The crisscross method to evaluate data quality in fertility surveys

Bruno SCHOUMAKER

Draft – 25 April 2014

Population Association of America Meeting, Boston, 1-3 May 2014

Centre de recherche en démographie et sociétés

Université catholique de Louvain

bruno.schoumaker@uclouvain.be

Phone : +32 10 474136

Fax : +32 10 472952

1. Introduction

Since the 1970s, birth histories have become a major source for estimating fertility levels and trends in developing countries. With more than 300 surveys conducted since the late 1970s, the WFS and the DHS program have allowed tremendous progress in the knowledge of fertility levels, trends, and determinants. Despite the usefulness of birth histories, fertility estimates based on birth histories are potentially affected by various types of data quality problems (Arnold, 1990; Blacker, 1994; Goldman et al., 1985; Potter, 1977; Pullum, 2006): displacements of births, omissions of births, under/over sampling of high fertility women, misreporting of age. The detection of data quality problems and the evaluation of their impact on fertility indicators are not straightforward. While some severe issues can be detected relatively easily, others are more difficult to identify. Disentangling the different types of data quality problems is important to improve data collection, and to evaluate the impact of these problems on fertility.

In this paper, I compare direct and indirect estimates of recent fertility to evaluate the quality of birth history data. The direct estimates are based on birth histories and computed in the same way as the published estimates in DHS. The indirect estimates are computed using the crisscross method (Schmertmann, 2002), based on cohort parity increments between successive surveys. Comparisons between direct and indirect estimates are used (1) to show that the data are affected data quality problems, and (2) to help identify the type of data quality problems (omissions of recent births, displacements, omissions of early births, differences in sample implementation). It could potentially be used to (3) help correct fertility estimates.

After this introduction, I present the crisscross method (and related cohort parity increments methods) and the direct method in the second section. In the third section, I discuss possible effects of data quality on direct and indirect estimates, and how cohort parity increments methods have been used to evaluate data quality. Data are presented in the fourth section. In the fifth section, direct and indirect estimates are compared in a series of countries where several DHS have been conducted. In section 6, simulations are used to map the links between data quality problems in surveys and the differences between direct and indirect estimates. In the seventh section, three case studies (Vietnam, Ethiopia and Nigeria) are analyzed to illustrate how the method can help identify data quality problems. Results are briefly compared with other methods to evaluate data quality. Section 8 concludes.

The approach presented in this paper is intended to be used as a rough diagnostic that can be performed easily. Preliminary results suggest the method can highlight specific types of data quality problems, or indicate surveys that are problematic. Used in combination with other methods, it could help making more detailed diagnostics about data quality of birth history data.

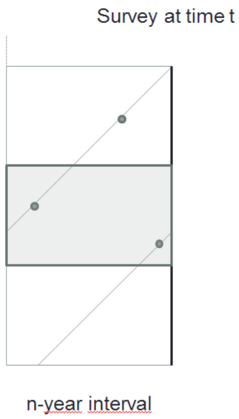
2. Computing recent fertility with direct and indirect methods

Recent age-specific fertility rates can be computed in a number of ways. Two methods are described below, and are compared in the rest of this paper in order to evaluate the quality of birth history data.

2.1. The direct method

The most common method with fertility surveys consists in computing fertility rates directly from birth history data (Moultrie 2013, Schoumaker 2013). The number of births that occurred in each age group over some period (e.g. three years preceding the survey) is computed, and is divided by the number of women-years (exposure) lived in each age group during the period (Rutstein and Rojas 2006, Moultrie 2013, Schoumaker 2013). Rates are computed between exact ages, and on periods of any length. In DHS, they typically refer to the three years or the five years preceding the survey, but they can be computed for shorter or longer periods (e.g. 6 or 7 years)[1].

Figure 1 : Illustration on the Lexis diagram of the computation of fertility rates with the direct approach.



2.2. Cohort parity increments approaches and the crisscross method

The estimation of period age-specific fertility rates from the increment of cohort parities is a classical indirect method for estimating fertility (Arretx 1973, Zlotnik and Hill 1981, UN Population Division 1983, Moultrie 2013). The central idea of that method is that the change in the parity in a cohort of women between two dates reflects the fertility rate during the time interval. Age-specific fertility rates can thus be inferred from changes in the average parity of women of the same cohorts over time (Arretx 1973, UN Population Division 1983, Moultrie 2013). For instance, comparing the average parity of women aged 30-34 in 2000 and women aged 25-29 in 1995 allows computing fertility rates for that cohort of women between 1995 and 2000. In the classical applications of cohort parity increments, surveys or censuses are five or ten years apart {Moultrie, 2013 #2}.

[1] In this paper, fertility rates are computed with the tfr2 module for Stata (Schoumaker, 2013), which allows computing fertility rates from birth histories in a flexible way on periods of any length.

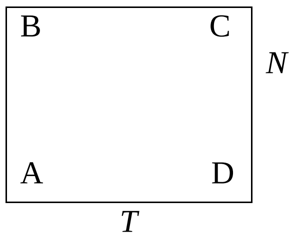
The crisscross method (Schmertmann 2002) can be viewed as a generalization of the earlier methods relying on cohort parity increments. In its application to fertility, the crisscross method also derives age-specific fertility rates form the comparisons of average parities at two points in time. It differs from the other methods in several respects. First, average parities are measured at exact ages (whereas it is usually measured in five-year age groups in the other methods). Secondly, the crisscross method is very flexible with regards to the time interval between surveys, and surveys need not be separated by 5 or 10 years. Third, age-specific fertility rates are computed between two exact ages and as a result are directly comparable to rates computed with the direct method, provided the same time period is used.

With the crisscross method, the fertility rate (λ) between two exact ages (x and x+*n*) over a period of any length *t*, is obtained in the following way (Eq. 1).

$$\lambda = \left(\frac{1}{2n} + \frac{1}{2t}\right).(C - A) + \left(\frac{1}{2n} - \frac{1}{2t}\right).(B - D) \qquad \text{(Eq. 1)}$$

A, B, C and D are the mean number of children ever born at exact ages and dates defined by the corners of the Lexis diagram, *t* is the time interval between the two surveys, and n is the width of the age group (Figure 2).

Figure 2 : Illustration of Lexis diagram and formula for estimating fertility rates with the crisscross approach (adapted from Schmertmann, 2002).



The mean number of children ever born is typically reported for age groups. Schmertmann (2002) suggests estimating the mean number of children ever born at exact ages as averages of parities of the neighboring 5-year age groups. In this paper, average parities by single year of age are used, and the relationship between age and parity is smoothed using Poisson regression with restricted cubic splines. Parities at exact ages are predicted from the regression coefficients[2].

3. Data quality and comparisons of crisscross and direct estimates

All measures of fertility – direct or indirect - are potentially affected by data quality problems. However, direct and indirect estimates are not affected in the same way by different types of problems. This is a key aspect of the method.

3.1. Types of data quality issues

In this paper, I distinguish six broad types of data quality problems that can affect fertility data and estimates. Only the first four will be taken into account in the method[3]. They cover problems that are considered to be frequent and important for our purpose. I briefly discuss these problems and the possible influence on direct and indirect estimates of recent fertility. The impact of data quality on fertility estimates is further discussed later with simulation results.

1) **Backward displacements of recent births**. This is a common issue in DHS surveys and is related to the lengthy health module (Pullum 2006), that encourages interviewers to displace some births backward to avoid administer this module. This may influence direct estimates of recent fertility if the period for which the rates are computed includes the years from which the births are displaced backward. However, the effect is likely to be small, because it is limited to a small portion of the period for which rates are computed. Crisscross estimates are not affected by displacements of births, since it relies solely on the total number of children ever born. The difference between the indirect and the direct estimate should be small.

---

[2] The results are not very sensitive to the way parities at exact ages are estimated.

[3] In this version of the paper, only the first four problems are taken into account.

2) **Omissions of recent births**. The same reasons that encourage interviewers to displace births may lead them to omit recent births. These data quality problems may lead to serious underestimation of recent fertility with direct estimates, although it is difficult to quantify this problem (Schoumaker 2011). Crisscross estimates will be little affected if omissions do not vary across surveys, but will be influenced if omissions vary across surveys. As shown later, the difference between indirect and direct estimates essentially depends on the degree of omissions in the first of the two surveys that are used for the crisscross method.

3) Omissions of early births. Early births may also be omitted, especially if the children have not survived. Such omissions are thought to increase with the duration between the births and the date of survey (UN Population Division 1983, Moultrie, Dorrington et al. 2013). Such omissions will have no or very limited influence on the direct estimates of recent fertility, since they concern early births. In contrast, omissions of early births will influence crisscross estimates. Constant omissions of early births across surveys will lead to underestimating fertility with the crisscross method, and indirect estimates will be lower than direct estimates. Varying omissions of early births may lead to indirect estimates that are lower or higher than direct estimates (and that true values).

4) **Sample implementation** may also influence fertility estimates. In DHS, great care is given to design representative random samples, but in the end some categories of women may be over- or underrepresented. If women with high fertility are underrepresented, direct estimates of recent fertility will be underestimated. Indirect estimates will also be influenced by sample implementation, and will be very sensitive to differences in sample implantation across surveys (Zlotnik and Hill 1981). Indirect estimates can be either greater or lower than direct estimates.

5) One specific problem of sample implementation found in DHS is the **exclusion of women from the eligible respondents**. Women aged 15-49

at the time of the survey are eligible for the individual interview. Some women aged 50 or over may be declared to be 50 in order not to be interviewed; in the same way, some women aged 15-19 may be declared to be younger, and will not be interviewed either (Pullum 2006). The impact on fertility estimates will depend on the link between the probability of being excluded and fertility. Direct estimates of fertility are likely to be little affected by these issues, because fertility levels at these ages are low. In contrast, crisscross estimates are likely to be more sensitive, especially at high ages. If women with higher parities are displaced to the 50+ age group, average parities in the last age group will be underestimated. If these issues vary across surveys, indirect estimates could be greater or lower than direct estimates.

6) **Incorrect age reporting** is also likely to affect fertility estimates (Zlotnik and Hill 1981). For instance, if women with high parities are declared to be older than they really are, parities will be underestimated at lower ages, and so will the crisscross fertility rates; in contrast, fertility rates will be overestimated at higher ages.

Another issue, which is not strictly speaking a data quality issue but may also influence fertility estimates is the selectivity of migration and mortality. Cohort parity increment methods consider that migration and mortality are not selective with regard to fertility (Zlotnik and Hill 1981)[4]. If the probability of dying increases with parity, high parity women will be underrepresented at high ages, and fertility estimates will be underestimated.

[4] Schmertmann (2002) presents applications of the crisscross method to other behavior (e.g. smoking cessation) for which this assumption is (and should) be lifted.

Table 1: Influences of different types of data quality problems on direct and indirect estimates of recent fertility.

	Method for estimating recent fertility	
	Direct	Crisscross
No problem	No influence	No influence
Backward displacements of recent births	Yes, probably slight underestimation	No influence
Omissions of recent births	Yes, underestimation	Yes (overestimation or underestimation), unless omissions are constant across surveys
Omissions of early births	No, or slight underestimation	Yes, underestimation or overestimation (underestimation if omissions are constant over time).
Sample implementation	Yes, underestimation or overestimation	Yes, underestimation or overestimation
Eligibility of women	Slight (probably underestimation)	Yes
Age reporting (overestimation of age of higher parity women)	Slight, underestimation at young ages, overestimation at higher ages	Slight, underestimation at young ages, overestimation at higher ages

3.2. Comparing direct and indirect estimates to evaluate data quality: a brief review

The idea of comparing direct and indirect estimates with the same data to evaluate data quality is not new. In the early 1980s, Zlotnik and Hill (1981, p.106), showed with simple simulations that estimates from cohort parity increments (called hereafter the hypothetical cohort) were very sensitive to changes in the completeness of information on births. Their discussion was based on a specific example, in which omissions were constant by age, and varied from one survey to the other[5]. They noted that "any error that which occurs equally in each observed data set will occur in equal magnitude in the synthetic data set; any change in error from one survey to the another,

[5] For instance, 1% of all births are omitted, regardless of age. As a result, the parity will be underestimated by 1% at all ages.

however, will be exaggerated in the synthetic data set" (Zlotnik and Hill 1981, p.106). Although the preceding assertion should be qualified – because it will depend on the type of data quality issue – their suggestion that this should not be "regarded entirely as a vice, since their sensitivity to error makes the technique proposed very useful in detecting it"[6] is a key point of this paper. The same idea was briefly discussed in Manual X[7].

Comparing fertility rates derived from the hypothetical cohort with parity increments and direct estimates to evaluate the quality of the data has also been performed by several authors. In the recent revision of the manual of indirect techniques, Moultrie (2013) compared estimates of the total fertility rates derived from parity increments and direct estimates from censuses, and found that indirect estimates were much higher than direct estimates suggesting the quality of reporting of recent fertility in the censuses was poor. Blacker (1994) also compared direct and indirect estimates from the same three surveys in Kenya to evaluate if the fertility decline was genuine. The way Blacker used the method – comparing recent fertility computed from birth histories (1978, 1984 and 1989 surveys) with indirect estimates for the same period using parities from the same three surveys – is very similar to what we do[8]. He showed that the fertility rates estimated from the parities were inconsistent with the recent fertility estimates, and clearly implausible, going from 9 children per woman to 5 children per woman in 5 years. He further showed that, when women in the 1989 survey were rejuvenated back to 1984, average parities were clearly lower in the latest survey. In other words, part of

[6] "Fairly small changes in the completeness of recording of parity from one survey to another can give rise to a set of parities for a hypothetical cohort that are clearly unacceptable" (Zlotnik and Hill, 1981, p.106).

[7] "Average parities for a hypothetical cohort are in fact very sensitive to changes in parity reporting from one survey to the other, and the calculation of such parities provides a useful consistency check of the raw data" (United Nations, 1983, p.58).

[8] The most notable difference is that we use the crisscross method instead of the classical cohort parity increment methods.

the issue came from reported parities that were lower than expected in the 1989 survey (or higher than expected in the previous survey), reflecting differences in sample composition (point 4) or omissions of early births (point 3) in the latest survey.

In the end, the existing applications indicate that comparing direct and indirect estimates is a useful device to check data quality. Since the two measures (direct and indirect) should match with good data, a difference between direct and indirect estimates point to data quality problems in one or two surveys. What is lacking – in my view – is a systematic investigation of the differences between direct and indirect estimates, and how such differences should be interpreted. For instance, is a higher TFR derived from indirect estimates than from direct estimates the sign of omissions of recent births, of displacements of births, of differences of sample composition?

A related question is whether crisscross estimates can be used to correct direct estimates. The idea that indirect estimates can be used to correct fertility levels obtained with direct methods has been developed in Manual X (UN Population Division 1983), as well as in the recent manual on tools for demographic estimation (Moultrie 2013). In essence, the approach is an extension of the P/F Method (Brass) for situations of changing fertility (Zlotnik and Hill 1981, p.107). P is the cumulated fertility in the hypothetical cohort computed from parity increments (P), and F is the cumulated fertility measured with direct estimates for the same period. The P/F ratios by age are used to compute an adjustment factor (e.g. average of the P/F ratios in the age range 20-29). The adjustment factor is then used to correct (upward) the fertility rates estimated with the direct method. The method presented in Moultrie (2013) is similar, and uses the relational Gompertz model to compare P and F. Both approaches rely on the assumption that the cumulated fertility in the hypothetical cohort up to age 30 or 35 is correct, i.e. that the parities reported by young women are accurate (Moultrie 2013). If one uses data from birth histories – as in this paper – there is however a contradiction between the assuming the correct reporting of parities and the use of the method to correct for omissions of recent births. By definition, if recent births have been omitted in the birth histories, parities will not be correctly reported. Moreover, while differences in sample composition do not matter with census data, it does with survey data. This means that using P/F ratios among young women to correct for the level of fertility is not necessarily appropriate – except under specific conditions. In this paper, we do not discuss further the correction of fertility estimates.

4. Data

We use Demographic and Health Surveys conducted since the early 1990s. Given than indirect estimates rely on two surveys, only countries where at least two surveys have been conducted can be included. In total 183 surveys from 49 countries can be used. The examples in this paper are selected from selected countries from various regions and with different types of data quality problems.

5. Comparisons of direct and indirect estimates

Direct and indirect estimates of fertility are compared for all the surveys in the selected countries. For each survey, age-specific fertility rates are computed directly from birth histories for the period between the survey and the preceding survey (or the last 5 years for the first survey). Age-specific fertility rates are computed for the same periods (except the first one) with the crisscross method. As in the P/F method, cumulated fertility (TFR) is computed for various age ranges with the direct method (F) and the indirect method (P). In this paper, we compute cumulated fertility (P and F) for three age ranges: 15-24, 15-34, and 15-49[9]. The P/F ratios are used to summarize differences between indirect and direct estimates.

---

[9] This could potentially be computed for the 7 age groups, and will be done in a future version.

Figure 3 and Figure 4 compare (in Rwanda and Madagascar) direct and indirect estimates of the age-specific fertility rates for different periods, as well as the fertility trends of cumulated fertility (15-24, 15-34 and 15-49).

In Rwanda, the crisscross estimates of the TFR vary erratically for the 15-49, and differ greatly from the direct estimates. In contrast, trends from direct and indirect estimates are in very good agreement for cumulated fertility up to 24 and 34. Age specific fertility rates show that rates are indeed very close for direct and indirect estimates up to age 35. After age 35, differences can become substantial, and rates estimated with the crisscross method are even negative between 2005 and 2008[10]. The Madagascar case is also characterized by large fluctuations of the indirect estimates of the TFR 15-49, with a very low TFR for the 1997-2003 period, probably reflecting an overrepresentation of low fertility women in the 2003 survey. Fluctuations of the indirect TFR 15-34 are much less pronounced, and estimates of fertility up to 24 are in close agreement for direct and indirect estimates.

Trends in direct and indirect estimates of cumulated fertility (15-24, 15-34 and 15-49) are represented for 12 other countries (Figure 5). Differences between direct and indirect estimates for fertility between 15 and 49 can be substantial. In Benin for instance, the most recent indirect estimate is implausibly low; the second indirect estimate of the TFR in Nigeria is also clearly implausibly high. In the end, direct and indirect estimates of fertility (15-49) match in very few cases, and fluctuations are much larger with the crisscross method than with the direct method. Indirect and direct estimates are in better agreement for cumulated fertility to lower ages.

[10] Negative rates at old ages arise when parities at a given age in the first survey are higher than parities at a higher age in the second survey, reflecting either omissions of early births in the second survey, or overrepresentation of low fertility women in the second survey (or overrepresentation of high fertility women in the first survey).

Figure 3 : Comparisons of direct and indirect estimates of age-specific fertility rates for four periods, and trends in cumulated fertility for three age ranges, Rwanda.

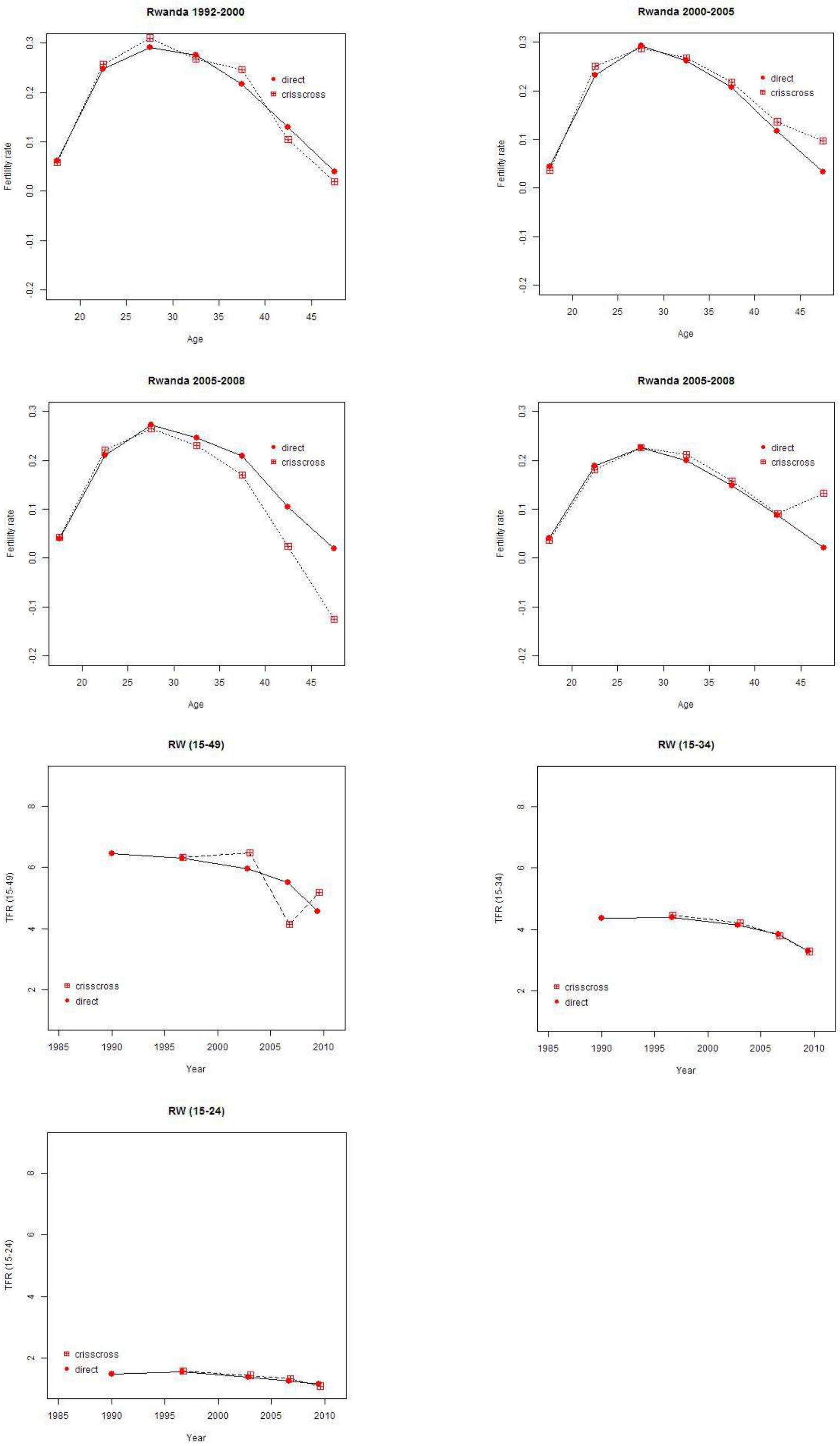
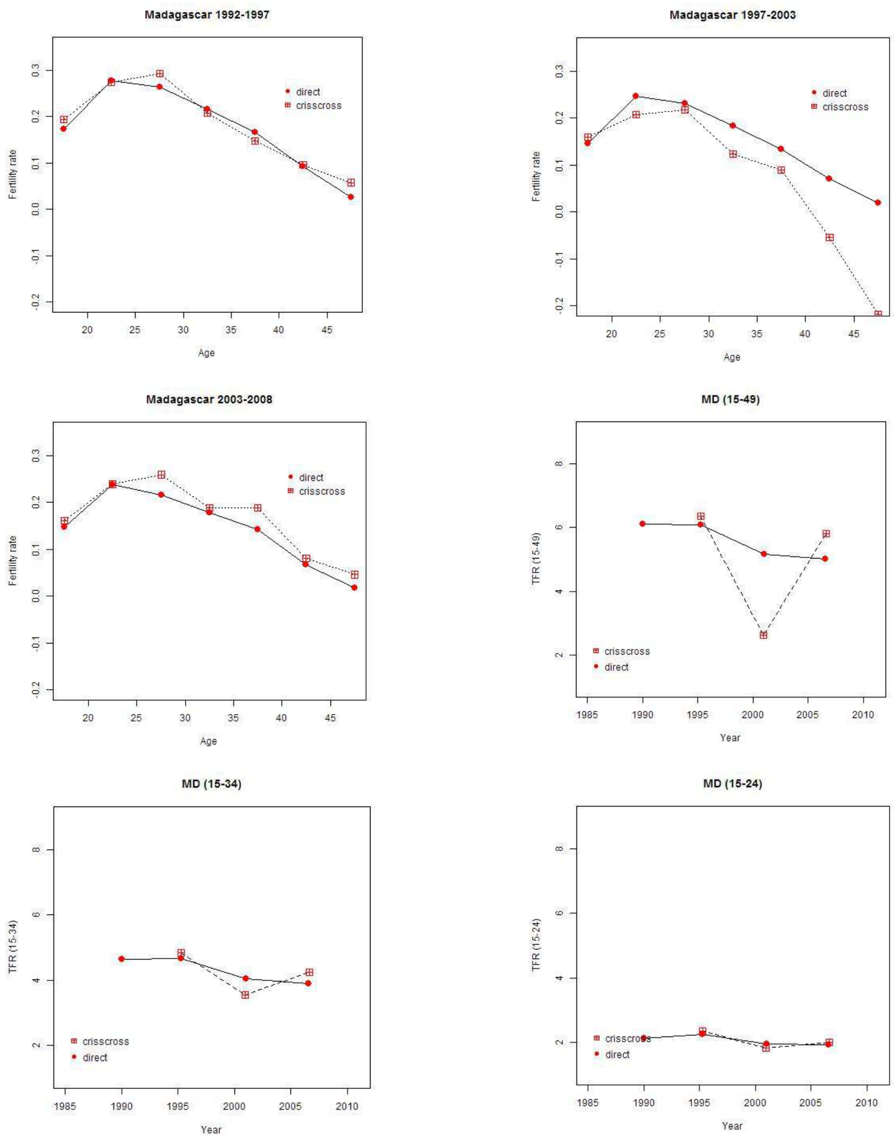


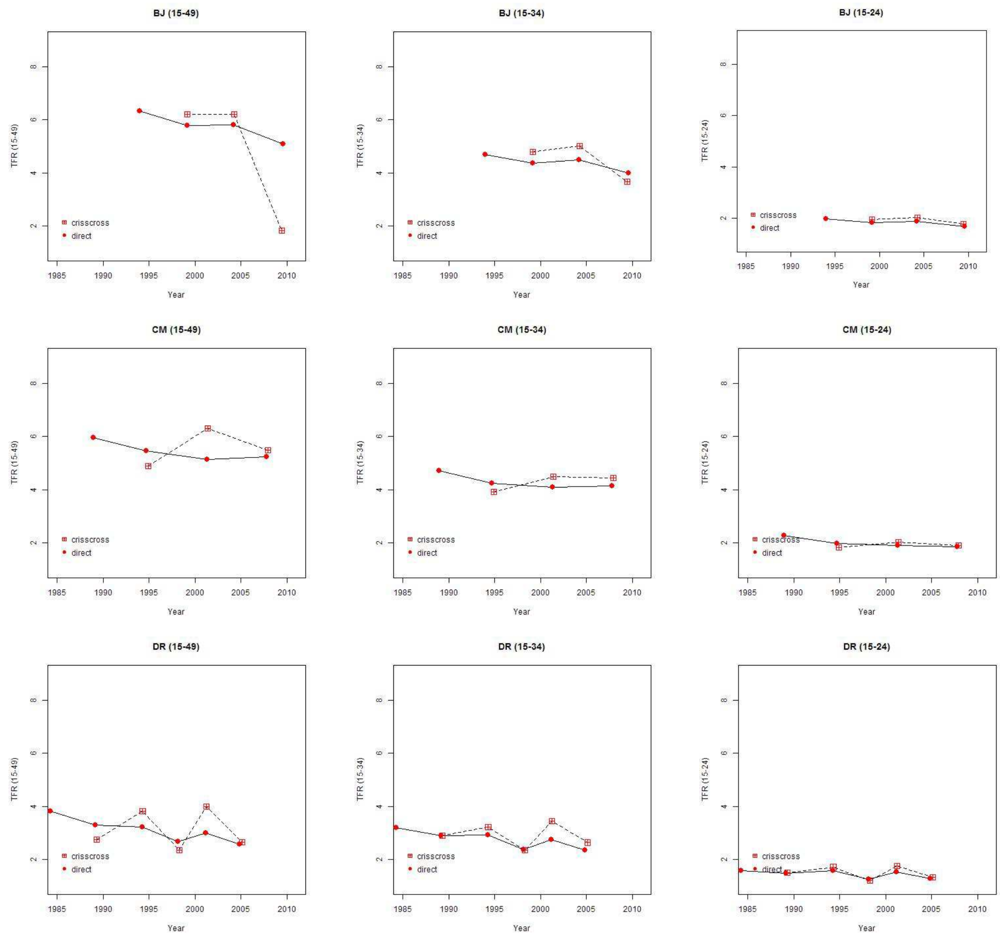
Figure 4 : Comparisons of direct and indirect estimates of age-specific fertility rates for three periods, and trends in cumulated fertility for three age ranges, Madagascar.

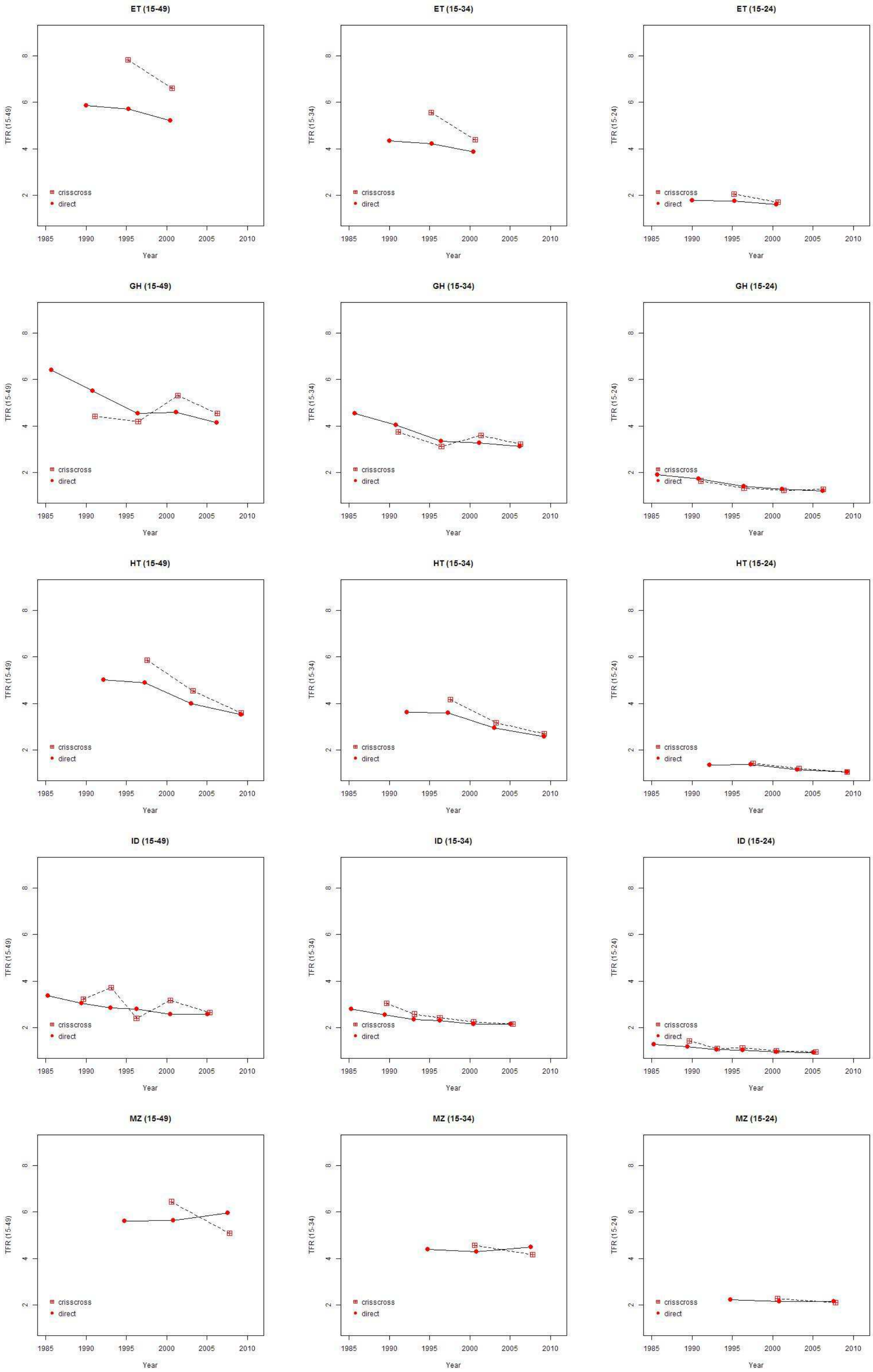


Analyses of all the surveys and all the countries (results not shown) further indicate that indirect estimates are on average higher than direct estimates (the P/F ratios tend to be positive). In addition, very few situations are found where indirect estimates are lower than direct estimates in two consecutive surveys. In contrast, a situation in which the indirect estimate is lower than the direct estimates is very likely to be followed by the opposite.

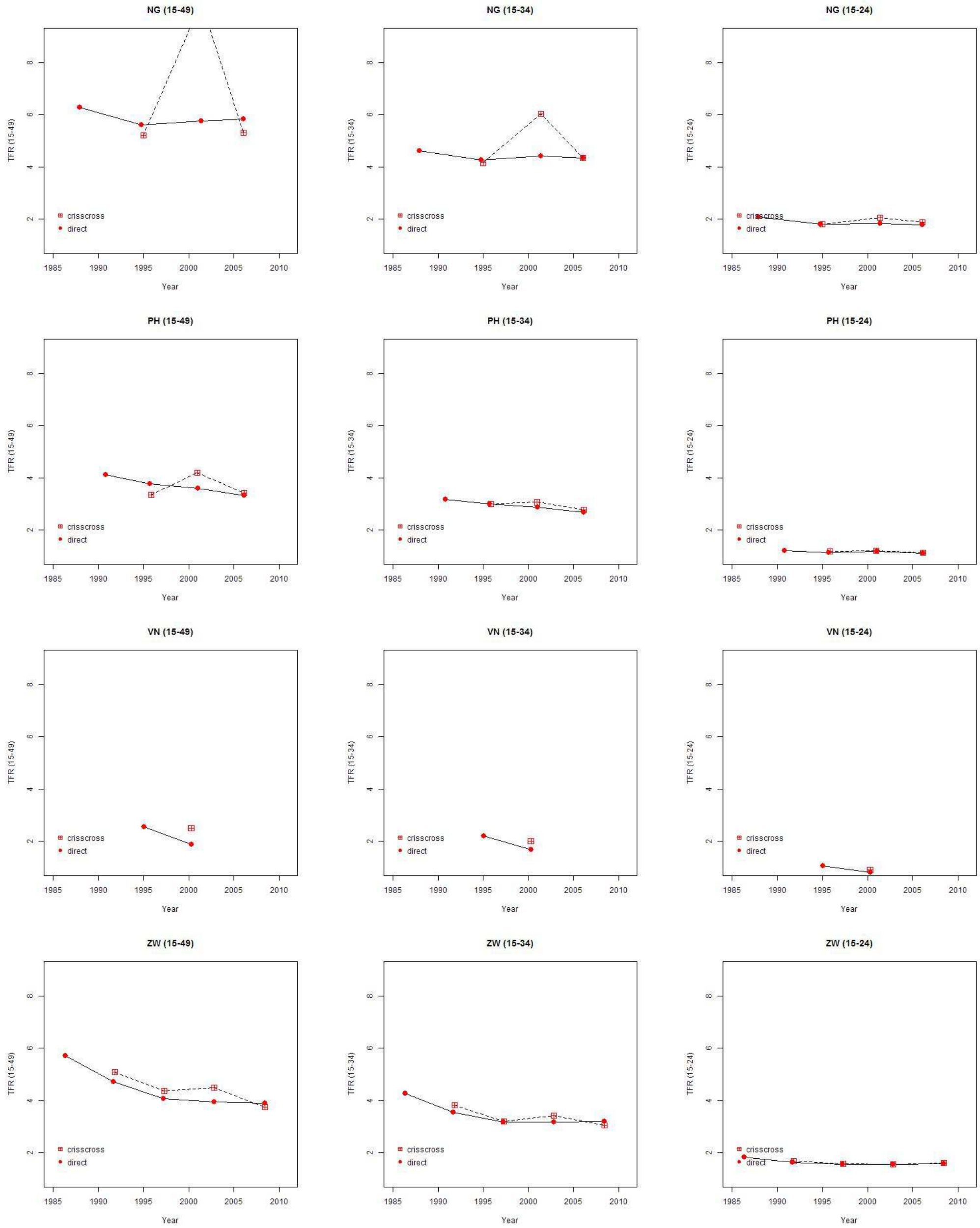
The key idea here is that these patterns reflect different types of data quality issues. A P/F ratio that increases or decreases strongly with age is likely to reflect differences in sample implementation or omissions of early births rather than recent omissions. Recent omissions, in contrast, would lead to P/F ratios that increase less steeply with age. Different types of data quality issues also translate into contrasting evolutions in the P/F ratios over time. For instance, varying omissions of recent births will lead to fluctuations of the P/F ratio, but the ratio will remain greater to one. Differences in sample composition may lead to wide fluctuations of the P/F ratio, which may be much below 1 or much greater than one.

Figure 5 : Trends of cumulated fertility rates for three age ranges, comparisons of direct and indirect estimates in 12 countries.





ET (15-49)
ET (15-34)
ET (15-24)
GH (15-49)
GH (15-34)
GH (15-24)
HT (15-49)
HT (15-34)
HT (15-24)
ID (15-49)
ID (15-34)
ID (15-24)
MZ (15-49)
MZ (15-34)
MZ (15-24)
TFR (15-49)
TFR (15-34)
TFR (15-24)
Year
crisscross
direct



BJ: Benin; CM: Cameroon; DR: Dominican Republic; ET: Ethiopia; GH: Ghana; HT: Haiti; ID: Indonesia; MZ: Mozambique; NG: Nigeria; PH: The Philippines; VN: Vietnam; ZW: Zimbabwe.

6. Mapping the links between data quality problems and P/F ratios

6.1. Simulations

Simulations are used to 'map' the links between data quality problems and differences between direct and indirect estimates of fertility. The objective of this exercise is to make an inventory of the combinations of data quality problems and the associated values of P/F ratios (indirect/direct). This can then be used to identify the combinations of data quality problems that are compatible with the observed P/F ratios.

This inventory is done in two stages. First, data quality problems were introduced in simulated individual birth histories[11]. Several types of data quality problems were combined randomly for each survey, and two consecutive surveys with different data quality problems were also combined randomly (see Box 1 for an explanation of the data quality problems). In total, 50 000 combinations of 2 surveys were created (25 000 combinations of the first and second survey, and 25 000 of the second and the third surveys).

For each of these 50 000 simulations, we know (1) the data quality problems that were introduced in the birth histories, (2) the direct and indirect estimates of age-specific fertility, as well as (3) the true value of age-specific fertility rates. P/F ratios are computed from the indirect and direct estimates of fertility for three age ranges (15-24, 15-34 and 15-49) to summarize differences between direct and indirect estimates.

The 50 000 combinations represents a relatively small sample of possible combinations of data quality problems. In a single survey, 10 000 combinations of data quality problems are possible if we consider 10 degrees of intensity for each of the four problems; 100 million combinations are possible for two surveys. The second stage consists of creating a much larger set of combinations of data quality problems and of the associated P/F ratios. Instead of simulating such a large number of combinations of birth histories –

[11] SOCSIM is used to create 3 pseudo-surveys conducted 5 years apart. Each simulated survey contains birth histories for a sample of around 10 000 women. These birth histories are free from data quality problems. Direct and indirect estimates of TFRs are identical, and are equal to the true value of fertility.

which is time consuming - it is possible to use a regression model to predict P/F ratios from data quality problems.

Box 1. Description of data quality problems in simulated birth histories

Four types of data quality problems are considered, corresponding to the first four problems discussed in point 3.1[12].

- **Displacements of recent births**. In these simulations recent births are those that occurred in the five years preceding the survey. Births are displaced from the fifth year before the survey (this corresponds roughly to the first year of the health module in DHS) to the year just before. In simulations, the percentage of displaced births varies from 0 to 40%.
- **Differences in sample implementation**. This is operationalized by randomly removing a percentage of births regardless of the age and the time at which they occurred. The percentage of removed births varies from 0 to 20%. A high percentage of removed births corresponds to an overrepresentation of low fertility women in the sample.
- **Omissions of early births**. Here, we consider that the percentage of omitted births increases linearly with the number of years before the survey. As a result, the percentage of omissions increases with the age of the mother, and omissions at a given age are also more frequent further back in time. It is maximum 35 years before the survey (at age 15 for women aged 50 at the time of the survey). The maximum percentage varies from 0 to 20%.
- **Omissions of recent births**. Some recent births (last 5 years) are omitted, as it is thought to occur in DHS as a consequence of the lengthy heath module. The percentage of omissions of recent births is independent from the age of the respondent, and ranges from 0 to 20%.

The results of the 50 000 simulations of pairs of surveys are used to fit three linear regression models of the P/F ratios (one for each age range) on the intensity of the four data quality problems in each of the two surveys (8 variables). These models have very large R^2 (close to 99%); in other words,

---

[12] The other 2 problems related to age declarations have not been taken into account be will be in further analyses.

the data quality problems predict almost perfectly the P/F ratios. A large number of combinations of data quality problems of varying intensity can be prepared (e.g. 10 million), and P/F ratios are predicted from the regression coefficients in a straightforward way.

Regression results are useful for predictions, but are also interesting in their own right, as they indicate how data quality problems influence the P/F ratios (Table 2). As expected, the P/F ratios for ages 15-49 are very sensitive to differences in sample composition: a difference of 1% in the level of fertility leads to an increase/decrease of the P/F ratio by more than 5%. If the sample composition is similar across surveys, the P/F ratios will not be affected (the sum of coefficients is very close to zero). The P/F ratios are also very sensitive to omissions of early births. Interestingly, even if the percentage of omissions of early births is similar across surveys, the P/F ratio will be lower than one (the sum of the coefficients is negative). The P/F ratio will also depend on the degree of omissions of recent births in the first survey, but very little in the second survey. This is simply explained: if recent births are omitted in the second survey, it will lead to underestimating both the direct estimate and the indirect estimate. In contrast, omissions of recent births in the first survey will lead to underestimating parities in the first survey, and thus overestimating fertility with the crisscross method. In contrast, the direct estimate will not be affected by omissions of recent births in the first survey. Finally, displacements of births have a limited impact on P/F ratios. Displacements have no effects on the indirect estimates, and lead to a small underestimation of the direct estimate of fertility. As a result, the P/F ratio will slightly increase with displacements of births in the second survey. The P/F ratios at 15-34 are less sensitive to omissions of early births, and are more sensitive to differences in sample composition and to omissions of recent births. Finally, the P/F ratios at 15-24 only depend on omissions of recent births and on differences in sample composition.

Table 2. Linear regressions of the PF ratios on indicators of data quality.

	P/F 15-24	P/F 15-34	P/F 15-49
Constant	0.963	0.963	0.9744
First survey			
Displacements	-0.004	-0.017	-0.025
Sample composition	0.430	1.777	5.631
Early omissions	0.025	0.266	1.804
Recent omissions	0.414	0.924	1.265
Second survey			
Displacements	0.211	0.227	0.229
Sample composition	-0.431	-1.778	-5.640
Early omissions	-0.080	-0.523	-2.605
Recent omissions	0.046	0.094	0.079
R^2	0.986	0.988	0.988

Note: indicators of data quality are expressed in percentages. The coefficients should be interpreted in the following way: a 1% increase in recent omissions in the first survey is associated with an increase of the P/F ratio at 15-49 of 1.265.

In summary, these results show that P/F ratios are more sensitive to some types of issues than to others, and that the effects of the data quality problems on P/F ratios depend on the age group.

7. Linking observed situations with simulations

With real data, we only observe direct and indirect estimates of age-specific fertility rates. The idea of the method is to use the simulations to infer data quality problems from the observed P/F ratios. This is illustrated with three countries. In the first country (Vietnam), two DHS were conducted, and P/F ratios are available only at one point. In the other two examples (Ethiopia and Nigeria), three or more surveys were conducted. Results for additional countries are presented in appendix 1.

7.1.1. Case study 1: two surveys in Vietnam

The Vietnam situation is shown on Figure 5, for the two surveys conducted in 1997 and 2002. The crisscross estimate is higher than the direct estimate at ages 15-49 and 15-34 and more slightly at ages 15-24 (Table 3).

Table 3: P/F ratios, two surveys (Vietnam)

Surveys	P/F 15-24	P/F 15-34	P/F 15-49
1-2	1.101	1.187	1.335

10 million combinations of data quality problems in pairs of surveys are generated randomly, and the regression models are used to predict the P/F ratios for each age range. The distance between the observed P/F ratios and the P/F ratios in simulations is computed as the square root of the sum of squares of the differences between these quantities. The 1000 simulations that lead to the P/F ratios that are the closest to the observed P/F ratios are selected[13]. The table below shows the PF ratios in the closest simulation and in the 1000th closest.

Table 4: P/F ratios in the closest simulation and the 1000th closest simulation, (Vietnam)

Surveys	P/F 15-24	P/F 15-34	P/F 15-49
Closest	1.100	1.186	1.335
1000th closest	1.104	1.193	1.333

Potential data quality problems in each survey are then inferred from these 1000 simulations. The distribution of the data quality problems in the 1000 simulations is represented for each survey (Figure 6, each line corresponds to a survey); the median value of the distribution and the 10th and 90th percentiles are also computed (Figure 7). In this case, distributions are not very concentrated, and no strong conclusion emerges. There are however signs of

[13] Given that several combinations of data quality problems (simulations) may lead to very similar patterns of P/F ratios, it is more relevant to select a sample of simulations rather than the closest simulation.

omissions of recent births in the first survey, and displacements of births in the second survey.

Figure 6 : Distribution of the data quality indicators (1000 simulations) in two surveys in Vietnam (DHS 1997, DHS 2002).

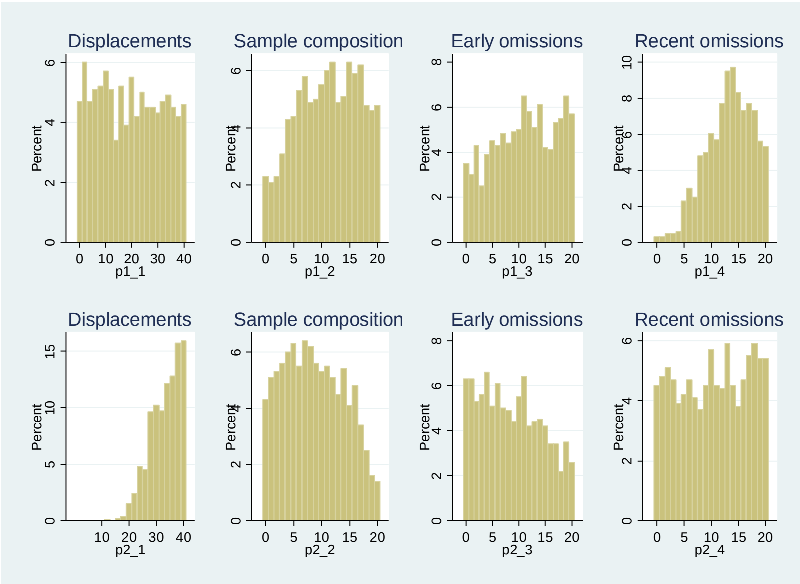
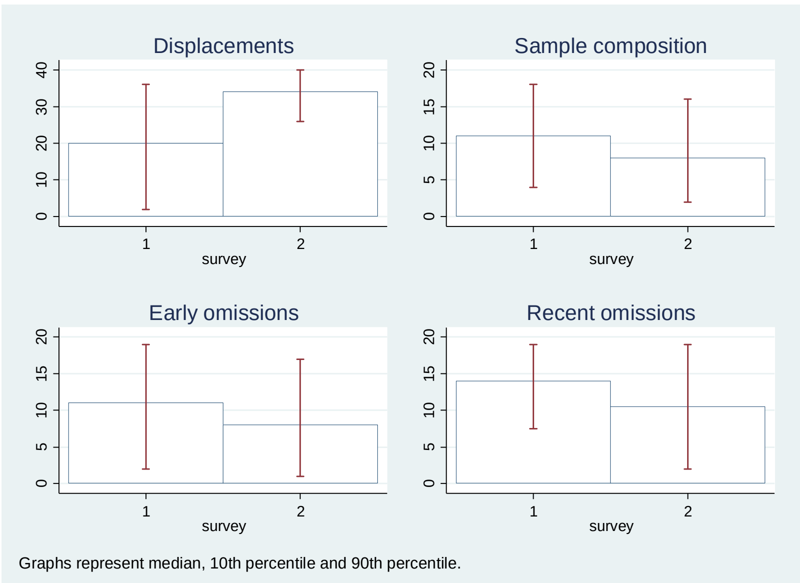


Figure 7 : Summary measures (median, p10 and p90) of the data quality indicators (1000 simulations) in two surveys in Vietnam (DHS 1997, DHS 2002).



7.1.2. Case study 2: Three surveys in Ethiopia

The second case study is Ethiopia. The indirect estimates of the TFR are greater than the direct estimates for the two points in time regardless of the age range considered (Figure 5). The P/F ratios are shown in Table 5.

Table 5: P/F ratios, two surveys (Ethiopia)

Surveys	Ratio 15-24	Ratio 15-34	Ratio 15-49
1-2	1.16	1.32	1.37
2-3	1.07	1.13	1.27

The following steps are used. As in Vietnam, 10 million combinations of data quality problems in pairs of surveys are generated randomly. The P/F ratios are predicted with the regression model. The distance between the observed P/F ratios and the P/F ratios in simulations is computed for the first pair of surveys, as in the Vietnam case, and the 1000 simulations that lead to the closest P/F ratios are selected[14]. These simulations provide 1000 possible combinations of data quality issues in the first and second survey that are compatible with the observed P/F ratios computed from these surveys. For each of these 1000 selected simulations, 10 000 combinations of data quality problems are randomly generated for the third survey (leading to a total of 10 million combinations), and P/F ratios are predicted with the regression models in the same way as in the first step. Among these 10 million combinations, the 1000 combinations that lead to P/F ratios that are closest to observed P/F ratios are selected. Combining three surveys helps reducing uncertainty in the diagnostic for the second survey, since it influences the two sets of P/F ratios.

[14] Given that several combinations of data quality problems (simulations) may lead to very similar patterns of P/F ratios, it is more relevant to select a sample of simulations rather than the closest simulation.

The distribution (1000 simulations) of the data quality indicators in each survey are shown on Figure 8. The median value of the distribution and the 10th and 90th percentiles are also computed (Figure 9). These results suggest several data quality issues. First, it seems the first survey was affected by severe omissions of recent births. Secondly, results suggest the sample implementation has varied across surveys, with low fertility women overrepresented in the first survey compared to the second and the third survey. Omissions of early births seem to have been strong in the second survey, as were displacements.

Although these results are only indicative, they suggest that recent fertility was underestimated in the first survey for two major reasons: omissions of recent births and overrepresentation of low fertility women. In the second survey, recent fertility may have underestimated (to a lesser extent) because of omissions of recent births, and the third survey seems less affected by omissions of recent births.

Figure 8 : Distribution of the data quality indicators (1000 simulations) in three surveys in Ethiopia (DHS 2000, DHS 200, DHS 2011).

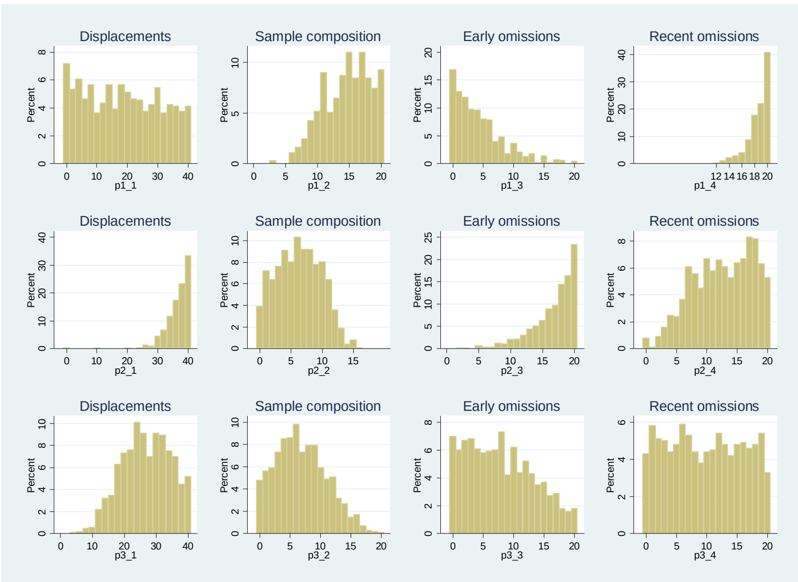
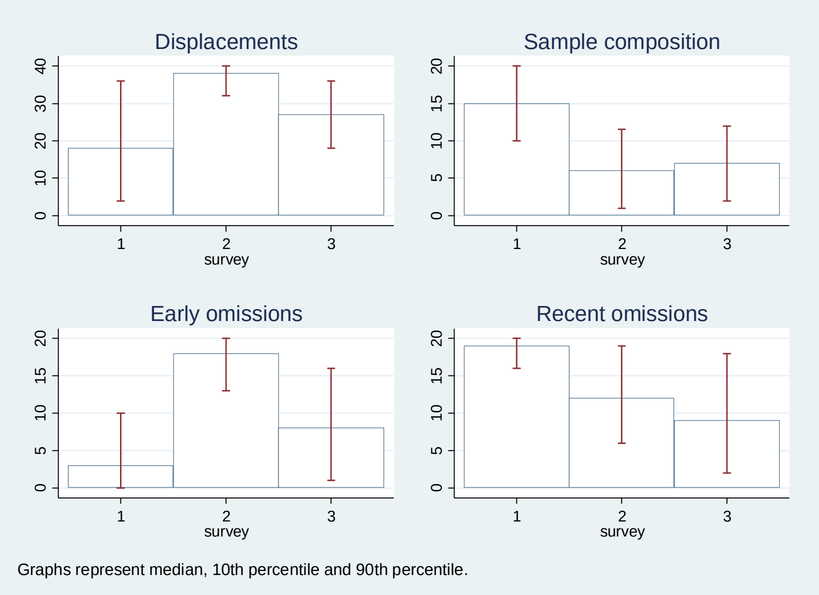
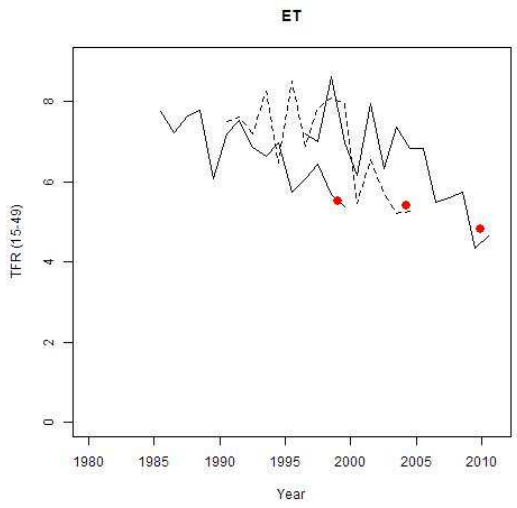


Figure 9 : Summary measures (median, p10 and p90) of the data quality indicators (1000 simulations) in three surveys in Ethiopia (DHS 2000, DHS 2005, DHS 2011).



Another way of evaluating the data quality consists in comparing fertility trends by single year from successive surveys (Schoumaker, 2013). Figure 10 shows the trend in the TFR for the 15 years preceding each of the three DHS in Ethiopia. Red dots indicate published fertility rates. Overall, these comparisons also suggest there were omissions of recent births in Ethiopia: recent fertility rates are lower than rates at the same dates computed from the next surveys. Omissions of recent births seem to be greater in the first than in the second survey. From these graphs it is difficult to ascertain whether there were omissions in the latest survey. The difference in sample implementation also seems plausible from this figure, with the overall level of fertility lower in the first survey. Greater displacements in the second survey also seem plausible. In further analyses, we will apply methods discussed in Schoumaker (2011) that allow estimating omissions of recent births, displacements and difference in sample implementation by pooling surveys together. This would allow confirming diagnostics based on the crisscross method.

Figure 10 : Reconstructed fertility trends (TFR 15-49) by single calendar year, three DHS, Ethiopia (red dots represent published TFRs)



7.2. Case study 3: Four surveys in Nigeria

Nigeria is the third case study. As shown on Figure 5, indirect estimates vary strongly, suggesting some serious data quality issues. P/F ratios are shown below.

Table 6: P/F ratios, four surveys (Nigeria)

Surveys	Ratio 15-24	Ratio 15-34	Ratio 15-49
1-2	1.00	0.97	0.93
2-3	1.12	1.36	1.79
3-4	1.06	1.00	0.91

The same procedure as in Ethiopia is used, with the second step repeated one more time because 4 surveys are available in Nigeria. The 1000 simulations that lead to P/F ratios that are closest to the observed P/F ratios are selected.

Figure 11 : Distribution of the data quality indicators (1000 simulations) in four surveys in Nigeria (DHS 1990, DHS 1999, DHS 2003, DHS 2008)

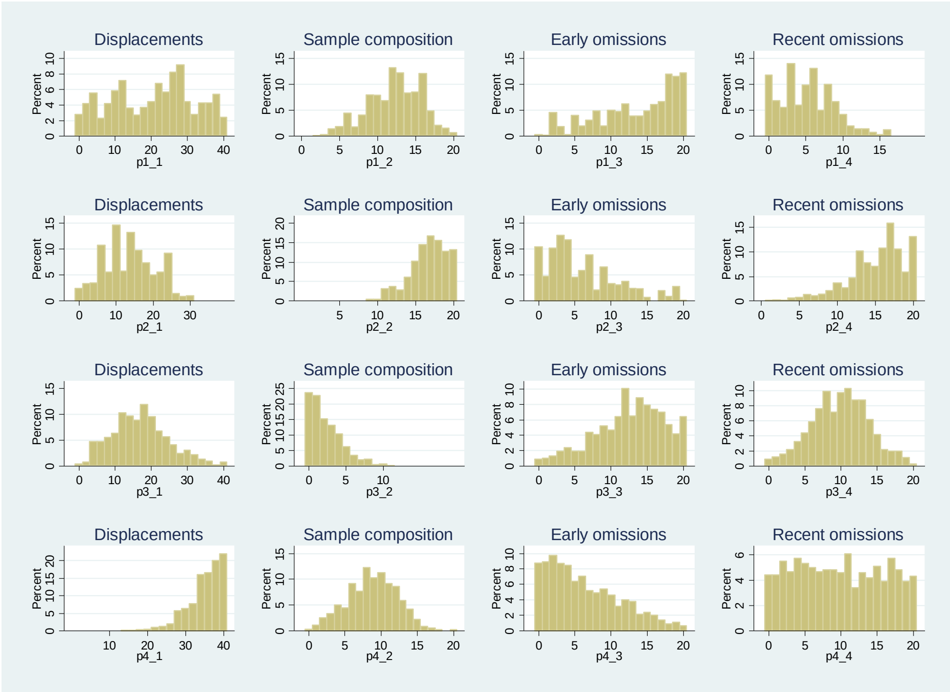
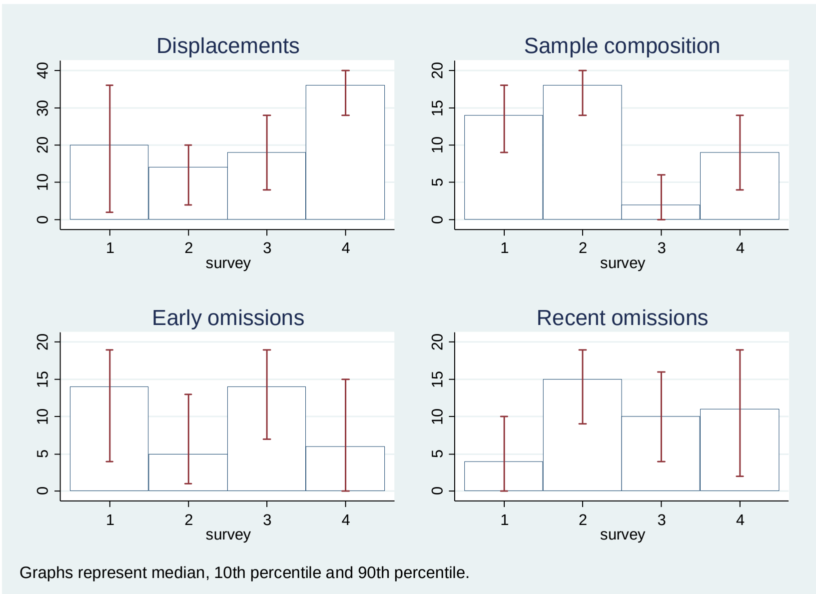


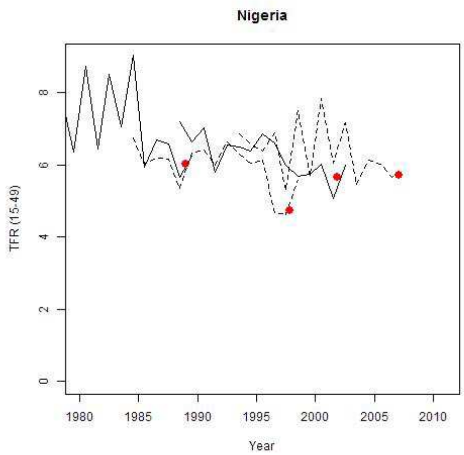
Figure 12 : Summary measures (median, p10 and p90) of the data quality indicators (1000 simulations) in four surveys in Nigeria (DHS 1990, DHS 1999, DHS 2003, DHS 2008).



Results strongly suggest that the sample composition has varied across surveys. Low fertility women were overrepresented in the second survey, while they were underrepresented in the third survey. It seems omissions of recent births were also more frequent in the second survey, contributing to the underestimation of fertility. Omissions of early births also varied across surveys, and differences between the third and the fourth surveys are strong (early births are less likely to have been omitted in the fourth survey than in the third survey). Finally, these results also suggest that displacements of births were strongest in the latest survey.

As in Ethiopia, this rough diagnostic is consistent with the reconstructed fertility trends (Figure 9). The second survey (1999) is below the others (reflecting a sampling issue), and the drop in fertility in the few years before the survey may reflect omissions of births. Recent omissions in the latest two surveys seem less pronounced than in the second survey; in contrast, the reconstructed trends suggest important omissions in the first survey. This was not clearly shown by the direct and indirect comparisons.

Figure 13 : Reconstructed fertility trends (TFR 15-49) by single calendar year, four DHS, Nigeria (red dots represent published TFRs)



8. Preliminary conclusions

The method described here consists in using ratios of crisscross and direct estimates of cumulated fertility up to different ages (P/F ratios) to provide a rough diagnostic of data quality issues. It relies on the fact that different data quality problems influence direct and indirect estimates in different ways.

Simulations are used to map the links between data quality problems and P/F ratios for different age ranges. Observed P/F ratios are interpreted using simulation results.

Tests of this method on several countries suggest that it is useful to provide a quick diagnostic of data quality. It points to potential data quality issues that can influence fertility levels and trends. This quick diagnostic can be complemented with other methods. At this stage, the method is still quite experimental. Further research will be done to validate the method. First, the method will be tested with simulated data for which data quality problems are known. The percentages of omissions, displacements etc. should fall between the 10^{th} and 90^{th} percentile of the distribution most of the time. In the same way, surveys with no data quality problems should ideally be identified as such. Testing the method in wide variety of situations, and comparing the diagnostics with other methods (as in Ethiopia and Nigeria) will also allow evaluating the validity of this approach. Applying to subpopulations (e.g. educated vs. uneducated) would be another research avenue. The way the selected simulations are interpreted can also be improved. Currently, distributions of data quality indicators are done separately. Other methods could be used to group patterns that are close to each other, and obtain distributions of *combinations* of data quality problems, rather than separate data quality problems. Using this method as a basis for correcting estimates of recent fertility estimates is another possible research avenue. Finally, for the method to be used, it should be easy to use. A user-friendly tool that performs the computation of the direct and crisscross TFRs, and that compares the observed pattern to the simulated patterns is under construction.

9. References

Arretx, C. (1973). Fertility estimates derived from information on children ever born using data from censuses. International Population Conference, Liège, International Union for the Scientific Study of Population.

Blacker, J. (1994). "Some thoughts on the evidence of fertility decline in Eastern and Southern Africa " Population and development review **20**(1): 200-205.

Moultrie, T. (2013). Direct estimation of fertility from survey data containing birth histories. Tools for Demographic Estimation. T. Moultrie, R. Dorrington, A. Hill et al. Paris, IUSSP**:** 110-117.

Moultrie, T. (2013). Fertility estimates derived from cohort parity increments. Tools for Demographic Estimation. T. Moultrie, R. Dorrington, A. Hill et al. Paris, IUSSP**:** 103-108.

Moultrie, T. (2013). Synthetic relational Gompertz models. Tools for Demographic Estimation. T. Moultrie, R. Dorrington, A. Hill et al. Paris, IUSSP**:** 96-102.

Moultrie, T., et al. (2013). Tools for Demographic Estimation. Paris, IUSSP.

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

Rutstein, S. O. and G. Rojas (2006). Guide to DHS Statistics. Demographic and Health Surveys. Calverton, ORC Macro.

Schmertmann, C. (2002). "A simple method for estimating age-specific rates from sequential cross sections." Demography **39**(4): 287–310.

Schoumaker, B. (2011). Omissions of recent births in DHS birth histories. Measurement and determinants. PAA Meeting. Washington DC.

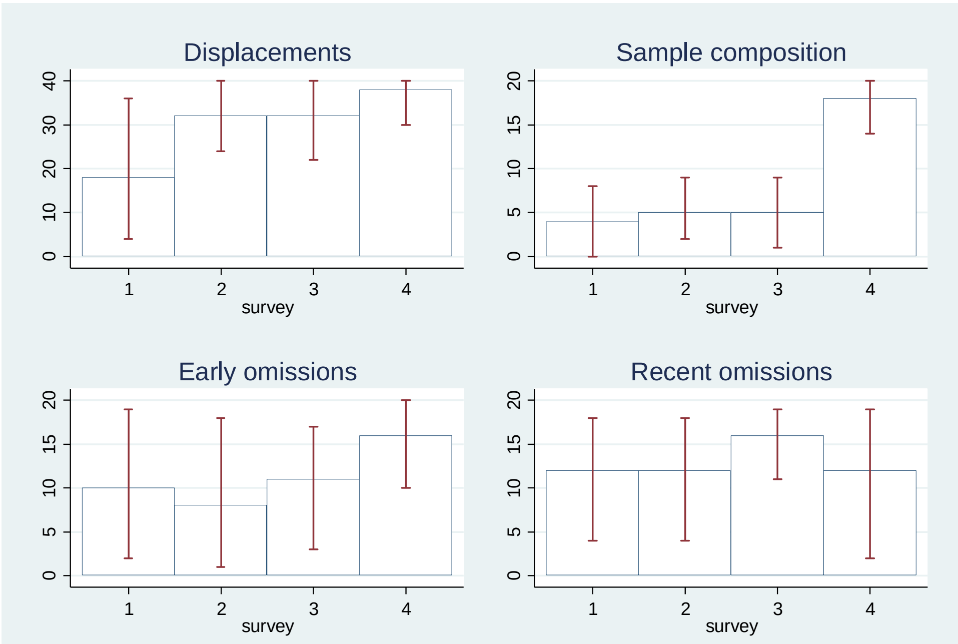
Schoumaker, B. (2013). "A stata module to compute fertility rates and TFRs from birth histories: tfr2." Demographic Research **28**(38): 1093-1144.

UN Population Division (1983). Manual X: Indirect Techniques for Demographic Estimation. New York, United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81.

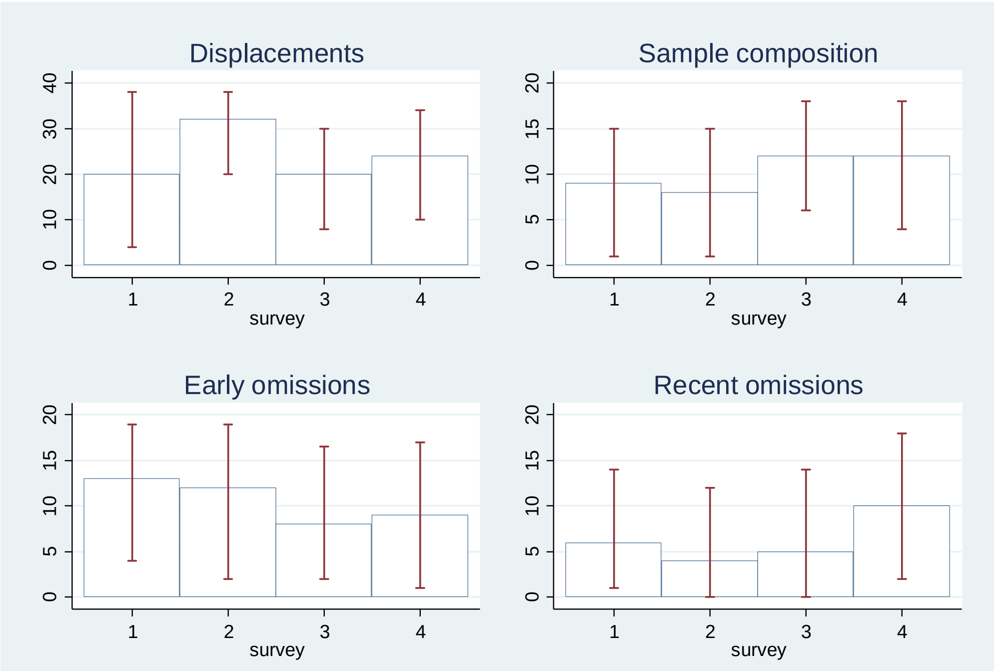
Zlotnik, H. and K. Hill (1981). "The Use of Hypothetical Cohorts in Estimating Demographic Parameters Under Conditions of Changing Fertility and Mortality." Demography **18**(1): 103-122.

Appendix 1. Diagnostics in selected countries

Benin

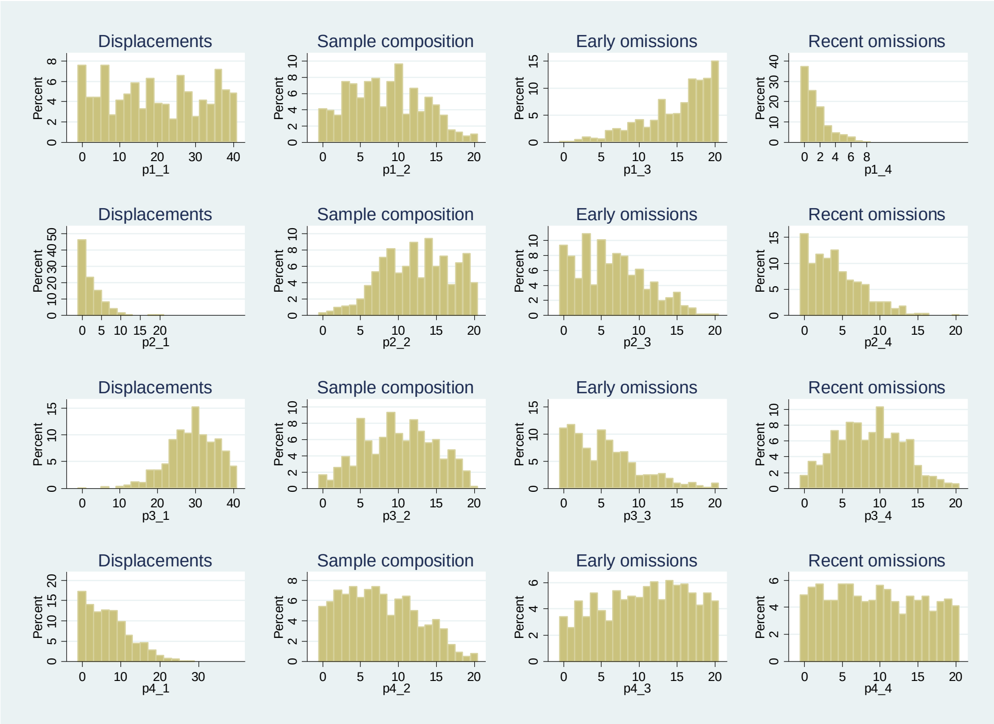


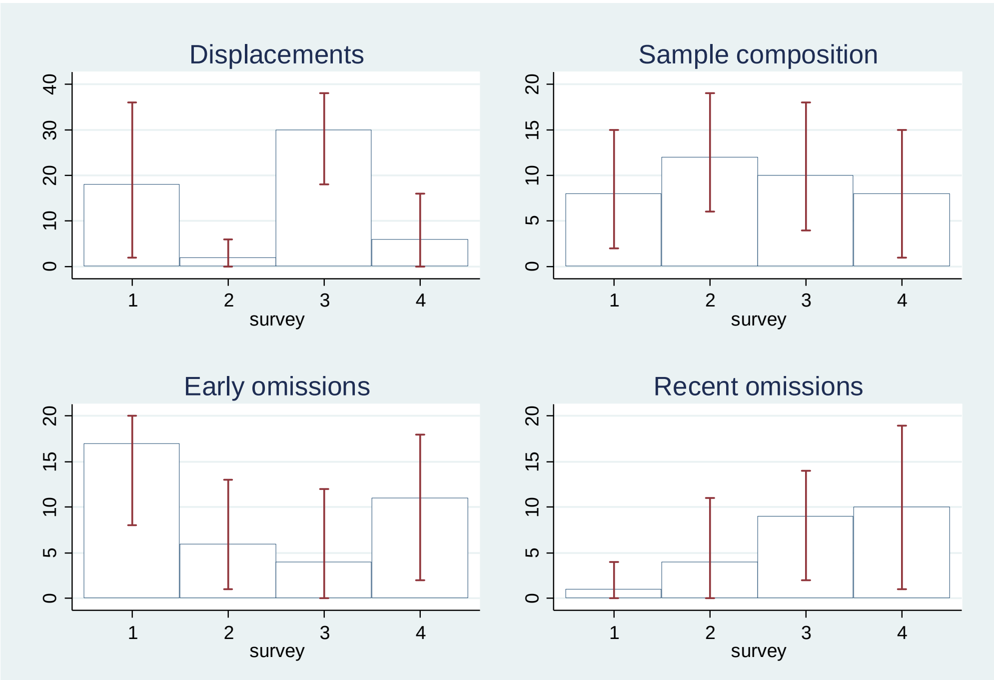
Graphs represent median, 10th percentile and 90th percentile.



Graphs represent median, 10th percentile and 90th percentile.

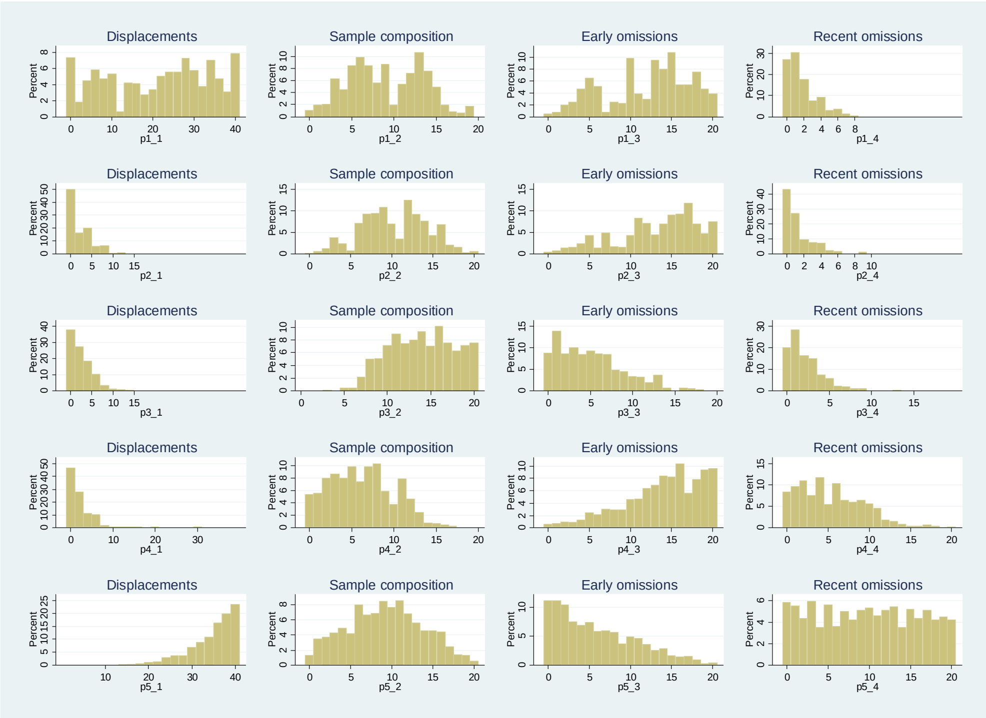
Cameroon

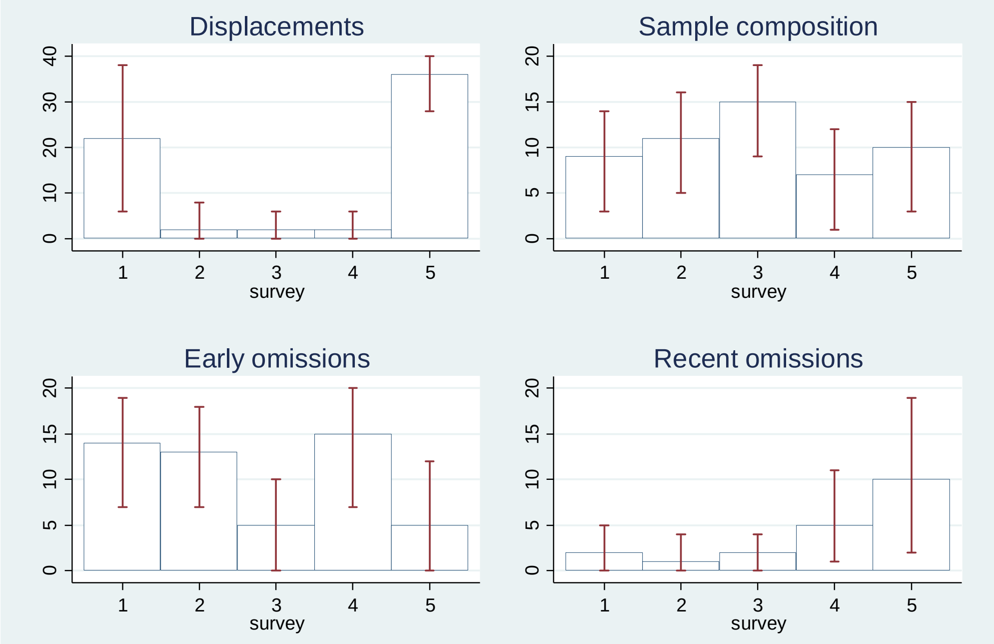




Graphs represent median, 10th percentile and 90th percentile.

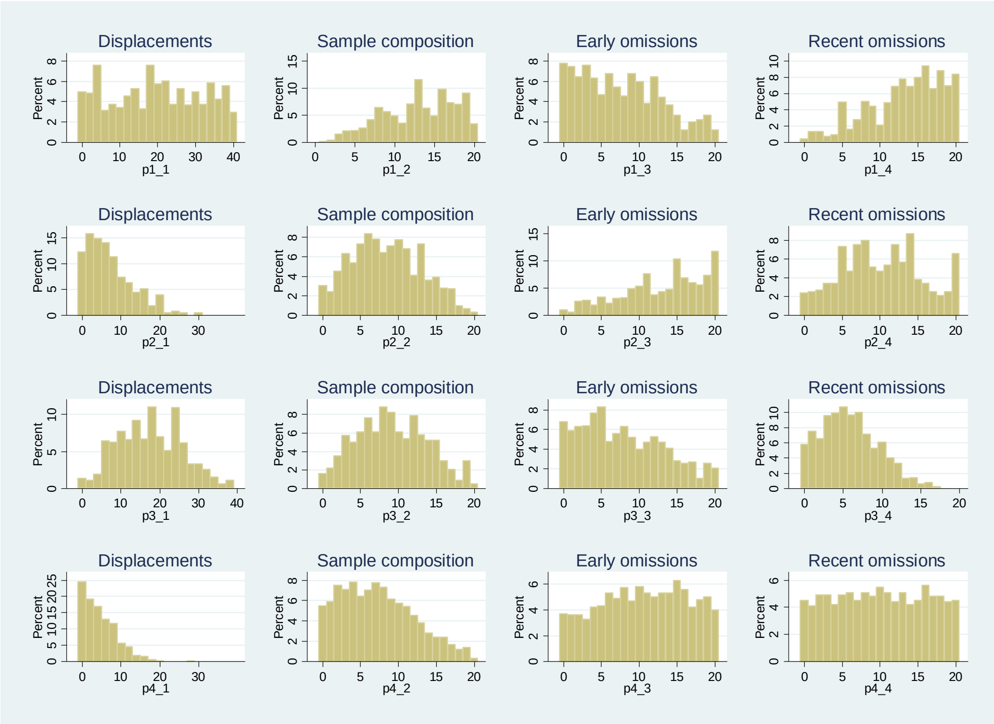
Ghana

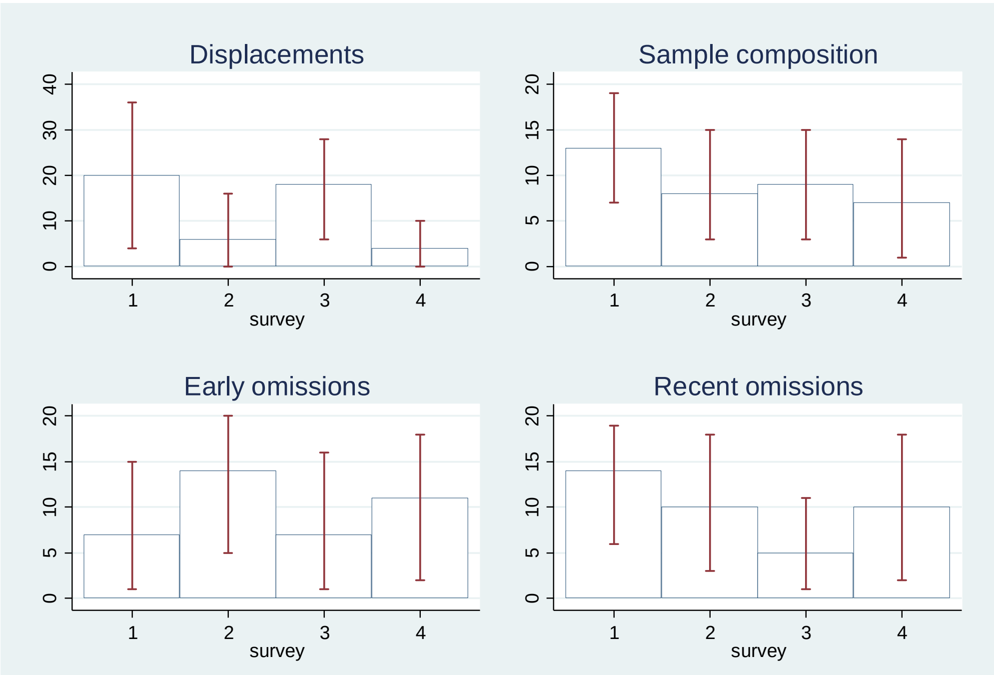




Graphs represent median, 10th percentile and 90th percentile.

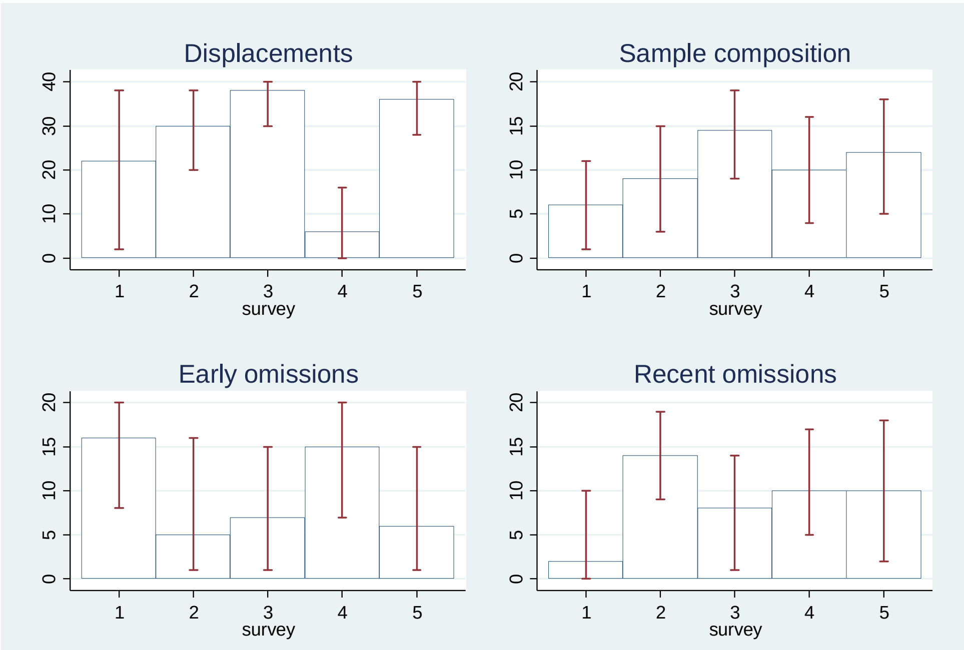
Haiti





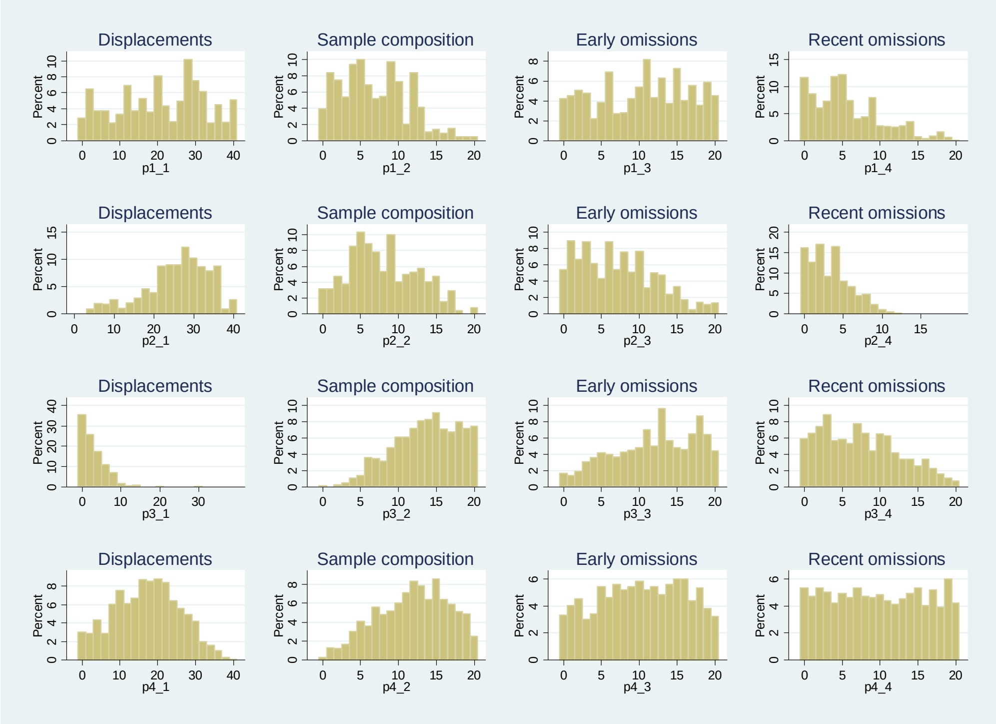
Graphs represent median, 10th percentile and 90th percentile.

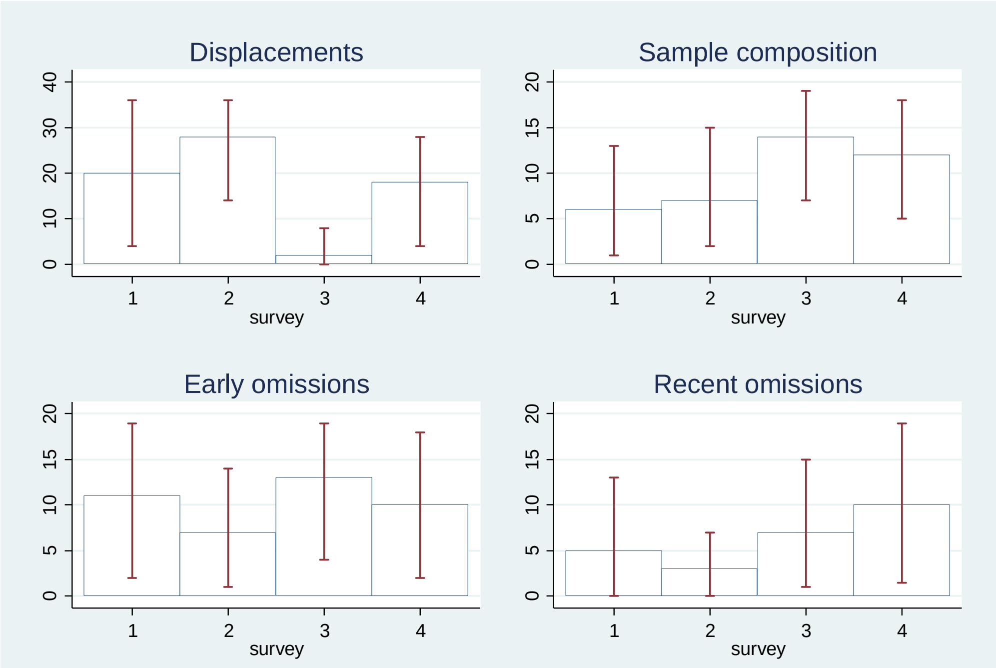
Kenya



Graphs represent median, 10th percentile and 90th percentile.

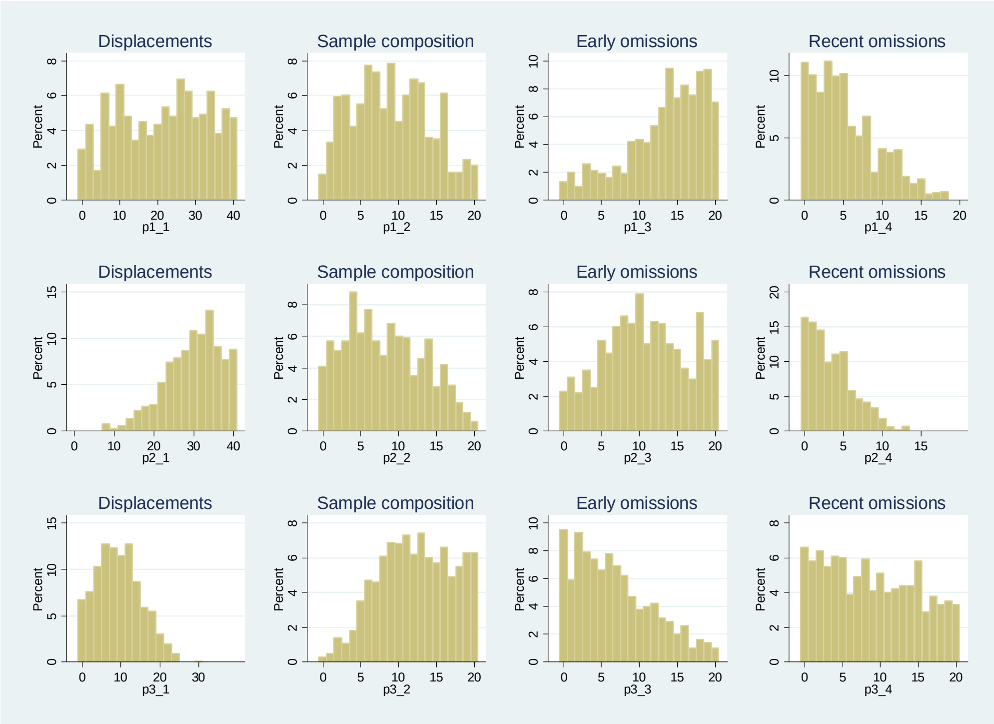
Madagascar

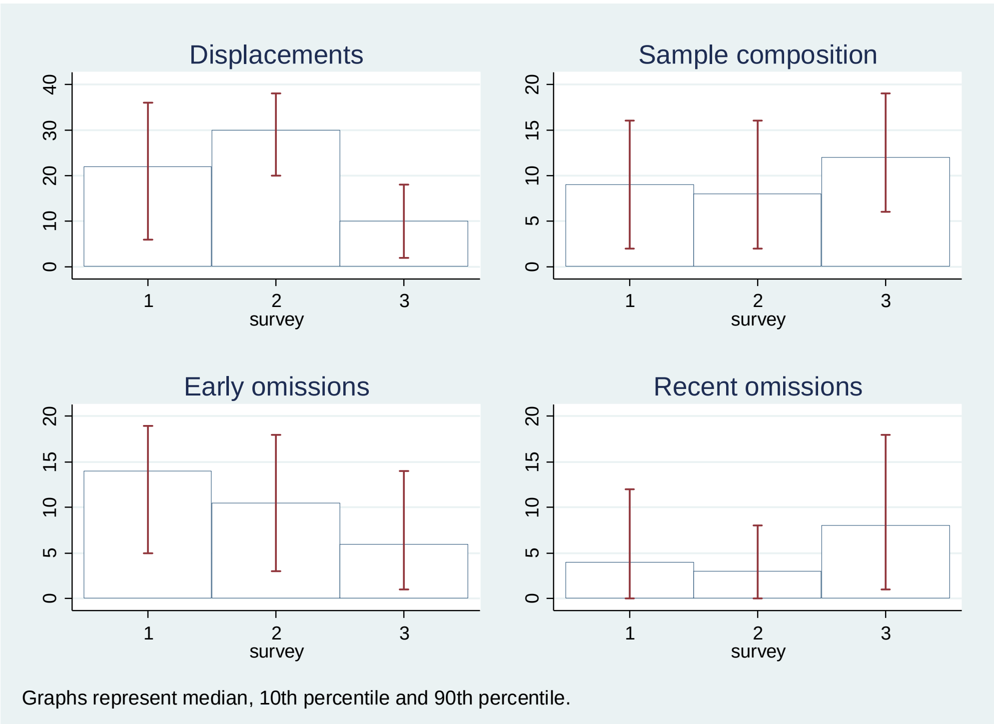




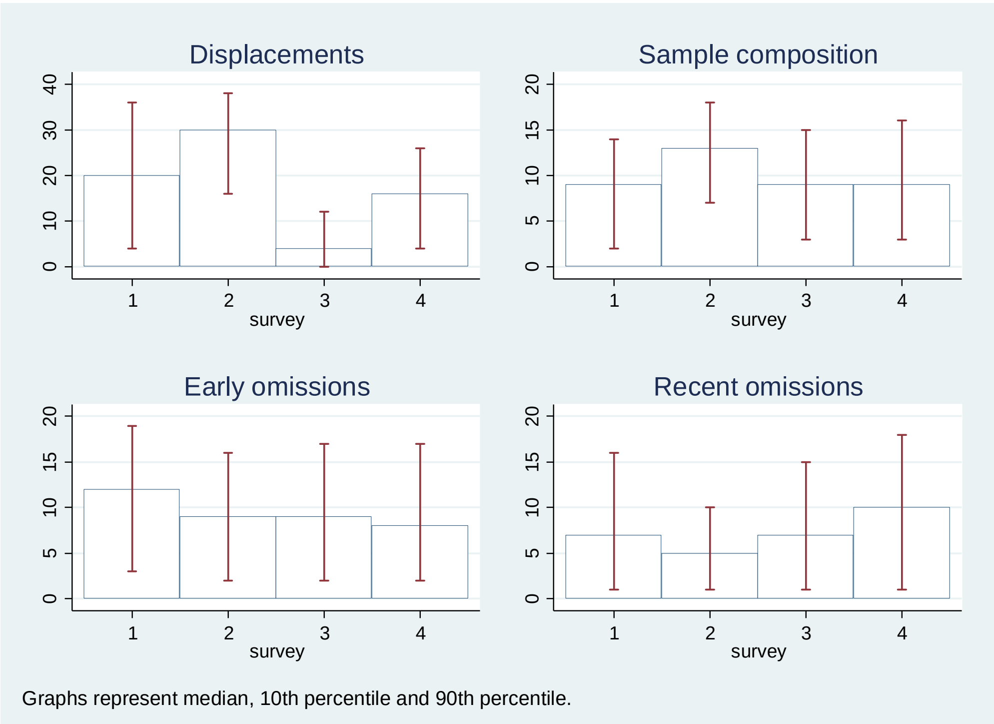
Graphs represent median, 10th percentile and 90th percentile.

Mozambique


Displacements
Sample composition
Early omissions
Recent omissions
Percent
p1_1
p1_2
p1_3
p1_4
p2_1
p2_2
p2_3
p2_4
p3_1
p3_2
p3_3
p3_4

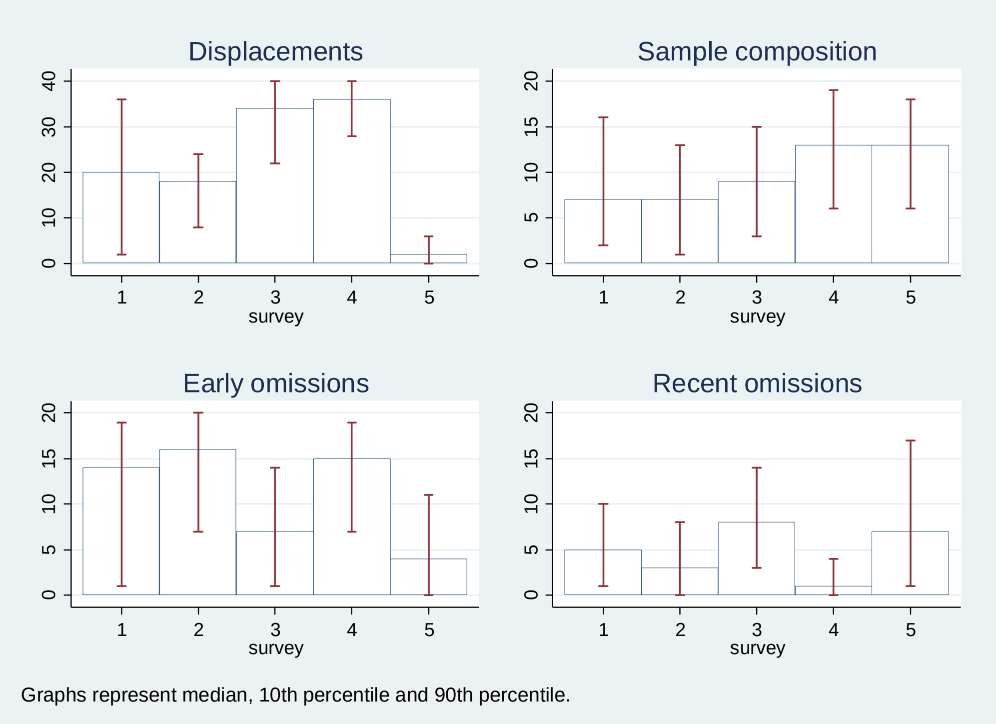

Displacements
Sample composition
Early omissions
Recent omissions
survey
Graphs represent median, 10th percentile and 90th percentile.

The Philippines



Rwanda


Displacements
Sample composition
Early omissions
Recent omissions
Percent
p1_1
p1_2
p1_3
p1_4
p2_1
p2_2
p2_3
p2_4
p3_1
p3_2
p3_3
p3_4
p4_1
p4_2
p4_3
p4_4
p5_1
p5_2
p5_3
p5_4


Displacements
Sample composition
Early omissions
Recent omissions
survey
Graphs represent median, 10th percentile and 90th percentile.