: Median estimates and confidence intervals for the TFR. The annual median estimates are shown by the solid line and the annual 95% confidence intervals (CI) by vertical lines. The grey lines are a random sample of TFR trajectories, given by the local smoother fits in the weighted likelihood bootstrap. The observations are displayed as black dots and the UN estimates by the grey line with squares.

Table 5: Validation results for the corrected and uncorrected methods: Bias (the average distance between left-out observations and the median TFR estimate), SAPE (the standardized absolute prediction error), PIT area (the area above one in the PIT histogram), and the proportion of left-out observations that falls outside their 80% and 95% confidence intervals.

Leaving out	Method	Bias	SAPE	PIT area	Proportion of observations <80	>80	<95	>95
50 Observations	Corrected	−0.02	1.06	0.05	0.08	0.11	0.04	0.04
	Uncorrected	−0.06	1.00	0.11	0.12	0.04	0.03	0.01
DHS's	Corrected	0.03	1.23	0.14	0.14	0.13	0.04	0.07
	Uncorrected	0.21	1.08	0.19	0.07	0.14	0.01	0.04

higher than the estimate from the corrected method. In Gambia, the only country without a DHS survey, the uncorrected method gave lower estimates than the corrected method for all years. The confidence intervals were generally much narrower for the corrected method — on average 40% narrower. In most cases, the UN estimates were inside the 95% confidence intervals of both methods.

To validate the methods, we left out different subsets of the observations, implemented the methods without them, and then compared the resulting predictive distributions with the observations themselves. The subsets left out for this cross-validation exercise were: (i) random subsets of observations: 10 different subsets of 50 observations, and (ii) one DHS at a time (there were 22 DHS's in total). Leaving out one survey at the time and then examining how the left-out observations fit into the uncertainty assessment is the most realistic scenario in terms of adding "new" observations to the data set, that are independent of the observations that were already in the data set. We did this for the uncorrected method as well as the corrected method.

The results are summarized for the two left-out categories in Table 5. The average bias was 0.03 children or less for both categories for the corrected method, and somewhat larger for the uncorrected method: 0.21 children for the left-out DHS's. Uncertainty was slightly underestimated in the corrected method, as shown by the standardized absolute prediction error (SAPE), which is larger than one for both outcomes, and by the proportions of observations that fell outside the confidence intervals.

The PIT histograms in Figure 6 do not show any systemmatic lack of calibration for the corrected method, but they clearly indicate the bias in the uncorrected method. This is confirmed by the better values of the PIT area for the corrected method compared to the uncorrected method. Overall, we conclude that our corrected method is reasonably well calibrated, and that taking account of data quality is worthwhile in that it removes the systematic biases in the uncorrected method.

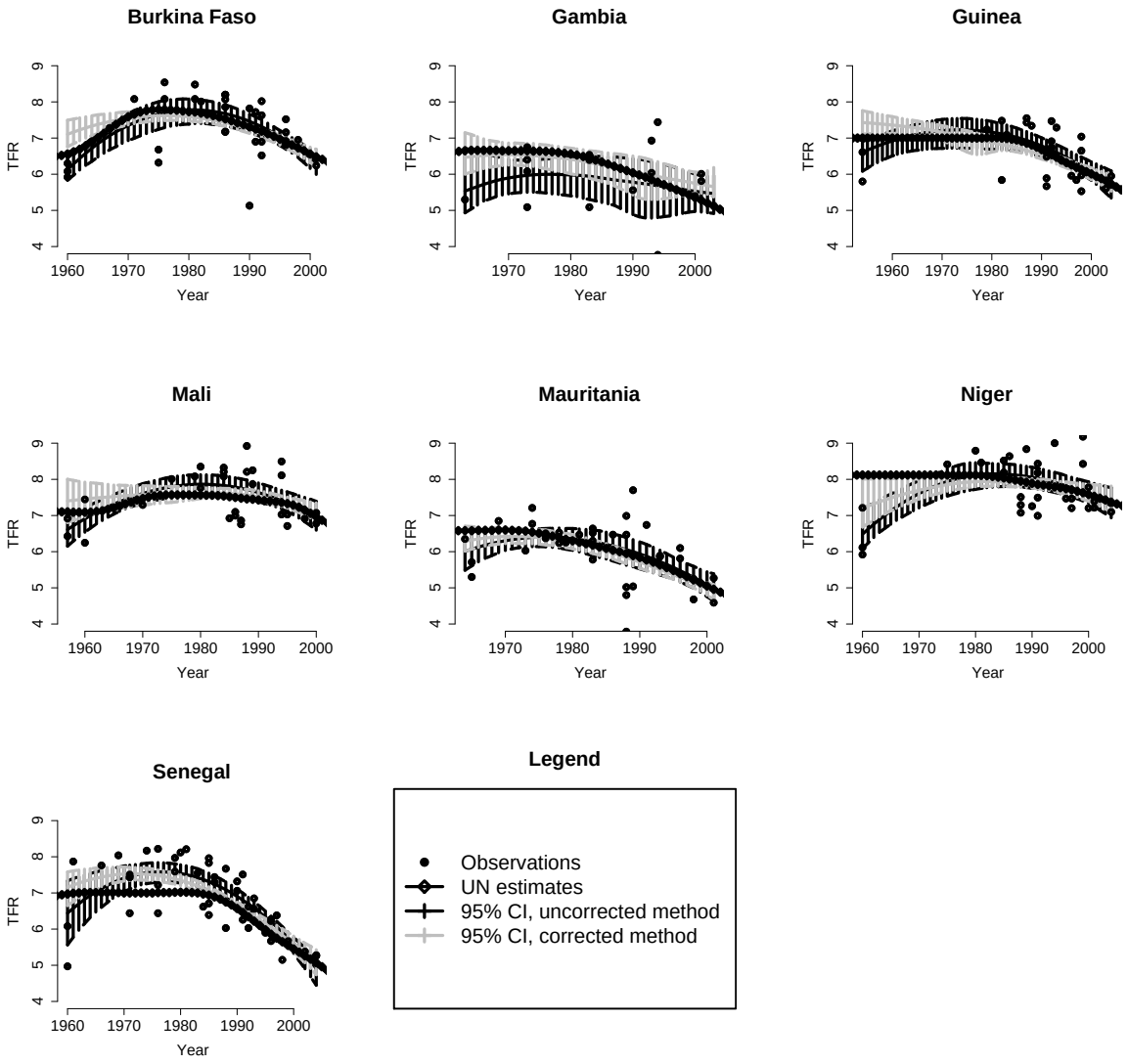

Figure 5: 95% Confidence intervals for the TFR for the corrected method (grey), and the uncorrected method (black). The solid line shows the annual median estimates and the 95% confidence intervals are plotted with dashed lines. The observations are displayed as black dots and the UN estimates by the grey line with squares.

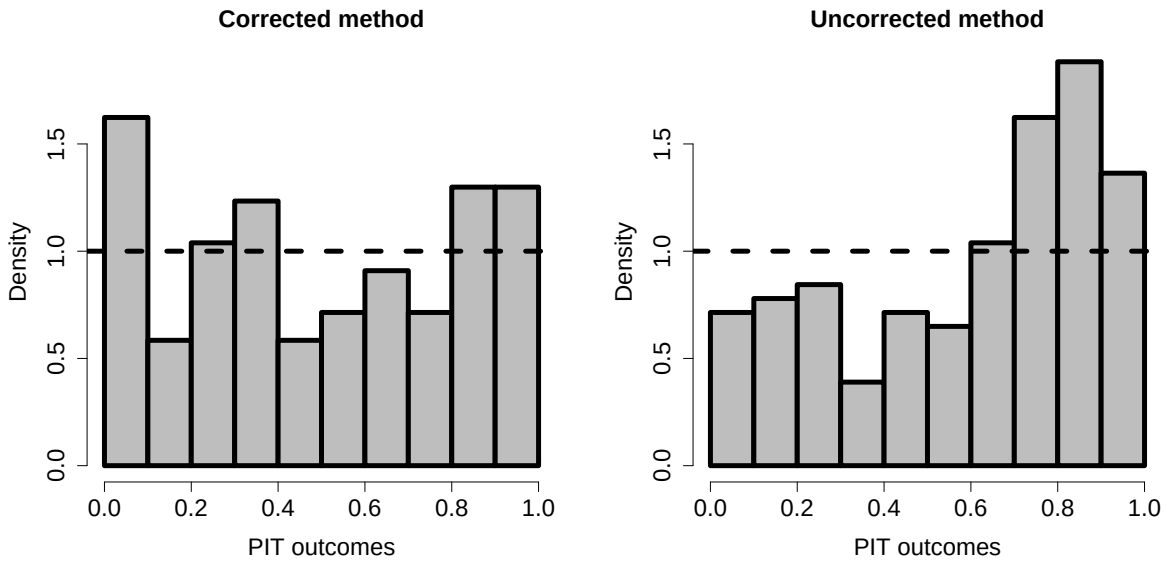

Figure 6: Histograms of the outcomes of the probability integral transforms for the left-out observations when leaving out one DHS at a time for the corrected and the uncorrected method.

5 Discussion

We have proposed a new approach to estimating the total fertility rate over time from multiple data sources of varying quality, and applied it to seven countries in western Africa. Our approach consists of four steps: bias adjustment, estimation of measurement error variance, local smoothing of the bias-adjusted values with weights based on the error variance, and uncertainty assessment using the weighted likelihood bootstrap. We assessed the results by cross-validation, and found that our method was reasonably well calibrated. Comparison with a similar method that excludes the first two steps showed that taking account of data quality removed clear biases and greatly reduced the average width of the confidence intervals.

One limitation of our method is that a data source that is deemed "unbiased" is needed to predict bias and measurement error variance. This data source does not have to be perfect or even of high quality, but is required to have no systematic tendency to substantially over- or underestimate TFR. We have used the existing UN estimates for this purpose. It would also be possible to identify one of the common data sources, most likely the DHS's with a short recall period, as unbiased. However, when using one single data source as a baseline for the analysis, data quality problems in that data source (such as those that we found for the 1990 DHS in Mauritania) may well be missed. Also, for estimating the TFR for multiple countries from the 1950s until the last year with observations, this data source will not be available for all years and all countries. In future work, we plan to investigate how to estimate the TFR and its uncertainty without using the UN estimates as an unbiased data source.

Our general approach could be applied with appropriate modifications to the estimation of other demographic quantities, including mortality. However, our methodology was developed for an aggregated rate (the TFR), while often age-specific rates are needed. The method could be applied directly to age-specific fertility rates, but this has the disadvantage that adherence to overall patterns is not guaranteed. A possible alternative would be to use our present methodology in combination with age-specific fertility schedules.

6 Acknowledgements

This research was partially supported by Grant Number 1 R01 HD054511 01 A1 from the National Institute of Child Health and Human Development. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Child Health and Human Development or those of the United Nations. Its contents have not been formally edited and cleared by the United Nations. This research was also partially supported by a seed grant from the Center for Statistics and the Social Sciences, by a Shanahan Fellowship at the Center for Studies in Demography and Ecology and by the Blumstein-Jordan professorship, all at the University of Washington. The authors are grateful to Thomas Buettner, Gerhard Heilig and Taeke Gjaltema for helpful discussions and insightful comments. Alkema thanks the United Nations Population Division for hospitality.

References

Alkema, L. (2008). *Uncertainty Assessments of Demographic Estimates and Projections*. Ph. D. thesis, University of Washington.

Becker, S. and S. Mahmud (1984). A validation study of backward and forward pregnancy histories in Matlab, Bangladesh. World Fertility Survey Scientific Reports 52, International Statistical Institute.

Berrocal, V., A. E. Raftery, and T. Gneiting (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review 135*, 1386–1402.

Brass, W. (1964). Uses of census or survey data for the estimation of vital rates. Paper presented at the African Seminar on Vital Statistics.

Brass, W. (1996). Demographic data analysis in less developed countries. *Population Studies 50*, 451–467.

Brass, W., A. J. Coale, P. Demeny, and D. F. Heisel et al. (eds.) (1968). *The Demography of Tropical Africa*. Princeton NJ: Princeton University Press.

Cleveland, W., E. Grosse, and W. Shyu (1992). Chapter 8.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*, 596–610.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics 7*, 1–26.

Ewbank, D. C. (1981). *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns, and Consequences for Demographic Analysis*. National Academy Press.

Feeney, G. (1998). A new interpretation of Brass' P/F Ratio method applicable when fertility is declining. http://www.gfeeney.com/notes/pfnote/pfnote.htm.

Gerland, P. (2007). Unpublished notes on Weighting scheme for emperical data on age-specific mortality and fertility rates, based on data quality covariates.

Hill, K., R. Pande, M. Mahy, and G. Jones (1998). Trends in child mortality in the developing world: 1960 to 1996. http://www.un.org/esa/population/publications/wwp2004/wwp2004_vol3_final/%6.pdf, UNICEF, New York.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Moultrie, T. A. and R. Dorrington (2008). Sources of error and bias in methods of fertility estimation contingent on the P/F Ratio in a time of declining fertility and rising mortality. *Demographic Research 19(46)*, 1635–1662.

Murray, C. J. L., T. Loakso, K. Shibuya, K. Hill, and A. D. Lopez (2007). Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. *Lancet 370*, 1040–1054.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B 56*, 3–48.

Potter, J. (1977). Problems in using birth-history analysis to estimate trends in fertility. *Population Studies 31*, 335–364.

Pullum, T. and S. Stokes (1997). Identifying and adjusting for recall error, with application to fertility surveys. In L. Lyberg and P. Biemer and et al. (Ed.), *Survey Measurement and Process Quality*, pp. 711–732. New York: John Wiley and Sons.

Pullum, T. W. (2006). An assessment of age and date in the DHS surveys, 1985 - 2003. Technical report. DHS Methodological Reports 5, http://www.measuredhs.com/pubs/pub_details.cfm?ID=664&srchTp=type.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology 25*, 111–196.

Raftery, A. E., I. Painter, and C. T. Volinsky (2005). BMA: An R package for Bayesian Model Averaging. *5*(2), 2–8. R News.

Silverwood, R. and S. Cousens (2007). Comparison of spline- and loess-based approaches for the estimation of child mortality. *Presentation at the Inter-Agency Coordination Group on Child Mortality Estimation, Geneva, 7 March 2007 Slides at http://www.who.int/whosis/mort/20080306mtg_Present_Day1_Session6_DrSilverwood.pdf*.

Som, R. (1973). *Recall Lapse in Demographic Enquiries*. New York: Asia Publishing House.

Trussell, T. (1975). A re-estimation of the multiplying factors for the Brass technique for determining children survivorship rates. *Population Studies 29*, 97–108.

UNICEF, the World Health Organization (WHO), World Bank, and United Nations Population Division (UNPD) (2007). Levels and trends of child mortality in 2006 estimates developed by the interagency group for child mortality estimation. *Working paper http://www.childinfo.org/files/infant_child_mortality_2006.pdf*.

United Nations (1982). National household survey capability programme. Non-sampling errors of household surveys: sources, assessment and control. *New York: United Nations Department of Technical Cooperation for Development and Statistical Office http://unstats.un.org/unsd/publication/unint/DP_UN_INT_81_041_2.pdf*.

United Nations (1983). *Manual X, Indirect Techniques for Demographic Estimation*. United Nations, New York. http://www.un.org/esa/population/publications/Manual_X/Manual_X.htm.

United Nations, Department of Economic and Social Affairs, Population Division (2007). *World Population Prospects. The 2006 Revision, Vol. I, Comprehensive Tables*. United Nations publication, Sales No. E.07.XIII.2, http://www.un.org/esa/population/publications/wpp2006/wpp2006.htm.