Max-Planck-Institut für demografische Forschung

Max Planck Institute for Demographic Research

Konrad-Zuse-Strasse 1 ▪ D-18057 Rostock ▪ Germany ▪ Tel +49 (0) 3 81 20 81 - 0 ▪ Fax +49 (0) 3 81 20 81 - 202 ▪ www.demogr.mpg.de

MPIDR Working Paper WP 2017-019 I November 2017

Estimating male fertility from vital registration data with missing values

Christian Dudel I dudel@demogr.mpg.de
Sebastian Klüsener I kluesener@demogr.mpg.de

© Copyright is held by the authors.

Working papers of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in working papers are attributable to the authors and do not necessarily reflect those of the Institute.

Estimating male fertility from vital registration data with missing values

Christian Dudel[*], Sebastian Klüsener

Abstract

Comparative perspectives on male fertility are still rare, in part because vital registration data often do not include paternal age information for a substantial number of births. We compare two imputation approaches that attempt to estimate male age-specific fertility rates and related measures for data in which the paternal age information is missing for a non-negligible number of cases. Taking births with paternal age information as a reference, the first approach uses the unconditional paternal age distribution, while the second approach considers the paternal age distribution conditional on the maternal age. To assess the performance of these two methods, we conduct simulations that mimic vital registration data for Sweden, the U.S., Spain, and Estonia. In these simulations, we vary the overall proportion and the age selectivity of missing values. We find that the conditional approach outperforms the unconditional approach in the majority of simulations, and should therefore generally be preferred.

Keywords: male fertility; birth register data; imputation methods; Sweden; United States; Spain; Estonia;

[*] Corresponding author; address: Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1, 18057 Rostock, Germany; email: dudel@demogr.mpg.de; phone: +49 381 2081221; fax: +49 381 2081521

1 Introduction

Much of the fertility literature has focused on female reproduction, while research on male reproduction has remained rare (Coleman 2000; Poston et al. 2006). However, interest in male fertility has been growing in response to the increasing involvement of men in fertility decision-making and parenting (Lappegård et al. 2011), and to concerns that have been raised about possible links between paternal ages and the health outcomes of children (Khandwala et al. 2017, Paavilainen et al. 2016). The number of studies that analyze male fertility has therefore been growing in recent years (e.g., Schoumaker 2017; Dudel & Klüsener 2016; Nordfalk et al. 2015; Nisén et al. 2014; Carmichael 2013; Lappegård et al. 2009).

One important reason why comparative perspectives on male fertility are rare is that in vital registration data, paternal age information is often missing for a substantial proportion of births, which makes it difficult to estimate age-specific fertility rates (ASFRs) and summary measures such as the total fertility rate (TFR). According to the UN Department of Economic and Social Affairs (2015), the proportion of missing values in birth registers in European countries has, in recent years, ranged from a low value of around 1% or 2% in countries such as Slovenia and Sweden, to a value of nearly 30% in Lithuania. In Latin America, even higher values have been recorded: e.g., 35% in Uruguay and 71% in Ecuador.

When estimating male fertility for countries in which the paternal age information is lacking for a substantial number of births, it is very common to impute the missing ages based on the paternal age distribution among births for which this information is available. This approach is used by the UN Department of Economic and Social Affairs (UN DESA 2015), among others. For example, if for 8% of births with complete paternal age information the father was aged 30, then it is assumed that this pattern also applies to births with missing paternal age data. This approach ignores that the paternal age information is not "missing completely at random" (Heitjan & Basu 1996). There is strong evidence for this, which relies on a variable that is virtually always available for registered births: the age of the mother. In many datasets, the maternal age at birth is highly correlated with both the probability of a missing value and the age of the father. More specifically, the paternal age is less likely to be recorded if the birth is to a young mother, while for most cases of a birth to a young mother for which the paternal age information is available the father is also relatively young (see section 2.3). It can therefore be assumed that fathers whose ages are not recorded are likely to be younger than fathers whose ages are known.

In this paper, we compare two nonparametric imputation approaches that are used to deal with missing paternal age data: the frequently applied "unconditional" approach, which is described above; and the "conditional" approach. The conditional imputation approach replaces the unconditional paternal age distribution with the paternal age distribution conditional on the maternal age (Dudel and Klüsener 2016; Carmichael 2013). It thus makes use of more information than the first approach, and explicitly assumes that maternal and paternal ages at birth are highly correlated. If, for example, we consider births to mothers aged 20 for which the paternal age information is not available, we impute ages for these births by taking the observed distribution or paternal ages for mothers aged 20 into account. If 1% of all births to mothers aged 20 are to fathers aged 30, then 1% of the births to mothers aged 20 for which the paternal age is unknown are assigned the paternal age 30.

To compare the performance of the unconditional and conditional imputation approaches, we conduct simulations which allow us to assess the bias in fertility estimates due to the proportion and age selectivity of missing values. These simulations mimic empirical birth register data for Sweden (1968-2014), the U.S. (1969-2015), Spain (1975-2014), and Estonia (1989-2013); and thus represent realistic settings. The countries and years cover very different demographic conditions, ranging from relatively early and high fertility to late and lowest low fertility. This variation in conditions allows us to explore the strengths and weaknesses of both approaches. The (female) TFR of Sweden over the last several decades has been analogized to a rollercoaster, with a succession of increases and decreases (Hoem & Hoem 1996; Andersson & Kolk 2016). In the U.S., the TFR was around two for most of our observation period, after an initial decline. The Spanish TFR plummeted from a rather high value in the late 1970s to a lowest-low fertility level (Kohler et al. 2002), and has been consistently below 1.5 since the 1990s. Estonias TFR dropped quickly after the collapse of the Soviet Union, in part due to the postponement of births to higher ages. All of the countries in our study saw marked increases in the mean age at childbirth for women, albeit from rather different baseline levels and with varying intensities. The proportion of births for which the paternal age was missing fluctuated considerably between countries and years. For instance, the proportion was around 1% for most years in Sweden, but was as high as 17% for the early 1990s in the U.S. For overviews of female fertility and the proportion of missing values in the countries studied, see the appendix.

The remainder of this paper is structured as follows. In section 2, we explain the unconditional and the conditional approaches. In section 3, we describe the data we use and the simulation setup. We present the results in section 4, followed by concluding remarks in section 5.

2 Imputation methods: Formal definitions and an illustrative example

2.1 Notation

We are interested in calculating age-specific fertility rates (ASFRs) for a given country and year t. $B(x,t)$ denotes the count of births to men aged x; i.e., in the age interval $[x, x+1)$, during the time interval $[t, t+1)$; $N(x,t)$ refers to the population exposure of men aged x. ASFRs are calculated as $f(x,t) = B(x,t)/N(x,t)$. To simplify the notation, we drop t and write, e.g., $f(x)$.

To avoid any issues with the denominator of fertility rates, we assume that $N(x)$ is always known. The number of births for which the paternal age is missing is denoted with $B(*)$. Since the paternal age is unknown for some births, it may be assumed that the data on the age-specific birth counts $B(x)$ are not complete. Instead, only $B^*(x)$ is observed, which captures the births for which the paternal age is known to be x. Given these missing values, $B^*(x)$ will be lower than $B(x)$.

2.2 The unconditional approach

The unconditional approach is based on the assumption that the paternal age at birth is "issing completely at random"; which means that the probability of a missing value does not depend on the age of the father or on other variables in the birth register. Births with missing values $B(*)$ are then distributed according to the observed paternal age distribution (see, e.g., UN DESA 2015). Written formally, let $P^*(x)$ denote the proportion of births to fathers aged x, calculated by ignoring missing values; i.e., $P^*(x) = B^*(x)/\sum_{i=\alpha}^{\beta} B^*(i)$, where α and β are, respectively, the first age and the last age of the reproductive phase of males. The unconditional approach then calculates ASFRs as:

$$f(x) = \frac{B^*(x) + B(*)P^*(x)}{N(x)}$$

2.3 The conditional approach

The conditional approach assumes that in many cases, paternal age information is not missing completely at random. For instance, for Sweden from 1968 to 2014, the Spearman's rank correlation between the proportion of missing values and the age of the mother is -0.4. Similar negative correlations can also be obtained for the other three countries. As we stated above, the conditional approach exploits the correlation between the maternal age at birth and both the paternal age at birth and the probability of a missing value. It does so by using the paternal age distribution conditional on the age of the mother (Dudel and Klüsener 2016; Carmichael 2013). This is based on the assumption that the age of the father is "missing at random" conditional on the age of the mother. More formally, let $B^*(x,y)$ denote the number of births to fathers aged x and mothers aged y. We assume that the maternal age is virtually always available, which based on the data documentation of the Human Fertility Database (2017) seems to be a quite realistic assumption. $P^*(x|y)$ is the paternal age distribution conditional on the maternal age, calculated as $B^*(x,y)/\sum_{i=\alpha}^{\beta} B^*(i,y)$; thus, the missing values are ignored. $B(*,y)$ represents the number of births for which the maternal age is known and equals y, but the paternal age is unknown. ASFRs are then calculated as:

$$f(x) = \frac{B^*(x) + \sum_{j=\gamma}^{\delta} B(*,j)P^*(x|j)}{N(x)}$$

where γ and δ denote the youngest and the oldest childbearing ages of women.

3 Simulation study setup

3.1 Data

To achieve realistic results, our simulation study emulates real data based on empirical birth counts, age distributions, and population exposures. To derive the birth counts and age distributions of mothers and fathers, we used birth register data from Sweden (1968-2014; supplied by Statistics Sweden), the U.S. (1969-2015; available through the National Bureau of Economic Research), Spain (1975-2014; provided by the Spanish Statistical Office), and Estonia (1989-2013; supplied by Statistics Estonia). For the U.S., the data for some years are based on 50% random samples of the birth registers of the individual U.S. states. For these years, we used weights to estimate the number of births by maternal and paternal ages. The population exposures for all of the countries were taken from the Human Mortality Database (2017).

3.2 Simulation setup

To assess the relative performance of both methods, we simulate data that mimic important characteristics of the obtained birth register data. These characteristics include observed paternal age distributions conditional on the maternal age at birth in each country and year. The simulation setup is briefly presented here, while further details and discussions are provided in the appendix. In our simulated datasets, we know the paternal ages for all births, but assume that they are unobserved for some births. We consider a wide variety of simulations in which we change two parameters: (1) the overall proportion of missing values; and (2) a parameter that alters the paternal age distribution for those births for which the paternal age is missing. This second parameter, called "age shift," explores how the performance of our methods changes if the unobserved paternal age distribution of births with missing values differs from the observed distribution. For instance, if the age shift parameter equals -1 years, this means that the average paternal age for births with missing values is one year less than for births for which the paternal age is known. If this age shift parameter is different from zero, then the missing values are "not missing at random," which implies that the assumptions of both the unconditional and the conditional approach are violated.

We refer to data for a specific country-year combination as a simulation setting, of which we have 159 in total. For each simulation setting, we apply 450 different combinations of simulation parameters, resulting in a total of $71,550$ simulations. For the proportion of missing values, we use values of 1% to 50% with 1% increments; while for the age shift, we consider values from -4 years to +4 years with one-year increments. Age shift values below -2 and above +2 can be considered rather extreme (see the appendix for a discussion).

For each of our 71,550 simulations, we simulate two datasets: one representing the "observed" data with missing values, and the other representing the "true" underlying data without missing information. We start with all births of a given country and year, and preliminary distribute them according to $P^*(x|y)$, with the paternal age distribution conditional on the maternal age. The latter has been calculated for the same country and year based on the empirical births for which the paternal age is known. Then we make use of the first parameter, the proportion of missing values, to set the age of the father to missing for some births. This share missing varies by the age of the mother, for which we again mimic characteristics in the specific country-year simulation setting. For the births for which the age of the father is "observed", we keep the empirical age distribution $P^*(x|y)$. The resulting dataset consists of the "observed" simulation data. To introduce age selectivity among the births for which the paternal age is "unobserved", we shift for each of these births the originally simulated paternal age depending on the age shift parameter (see the appendix for a discussion of alternative strategies). The combined "observed" and (shifted) "unobserved" part of the data then become the "true" dataset that we can use as a benchmark to measure the performance of the two imputation approaches.

To assess the bias of each imputation method, we calculate the "true" male TFR (MTFR) and the paternal mean age at childbirth (PMAC) based on the "true" data, as well as the MTFR and PMAC based on the "observed" data where missing paternal ages have been imputed with one of the two approaches. We then calculate the absolute bias of both imputation approaches for the MTFR and the PMAC as $|\text{MTFR}_{true} - \text{MTFR}_{observed}|$ and $|\text{PMAC}_{true} - \text{MAC}_{observed}|$, respectively. Bias in the MTFR captures whether the level of fertility is estimated correctly, while bias in the PMAC allows us to assess whether the timing of fertility is estimated correctly.

4 Results

4.1 Illustrative example: Sweden 2014

We use the simulation results based on the most recent Swedish data for 2014 as an illustrative example that allows us to explain how the simulations and the imputation approaches work. In other years and countries, both higher and lower values of bias can be found (see section 4.2 and appendix C).

The simulation results for Sweden in 2014 are shown in Figures 1 to 4. Figure 1 displays the absolute bias of the MTFR estimate derived using the unconditional imputation approach, and the extent to which it depends on the proportion of missing values and the selectivity of missing values (age shift). Figure 2 shows the same information for the conditional imputation approach. Absolute bias never exceeds 0.012 MTFR points; i.e., the difference between the true MTFR and the MTFR based on imputations is never higher than 0.012 for either approach, and is thus small. If the proportion of missing values is below 10%, bias is negligible for both approaches, irrespective of the age shift. In the case of the unconditional method, the results are biased to some extent above this threshold, even if there is no selectivity at all among the missing values.

In contrast, the conditional approach performs considerably better and the degree of bias is minimal in this simulation setting, even with high age shift values. This is because the conditional approach correctly accounts for the higher probability that missing values will occur among younger mothers who, on average, have younger partners. If the fathers are a few years older or younger than is estimated by the conditional approach – as indicated by the age shift parameter – these discrepancies have little effect on the degree of bias in the MTFR estimate. The overall performance of the conditional approach is considerably better than that of the unconditional approach, as in 97% of our simulations, the former approach is less biased than the latter approach.

Figures 3 and 4 show the absolute bias for the estimates of the PMAC. The overall pattern is quite different from the pattern found for the MTFR, as the conditional approach outperforms the unconditional approach in only 56% of the simulations. This is because the MTFR is sensitive to inaccuracies in the paternal age imputation only if successive cohorts vary substantially in size, which is rarely the case. For the PMAC, by contrast,

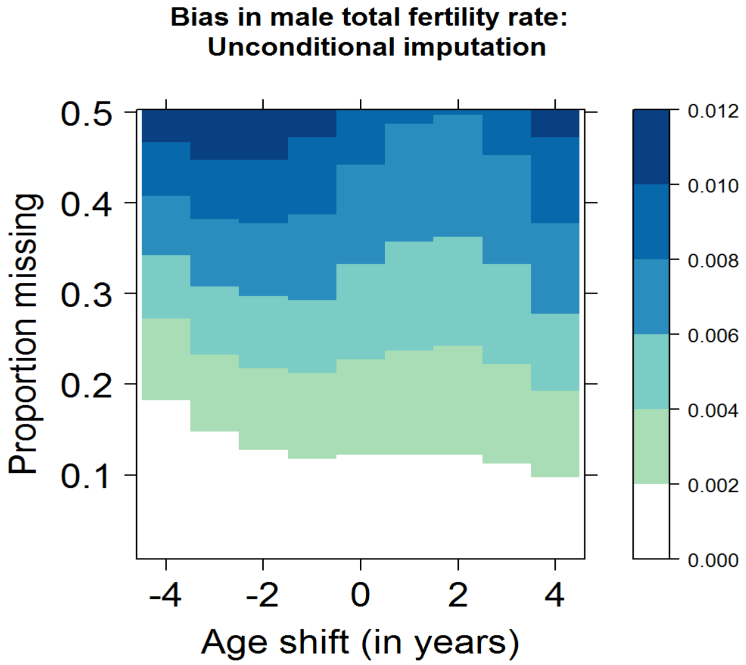

Figure 1: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2014. Source: Own calculations.

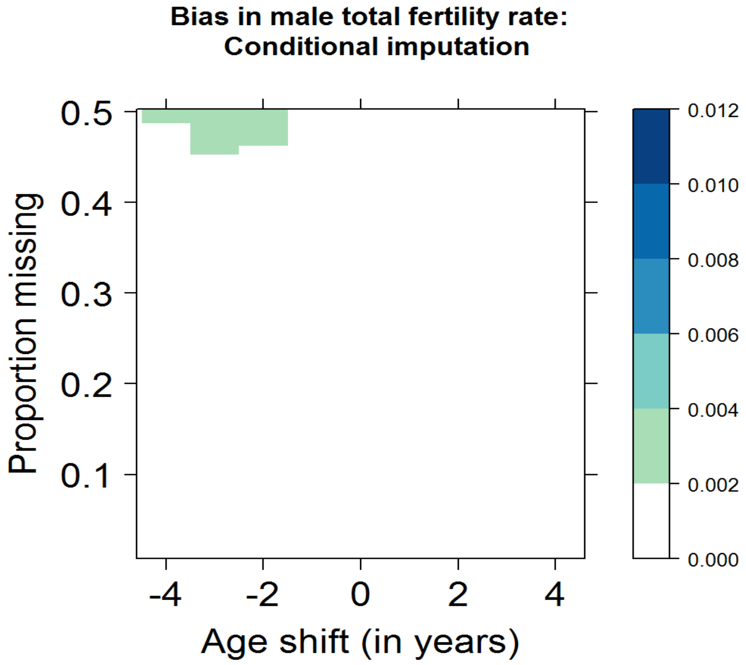

Figure 2: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 2014. Source: Own calculations.

the paternal age information is the key determinant of the calculations. The level of bias in the PMAC can be considerable for both methods, amounting to a difference of 2.2 years between the "observed" and "true" values for the unconditional approach, and to a difference of 1.9 years for the conditional approach. For the conditional approach, the level of bias seems to be more balanced around an age shift of zero, and is high only for more extreme values of the age shift parameter. Meanwhile, for the unconditional approach the bias can be more than one year for an age shift of two.

4.2 Simulation results for Sweden, the U.S., Spain, and Estonia

For each of our 159 country-year combinations, we generated results like those presented in Figure 1 to 4. A more detailed breakdown of these results is available in the appendix, while an overview is given in Table 1. For each country, the proportion of simulations in which the conditional approach outperforms the unconditional approach is shown for both the MTFR and the PMAC. In columns 1 and 4 we provide the outcomes for all of the parameter combinations considered. In addition, we present the results for more limited parameter spaces in which we exclude extreme cases of age selectivity of missing values, and focus only on the simulations for which the proportion of missing values is not negligible so that the choice of approach is more relevant. Based on the detailed outcomes presented in the appendix, we have chosen to focus on the simulations for which the paternal age information is simulated to be missing for at least 10% of the births. For these restricted summaries, we consider age shift parameters between -2 and +2 (columns 2 and 5), and between -1 and +1 (columns 3 and 6). The last row of Table 1 provides the average results for all of the 159 country-year simulation settings considered.

While the results differ somewhat by country, the conditional approach outperforms the unconditional approach in roughly two out of three cases if our entire space of parameter combinations is considered (columns 1 and 4). The outperformance of the conditional approach becomes even more pronounced if we consider only simulations in which at least 10% of the values are missing, and restrict the age shift to -2 and +2 years (columns 2 and 5), and then to -1 to +1 year (columns 3 and 6). Especially in the latter case, the conditional approach is highly dominant.

5 Summary and conclusion

We conducted simulations to compare two nonparametric imputation approaches that can be used to calculate age-specific fertility rates for males if paternal age information is missing for some births: the unconditional approach and the conditional approach. The

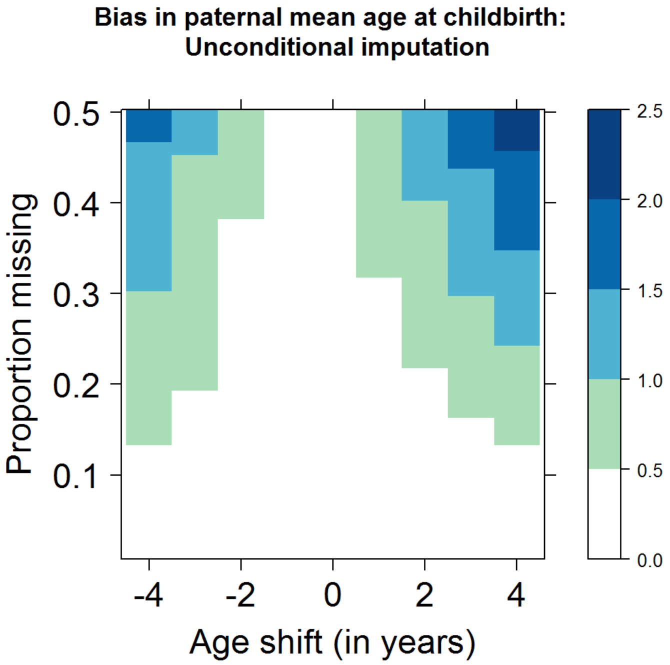

Figure 3: Absolute bias of the paternal mean age at childbirth (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2014. Source: Own calculations.

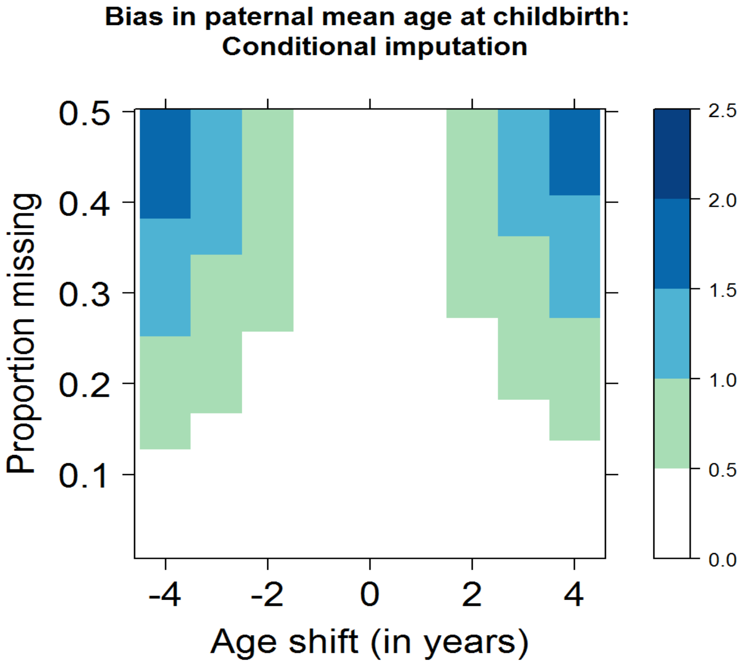

Figure 4: Absolute bias of the paternal mean age at childbirth (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 2014. Source: Own calculations.

Table 1: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the male total fertility rate (MTFR) and the paternal mean age at childbirth (PMAC); by country and total (in %).

		MTFR			PMAC	
	(1)	(2)	(3)	(4)	(5)	(6)
Age shift:	-4/+4	-2/+2	-1/+1	-4/+4	-2/+2	-1/+1
% Missing	$\geq 0\%$	$\geq 10\%$	$\geq 10\%$	$\geq 0\%$	$\geq 10\%$	$\geq 10\%$
Sweden (1968-2014)	76	83	91	56	62	69
USA (1969-2015)	75	92	97	73	94	100
Spain (1975-2014)	72	83	95	66	78	97
Estonia (1989-2013)	55	66	78	59	67	78
Average	71	83	92	64	76	86

Notes: The first and fourth columns take age shift values between -4 and +4 into account. The second and fifth columns are restricted to simulations with age shift values between -2 and +2, and also consider only simulations in which the paternal age is missing for at least 10% of births. The third and the sixth columns cover age shift values between -1 and +1 and a proportion of missing values of at least 10%.

Source: Own calculations.

unconditional approach, which is often adopted in the literature, can perform poorly if the proportion of missing values is non-negligible, and can even be biased if there is no selectivity of missing values. The conditional approach, on the other hand, performs better in the majority of simulations. Both approaches work well if the proportion of missing values is small. Generally, the conditional approach should be preferred.

While our study covers several different countries and periods, one limitation is that we only mimicked data from high- and middle-income countries. It is thus unclear whether our findings are transferable to low-income countries – and especially to low-income countries where very different patterns of male fertility have been reported (polygamy, etc.). In addition, in these countries reliable register data are often unavailable, and the data quality can be low. Schoumaker (2017) has discussed how male fertility can be estimated in such settings.

This paper focused on the maternal and paternal ages at birth only, but additional attributes can be included in both the unconditional and the conditional approach. For instance, if for every birth the legal status of the mother is available and has been shown to be related to the probability that the paternal age is missing, the data can be split into marital and non-marital births, and the imputation can proceed separately for each part. In such a case, however, the additional covariates must cover virtually all births with no or only a few missing values. In addition, the presented imputation approaches might also be used to impute other missing information on the father (such as education or other socioeconomic status attributes), provided this information is available for most of the mothers.

Acknowledgments

We thank Pavel Grigoriev, conference participants in Turku (Nordic Demographic Symposium), and conference participants in Cape Town (International Population Conference) for very helpful suggestions and comments on earlier versions of the manuscript. We also thank Miriam Hils for language editing.

References

Andersson, G., Kolk, M. (2016). Trends in childbearing, marriage and divorce in Sweden: An update with data up to 2012. *Finnish Yearbook of Population Research* 50: 21-30.

Carmichael, G. A. (2013): Estimating Paternity in Australia, 1976-2010. *Fathering* 11: 256-279.

Coleman, D. (2000): Male fertility trends in industrial countries: Theories in search for some evidence. In: Bledsoe, C., Lerner, S., Guyer, J. (eds.). Fertility and the male life cycle in the era of fertility decline. Oxford: Oxford University Press: 29-60.

Devolder, D., Ortiz, E., Zeman, K. (2016): Human Fertility Database Documentation: Spain. Available online at `http://www.humanfertility.org/Docs/ESP/ESPcom.pdf`.

Dudel, C., Klüsener, S. (2016): Estimating male fertility in eastern and western Germany since 1991: A new lowest low? *Demographic Research* 35: 1549-1560.

Gustafson, P., Fransson, U. (2015): Age differences between spouses: Sociodemographic variation and selection. *Marriage & Family Review* 51: 610-632.

Heitjan, D.F., Basu, S. (1996): Distinguishing missing at random and missing completely at random. *The American Statistician* 50: 207–213.

Hoem, B., Hoem, J. M. (1996): Swedens family policies and roller-coaster fertility. *Journal of Population Problems* 52: 1-22.

Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available online at `www.humanfertility.org`.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available online at `www.mortality.org` or `www.humanmortality.de`.

Khandwala, Y. S., Zhang, C. A., Lu, Y., Eisenberg, M. L. (2017): The age of fathers in the USA is rising: an analysis of 168 867 480 births from 1972 to 2015. *Human Reproduction* 32: 2110-2116.

Kohler, H. P., Billari, F. C., Ortega, J. A. (2002): The emergence of lowestlow fertility in Europe during the 1990s. *Population and Development Review* 28: 641-680.

Lappegård, T., Rønsen, M., Skrede, K. (2009): Fatherhood and fertility. *Fathering* 9: 103-120.

Nisén, J., Martikainen, P., Silventoinen, K., Myrskylä, M. (2014): Age-specific fertility by educational level in the Finnish male cohort born 1940–1950. *Demographic Research* 31: 119-136.

Nordfalk, F., Hvidtfeldt, U.A., Keiding, N. (2015): TFR for males in Denmark: Calculation and tempo correction. *Demographic Research* 32: 1421-1434.

Schoumaker, B. (2017): Measuring male fertility rates in developing countries with Demographic and Health Surveys: An assessment of three methods. *Demographic Research* 36: 803–850.

Paavilainen, M., Bloigu, A., Hemminki, E., Gissler, M., Klemetti, R. (2016): Aging fatherhood in Finland – first-time fathers in Finland from 1987 to 2009. *Scandinavian Journal of Public Health* 44: 423-430.

Poston, D.L., Baumle, A.K., Micklin, M. (2006): Epilogue: Needed research in demography. In: Poston, D.L., Micklin, M. (eds.). Handbook of population. New York: Springer: 853–881.

UN Department of Economic and Social Affairs (2014): Demographic Yearbook 2014. New York: United Nations.

A Implementation of the simulations

For each country and year, two parameters are used to alter our simulated birth register data: the proportion of missing values, and the degree to which the age distribution of the simulated "unobserved" paternal age distribution differs from the paternal age distribution among the "observed" cases. These two parameters can be applied once a basic dataset has been created. As we stated in the main text, we are simulating basic datasets by taking all births of a given country and year, and distributing them on a preliminary basis according to $P^*(x|y)$, with the paternal age distribution conditional on the maternal age distribution. The latter has been calculated for the same country and year based on the empirical births for which the paternal age is known.

The first parameter (proportion missing) is then applied by scaling the conditional proportion of missing values by the age of the mother up or down; for births for which the paternal age is known, $P^*(y|x)$ is used. This procedure creates the simulated "observed" dataset with missing values. In Figure 5, the pattern found in the empirical data for Sweden in 2014 is shown as a solid line. The dashed lines correspond to overall proportions of missing values of 10% and 20%.

Generally, a desired proportion of missing values $P_{des}(*)$ is achieved by iterative upscaling: If the proportion of missing values for a country-year setting is $P_{obs}(*)$, then a scaling factor $\lambda = P_{des}(*)/P_{obs}(*)$ is calculated and applied to the observed conditional proportion of missing values, resulting in a new set of adjusted values $P_{adj}(*|y) = P_{obs}(*|y)\lambda$. One challenge we face is that in cases in which there are no observed paternal ages at birth, neither the conditional nor the unconditional approach can be applied. For the conditional approach this is even true if for one or more maternal ages no births with paternal age information are available. We thus implement a ceiling so that if we derive values of $P_{adj}(*|y)$ larger than 90%, the corresponding proportions are set to 90%. In order to still reach the desired share of births with missing paternal age information, we use an iterative procedure in which we start anew by calculating and applying λ, potentially setting some values to 90%; and again starting anew until $P_{des}(*) \approx P_{obs}(*)$. This procedure guarantees that first, no conditional proportion of missing values exceeds 90% if the proportion is not already in the data. Moreover, in most cases, the upscaling procedure leaves the ratios of missing values unchanged; i.e., $P_{adj}(*|y)/P_{adj}(*|y') = P_{obs}(*|y)/P_{obs}(*|y')$. For instance, if the proportion of missing values for 20-year-old women is twice as high as it is for 30-year old women, the ratios will not be changed (except in certain cases because of the 90% rule described above).

An example of the effect of the second parameter (age shift) can be seen in Figure 6; the "true" but unobserved dataset is created using this simulation parameter. The solid line shows the distribution of the paternal age conditional on the mother being aged 30

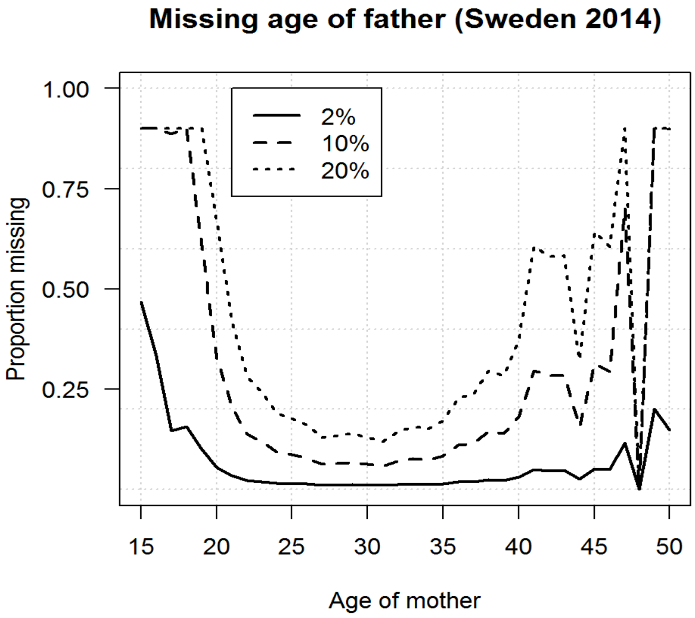

Figure 5: Proportion of missing values for the age of the father conditional on the age of the mother, Sweden 2014. Original data (solid line) and upscaled distributions (dotted and dashed lines); for young ages, the 90% threshold was reached and the proportion of missing values was not further increased. Source: Statistics Sweden; own calculations.

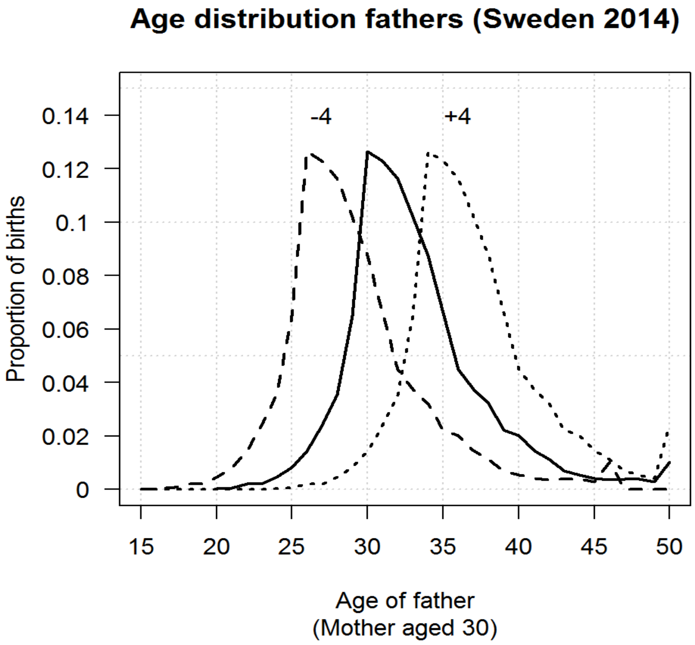

Figure 6: Distribution of the age of the father conditional on the mother being aged 30, Sweden 2014; original distribution (solid line) and shifted distributions (dotted and dashed lines). Source: Statistics Sweden; own calculations.

for the original data in 2014; i.e., $P^*(x|y=30)$. As was noted above, this is also the age distribution found in the simulated "observed" data, irrespective of the overall proportion missing. The dashed line was derived by shifting the distribution of births for which we set the paternal age to missing four years to the left. This makes the fathers of births with "unobserved" paternal age information younger; i.e., by applying an "age shift" of -4 years. The dotted line shows another scenario in which the fathers for whom the paternal age is not known are made older by shifting the distribution +4 years to the right. Formally, this shifting process can be described as follows: Given an age shift of ϕ, the distribution of the paternal ages of births for which this information is assumed to be unobserved, $P_{un}(x)$, is defined as $P_{un}(x+\phi|y) = P_{obs}(x|y)$. If this shift by ϕ years leads to ages above β, the ages are set to β and summed up; i.e., β functions as a ceiling. A similar procedure is applied with the bottom age α. This shifting approach was applied to all conditional age distributions for the age shift values from 4 to +4 in one-year increments. A value of zero implies that missing values are "missing at random," while higher or lower values indicate different degrees of age selective missingness. An age shift of -4 or +4 can be seen as rather extreme (also see Dudel & Klüsener 2016); the three distributions shown in Figure 1b, for instance, imply an average age at childbirth of roughly 33.1 years (original data), 29.1 years (age shift of -4), and 37.1 years (age shift of +4). To put these results in perspective, it should be noted that the difference between the observed mean age and the ages implied by the shifted distributions is slightly larger than the change in the mean age at childbirth for Swedish women from 1968 to 2014.

The age shift parameter could also be implemented with distributions of shapes other than the one observed. For instance, it would be possible to combine the two shifted distributions in Figure 6 (both the dashed and dotted lines) to a bimodal mixture distribution and to use it as $P_{un}(x)$; or a more skewed distribution could be chosen. However, experimenting with the shape of the distributions showed that the main driver of bias is the average age and the extent to which it differs from the observed distribution; i.e., how large the age shift is. We therefore proceeded in the implementation of the age shift as described above.

B Additional simulation results

The main text focuses on the results for Sweden for 2014, and includes only a summary of the results for other years and countries. A more detailed overview is given in Figures 7 to 10, and in Figures 11 to 18 further below.

Figures 7 and 8 show the maximum absolute bias by year and country in simulations with age shift values between -2 and +2. Taken together with proportions of missing values between 1% and 50%, these age shift values imply a total of 495 simulations per country and year; e.g., there are 495 simulations for the U.S. in 2000. The maximum absolute bias in the PTFR of these 495 simulations is displayed for the unconditional approach in Figure 7, and for the conditional approach in Figure 8. When we compare the results in the two figures, we see that for the PTFR the conditional approach performs better for all years and countries; that is, that the worst-case bias for the conditional approach is always smaller than for the unconditional approach. Using the average absolute bias instead of the maximum bias generates similar patterns, albeit at lower levels (results available upon request).

Figures 9 and 10 show the maximum bias by year and country for the PMAC, also restricted to age shift values between -2 and +2. More specifically, Figure 9 displays the results for the unconditional approach, while Figure 10 presents the findings for the conditional approach. Again, the worst-case performance of the conditional approach is considerably better than that of the unconditional approach. Interestingly, for the conditional approach the maximum bias is roughly bounded by the proportion of missing values times the age shift for all country-year combinations. As the age shift is up to -/+2 and the proportion of missing values is up to 50%, the maximum bias in the PMAC is around one year.

In Figures 11 to 18, the proportion of simulations in which the conditional approach outperforms the unconditional approach is shown by parameter values, and separately for each country. For instance, Figure 11 shows for Sweden and for the MTFR that the conditional approach is better in 98% of the simulation settings (1968-2014) when the proportion of missing values is 20% and the age shift parameter is set to -1. Similar results for the PMAC are given in Figure 15.

The results indicate that overall, the age shift parameter is more important than the proportion of missing values in determining which approach works better. Moreover, for age shift values from -4 to +1 or +2 depending on the country, the conditional approach performs better than the unconditional approach in the majority of cases. For Estonia, the conditional approach performs slightly worse than it does for the other countries, but it still outperforms the unconditional approach if our whole parameter space is considered. The finding that the conditional approach is not much better than the unconditional approach

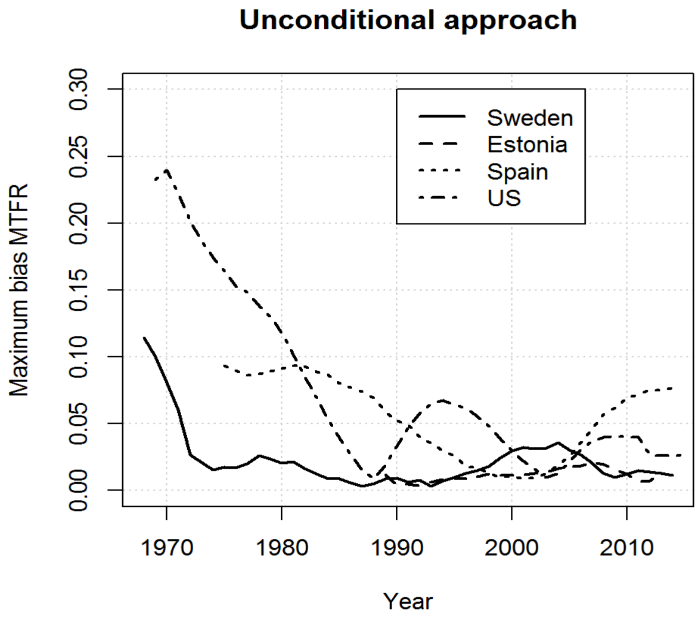

Figure 7: Maximum absolute bias in the male total fertility rate (MTFR) estimates in MTFR points by year and country, unconditional approach. Source: Own calculations.

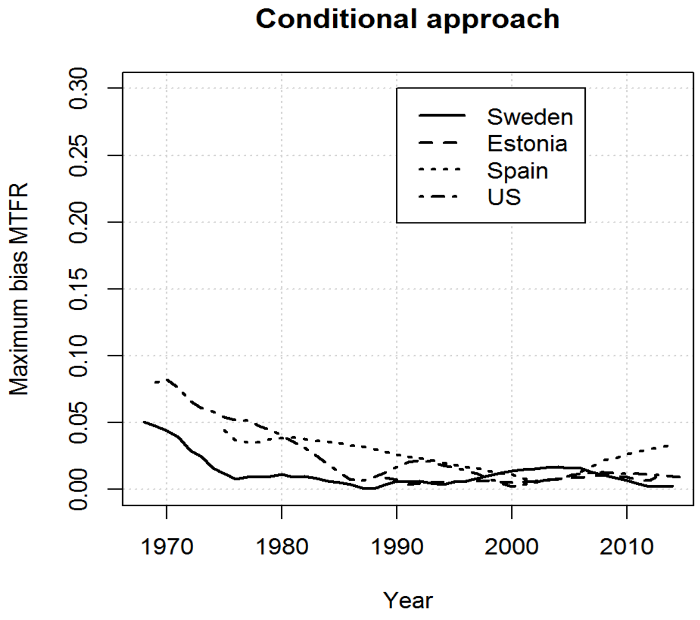

Figure 8: Maximum absolute bias in male total fertility rate (MTFR) estimates in MTFR points by year and country, conditional approach. Source: Own calculations.

can be partly attributed to the fact that for high positive age shift values, the conditional approach underestimates the paternal age, whereas the unconditional approach estimates the paternal age quite accurately due to its tendency to impute quite high paternal ages. For the other countries, the unconditional approach tends to overestimate the paternal age (detailed results are available upon request). Overall, only for high positive age shift values the unconditional approach is able to perform better than the conditional approach. When the age shift is around or below zero, on the other hand, the unconditional approach is almost always more biased than the conditional approach.

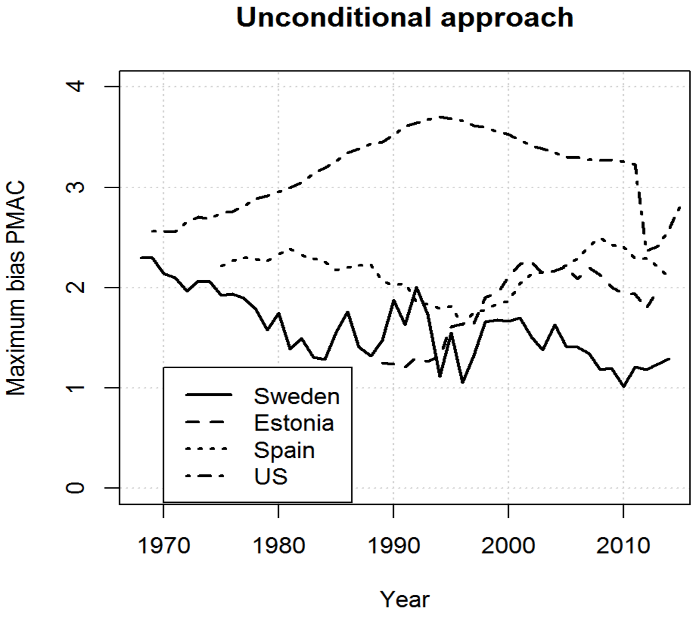

Figure 9: Maximum absolute bias in the paternal mean age at childbirth (PMAC) estimates in years by year and country, unconditional approach. Source: Own calculations.

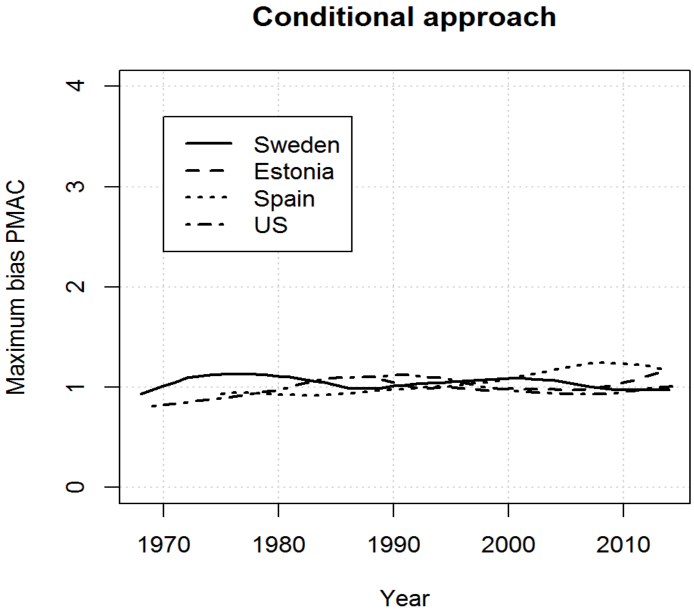

Figure 10: Maximum absolute bias in the paternal mean age at childbirth (PMAC) estimates in years by year and country, conditional approach. Source: Own calculations.

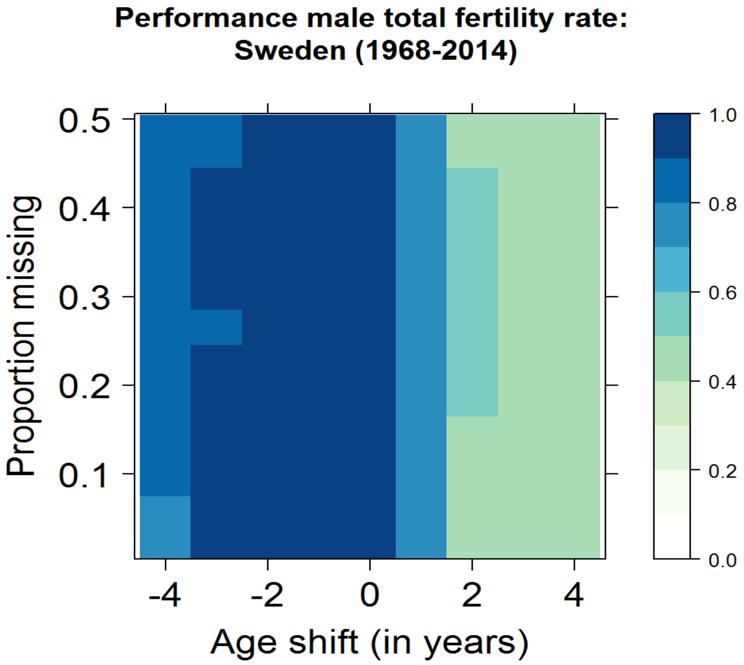

Figure 11: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the MTFR by simulation parameters, Sweden. Source: Own calculations.

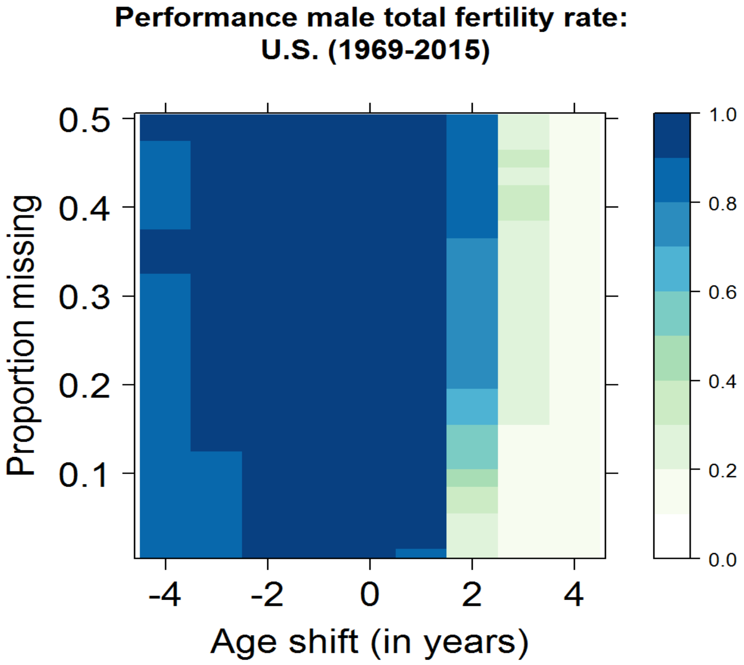

Figure 12: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the MTFR by simulation parameters, US. Source: Own calculations.

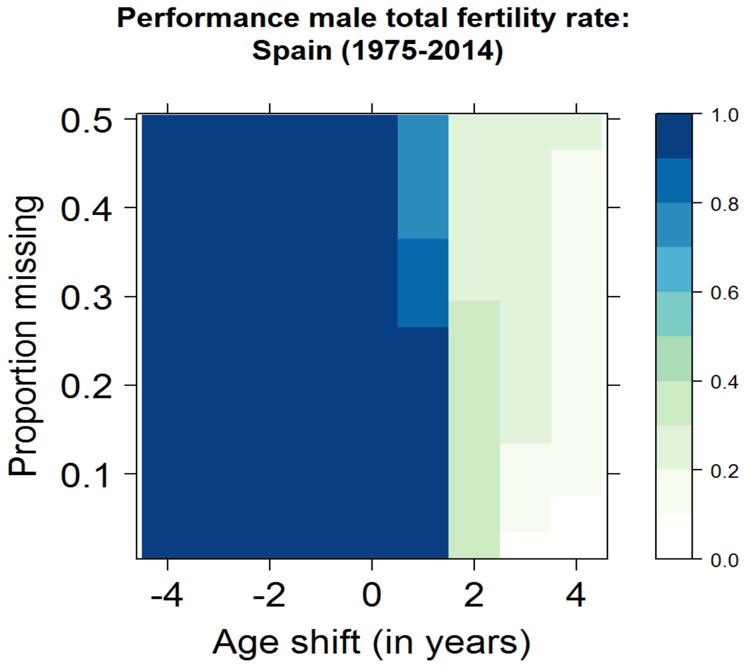

Figure 13: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the MTFR by simulation parameters, Spain. Source: Own calculations.

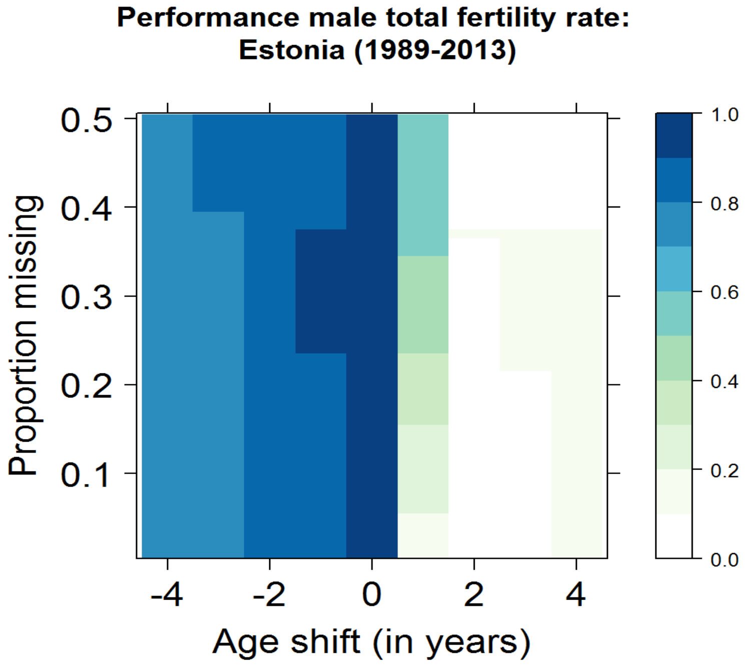

Figure 14: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the MTFR by simulation parameters, Estonia. Source: Own calculations.

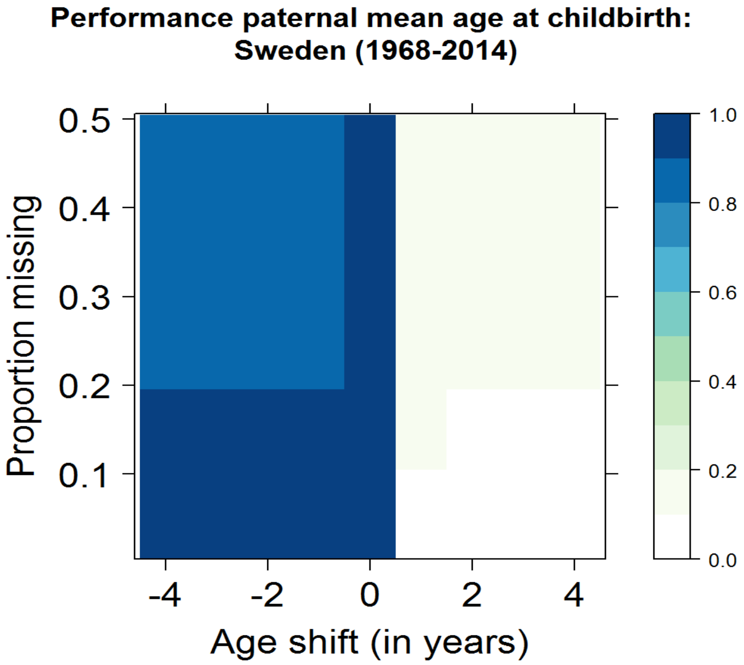

Figure 15: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the PMAC by simulation parameters, Sweden. Source: Own calculations.

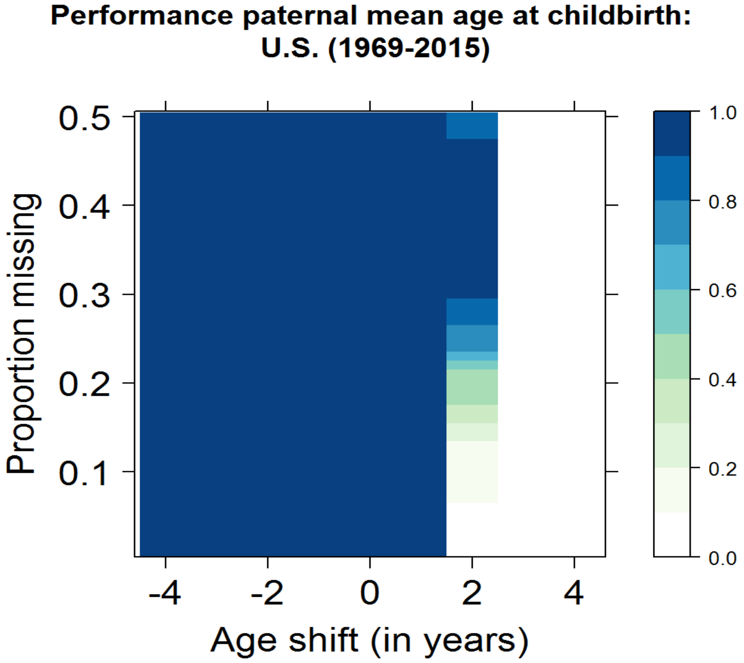

Figure 16: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the PMAC by simulation parameters, US. Source: Own calculations.

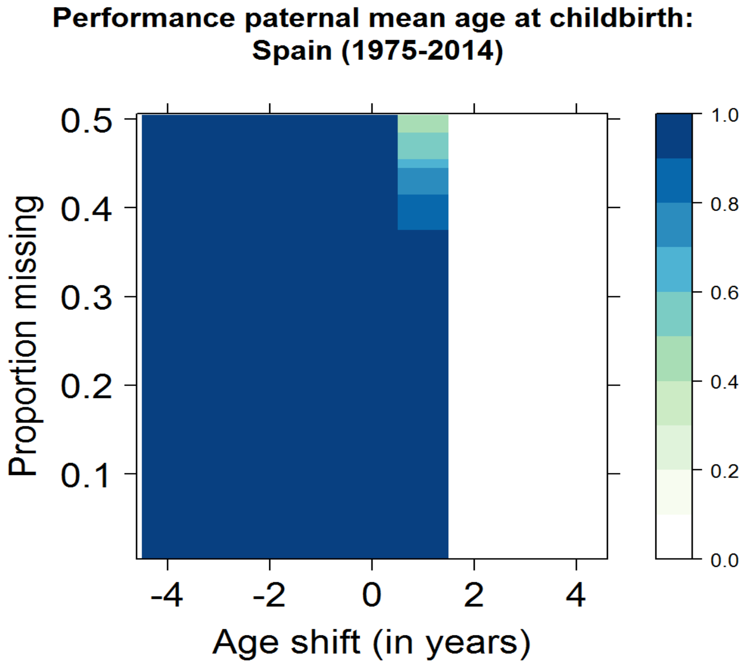

Figure 17: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the PMAC by simulation parameters, Spain. Source: Own calculations.

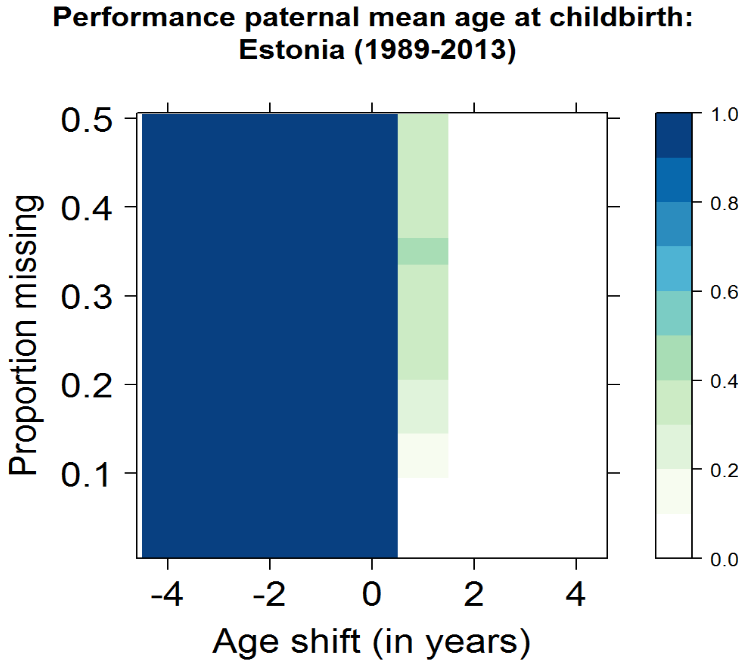

Figure 18: Proportion of simulations in which the conditional approach outperforms the unconditional approach for the PMAC by simulation parameters, Estonia. Source: Own calculations.

C Additional simulation results by year and country

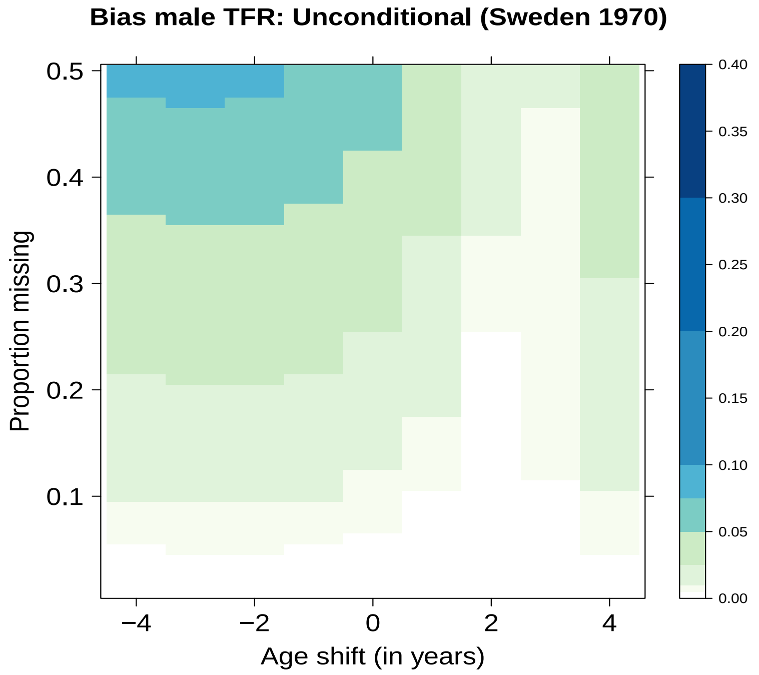

Figure 19: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1970. Source: Own calculations.

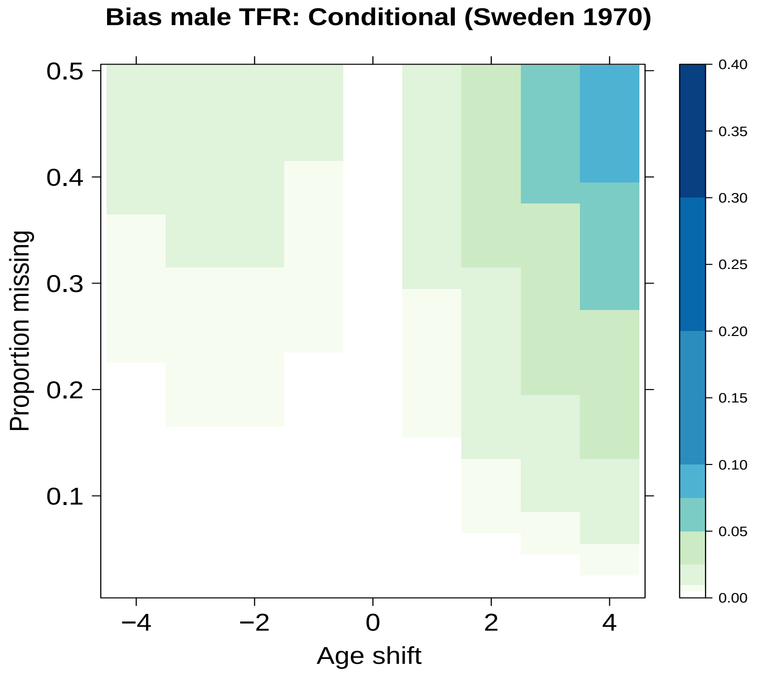

Figure 20: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 1970. Source: Own calculations.

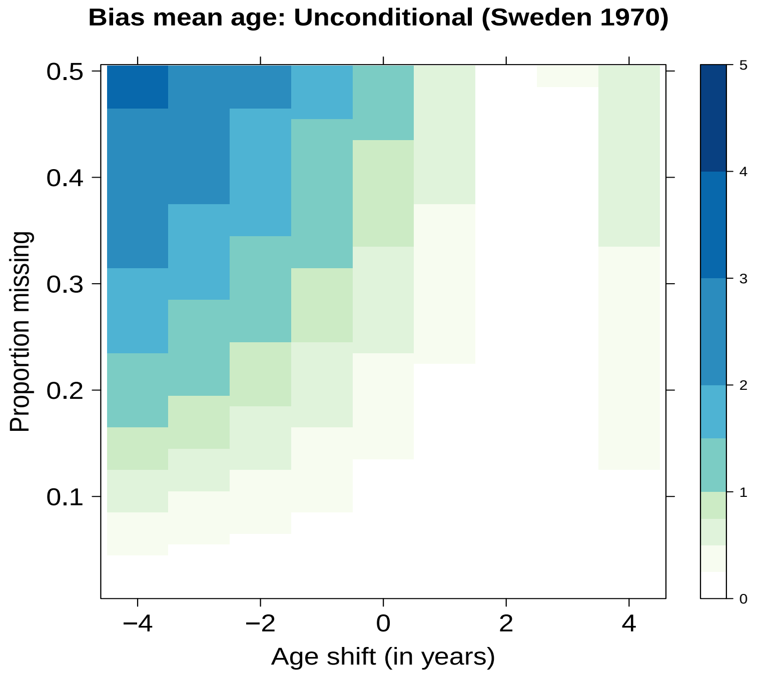

Figure 21: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1970. Source: Own calculations.

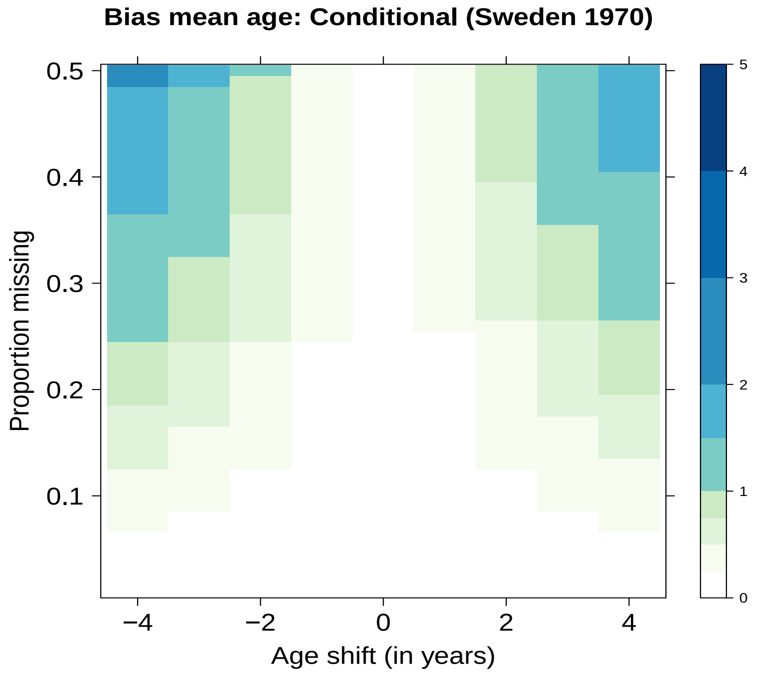

Figure 22: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 1970. Source: Own calculations.

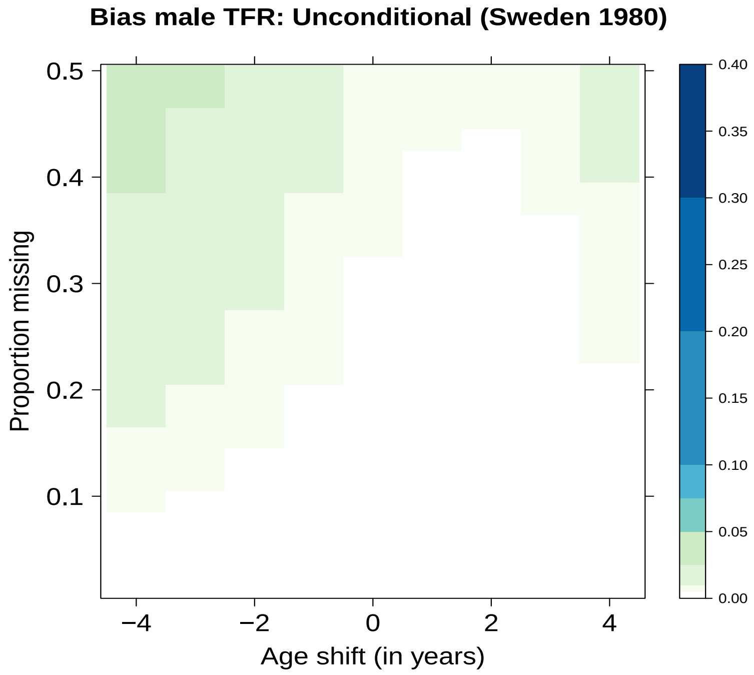

Figure 23: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1980. Source: Own calculations.

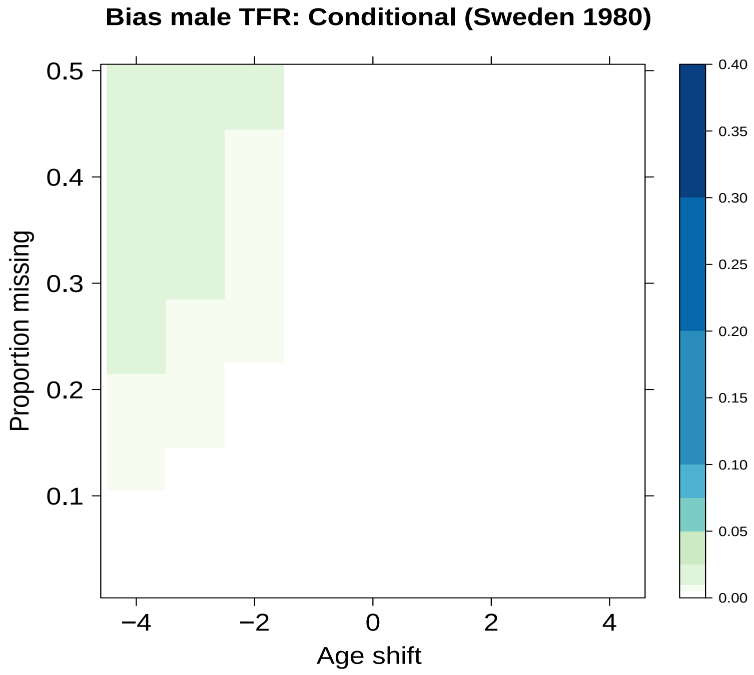

Figure 24: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 1980. Source: Own calculations.

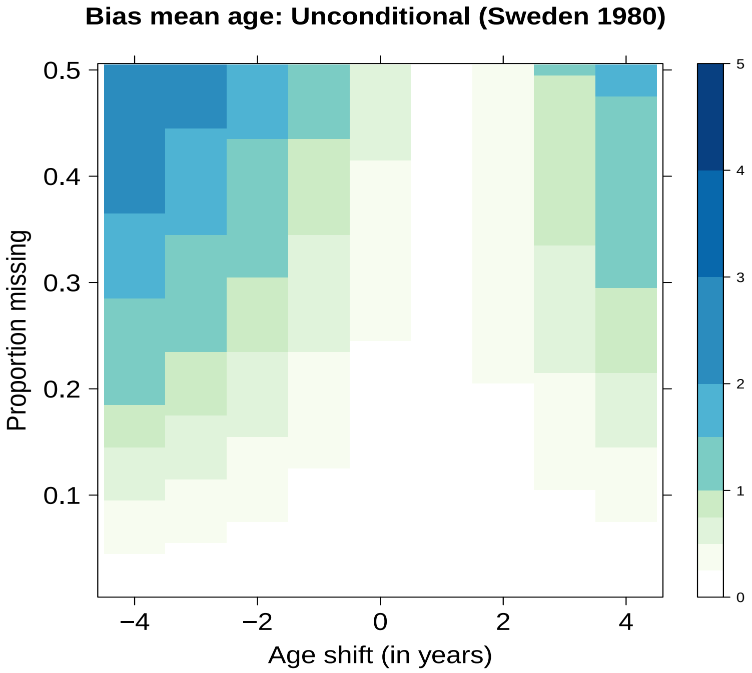

Figure 25: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1980. Source: Own calculations.

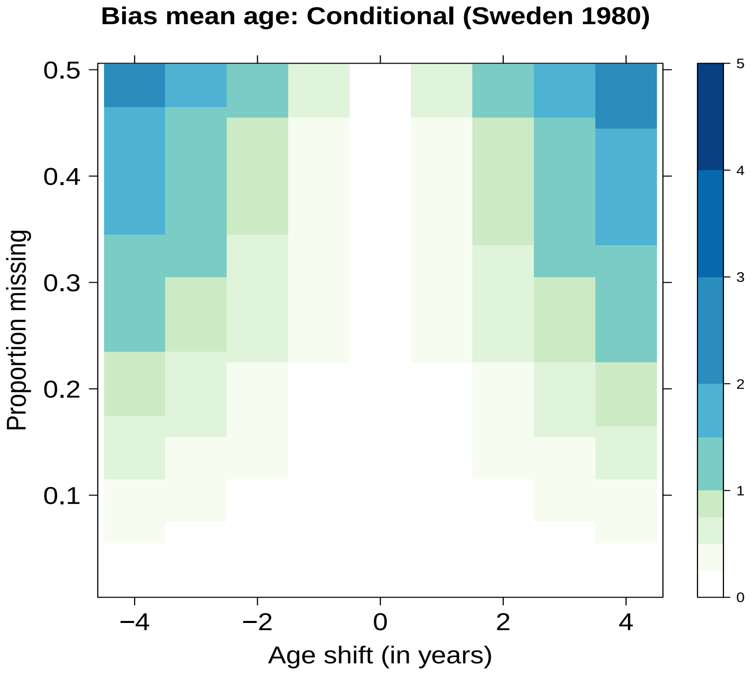

Figure 26: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 1980. Source: Own calculations.

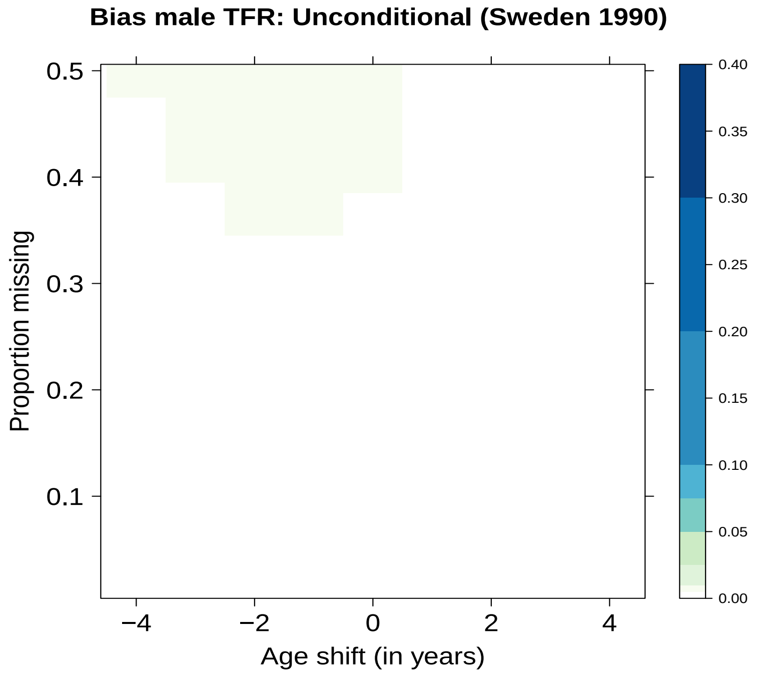

Figure 27: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1990. Source: Own calculations.

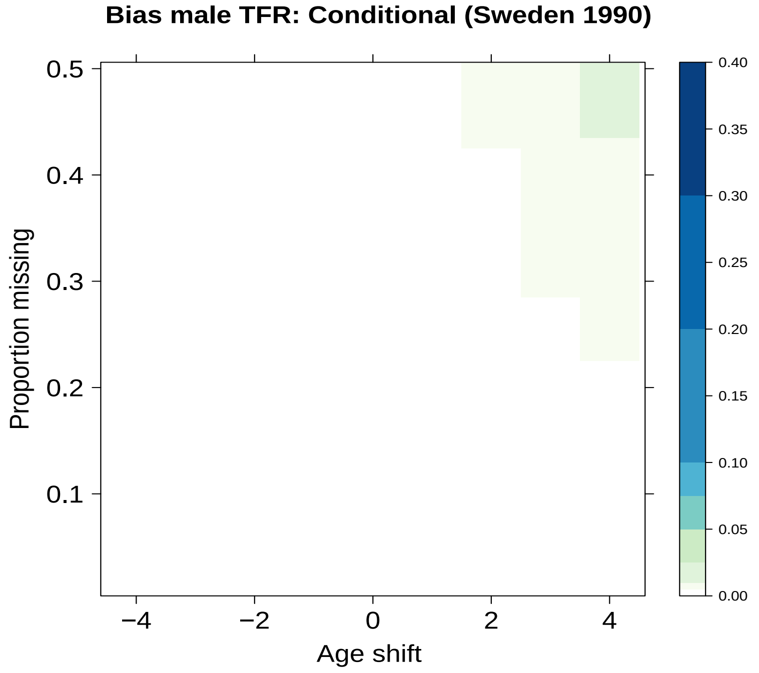

Figure 28: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 1990. Source: Own calculations.

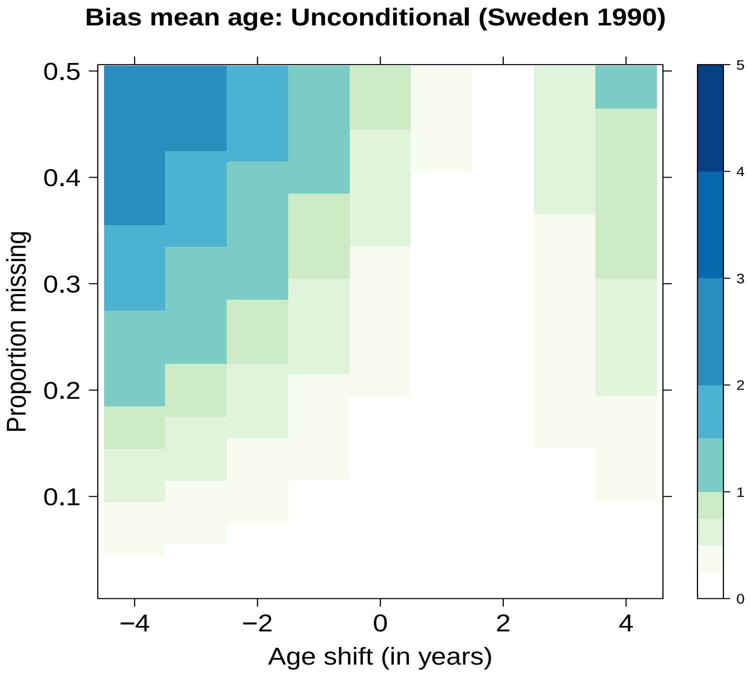

Figure 29: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 1990. Source: Own calculations.

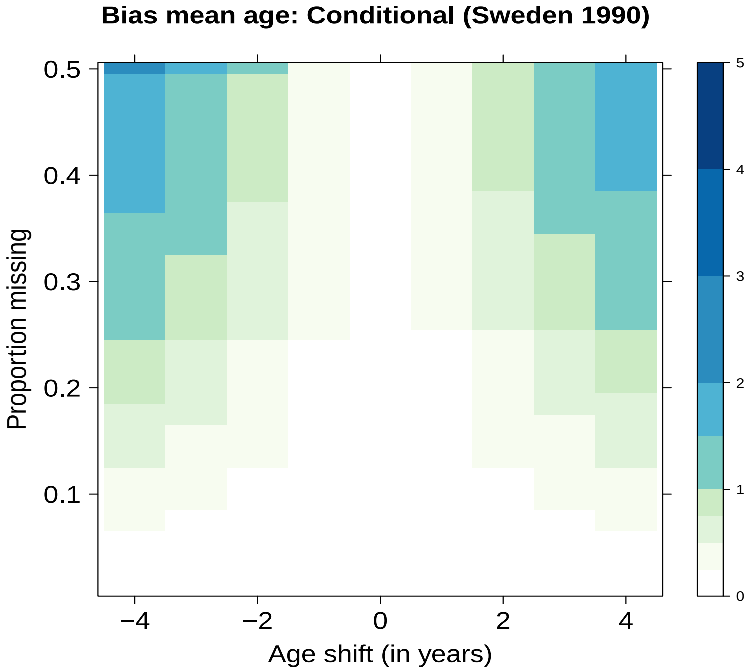

Figure 30: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 1990. Source: Own calculations.

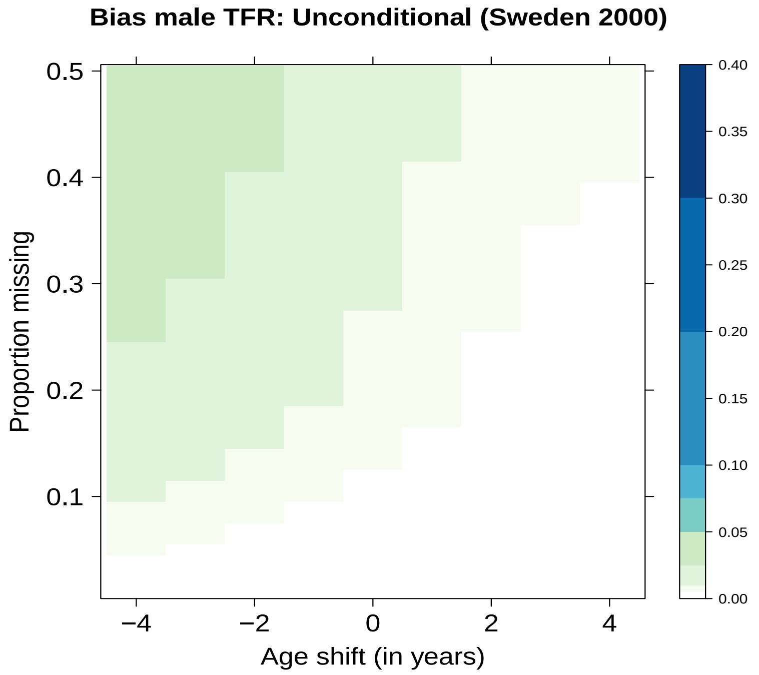

Figure 31: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2000. Source: Own calculations.

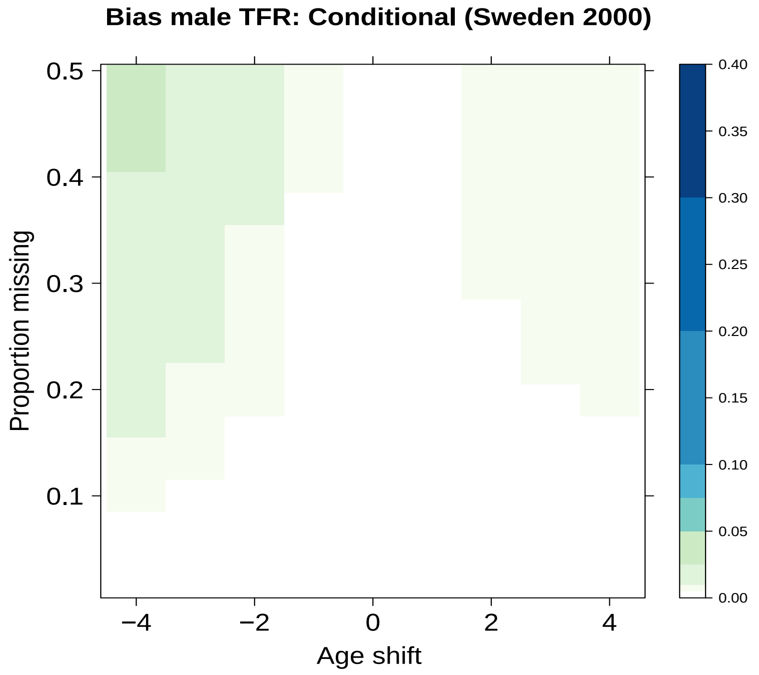

Figure 32: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 2000. Source: Own calculations.

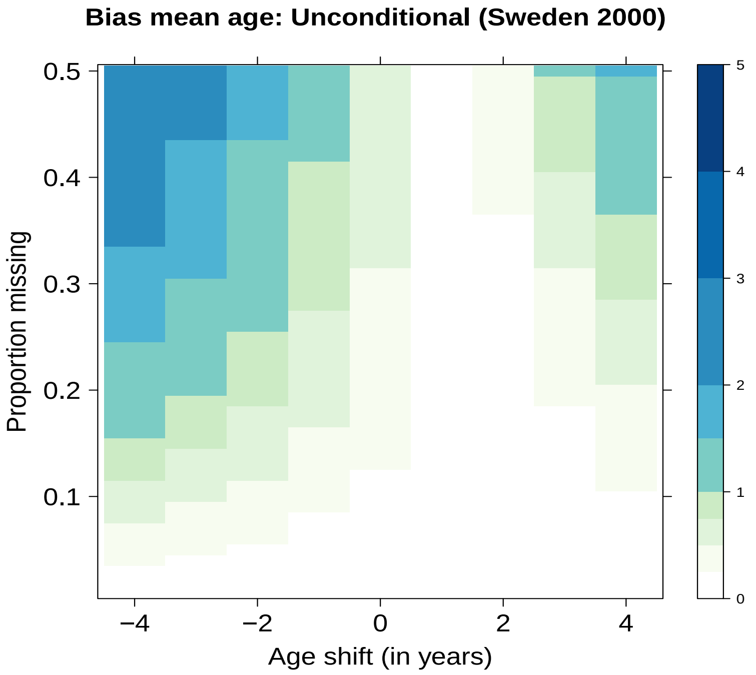

Figure 33: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2000. Source: Own calculations.

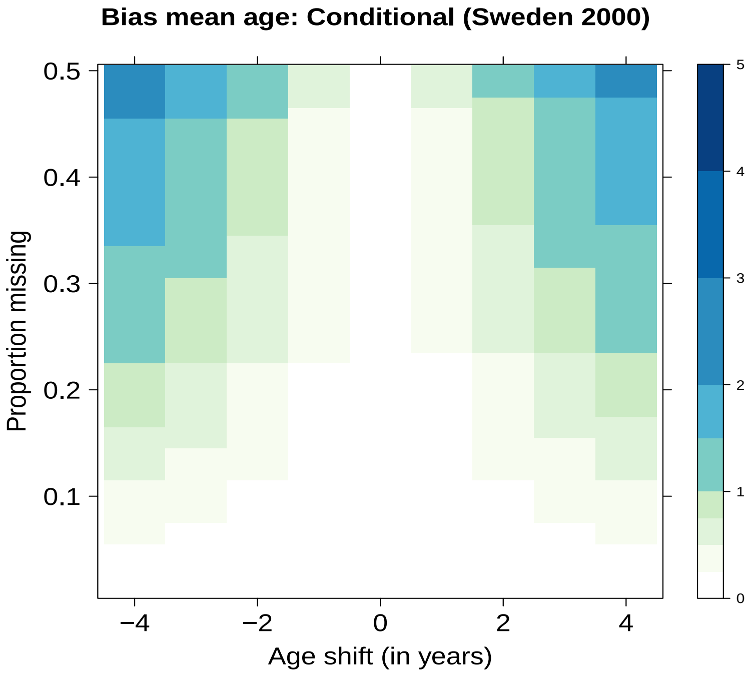

Figure 34: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 2000. Source: Own calculations.

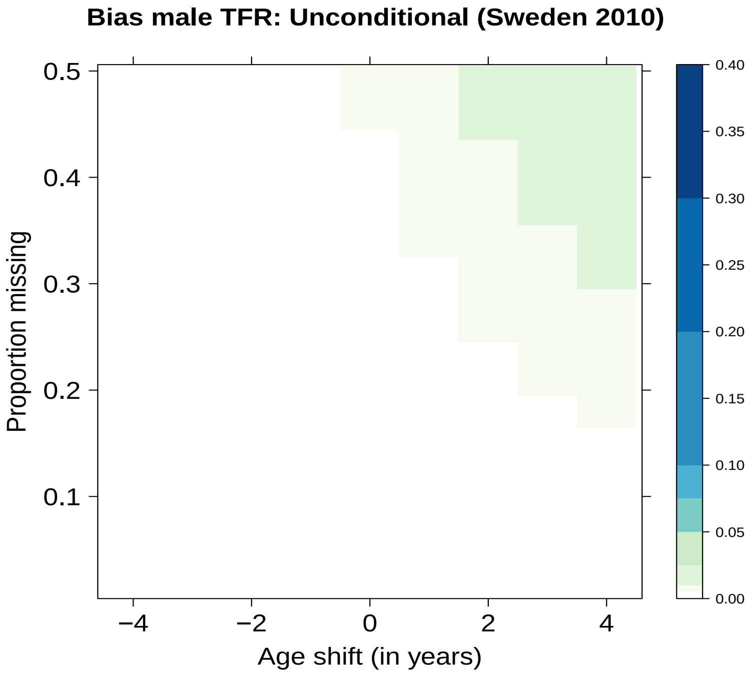

Figure 35: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2010. Source: Own calculations.

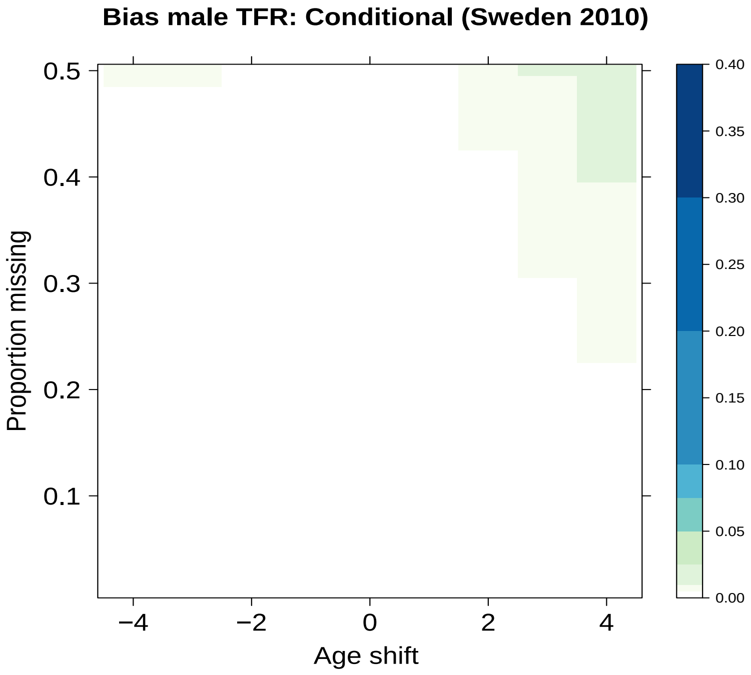

Figure 36: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Sweden 2010. Source: Own calculations.

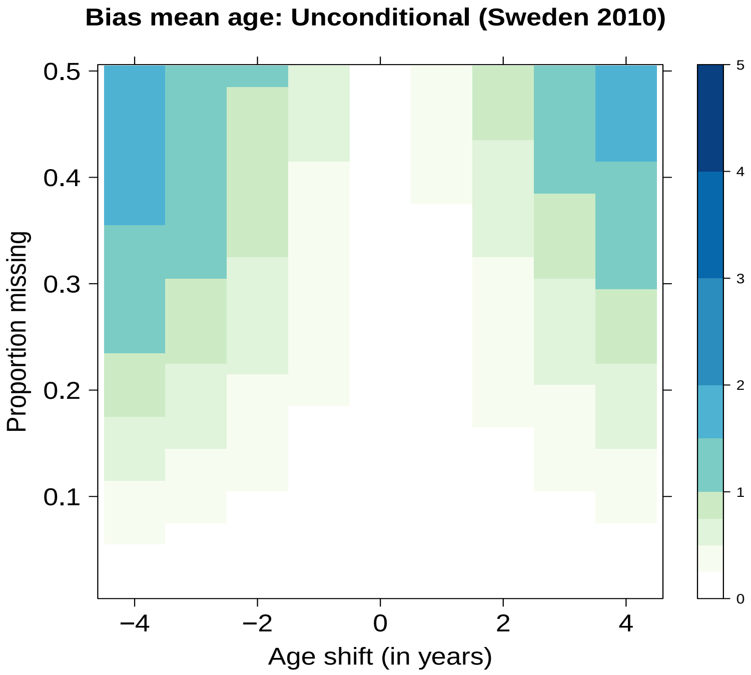

Figure 37: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Sweden 2010. Source: Own calculations.

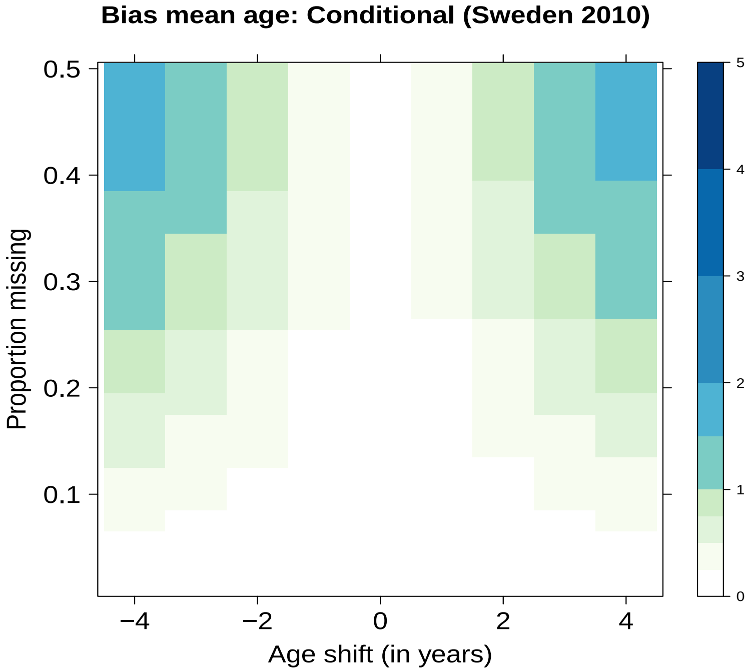

Figure 38: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Sweden 2010. Source: Own calculations.

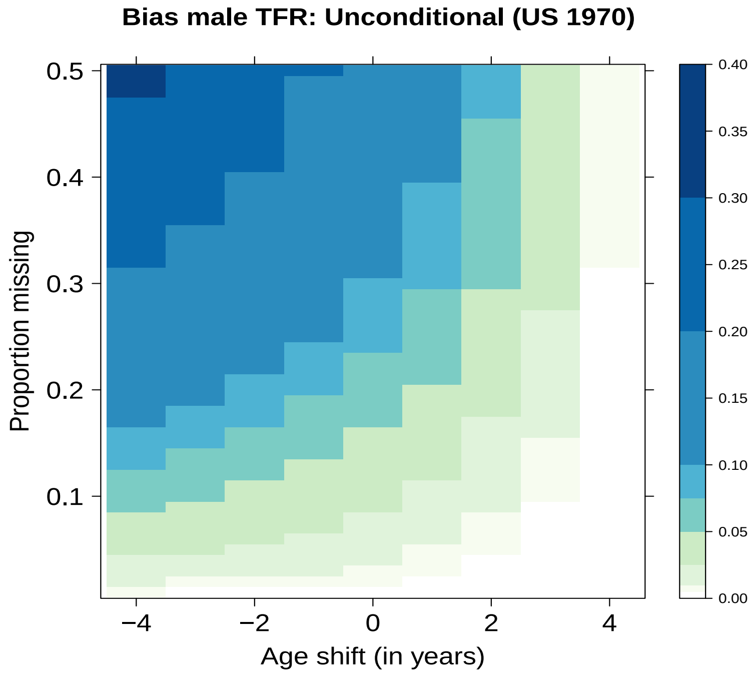

Figure 39: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1970. Source: Own calculations.

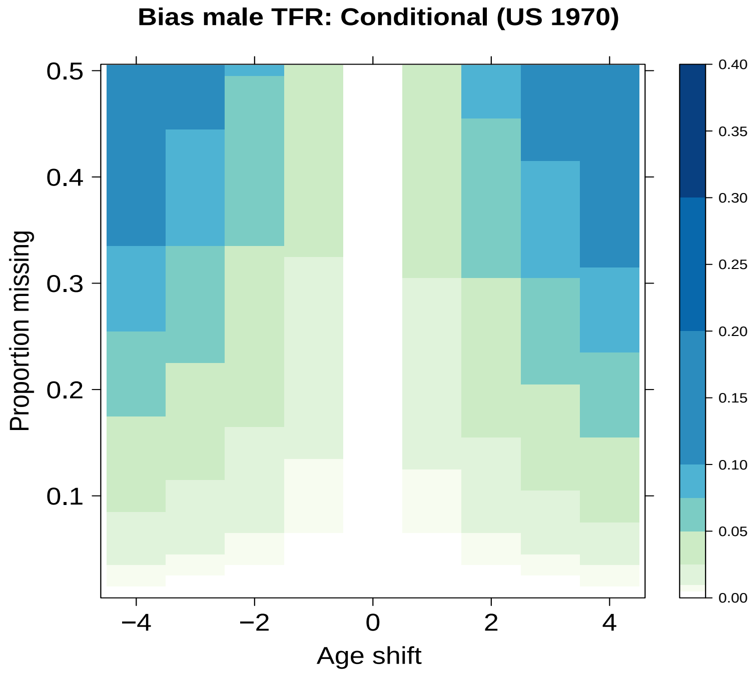

Figure 40: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1970. Source: Own calculations.

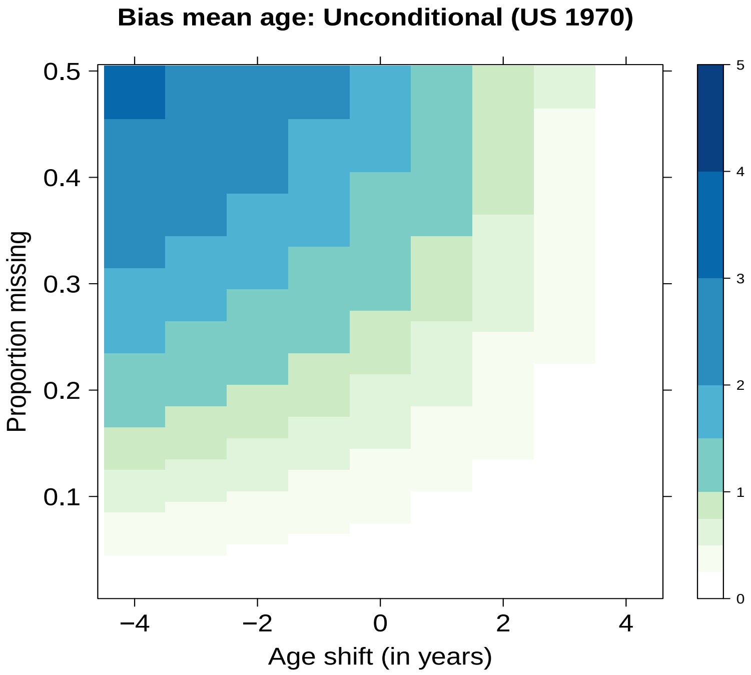

Figure 41: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1970. Source: Own calculations.

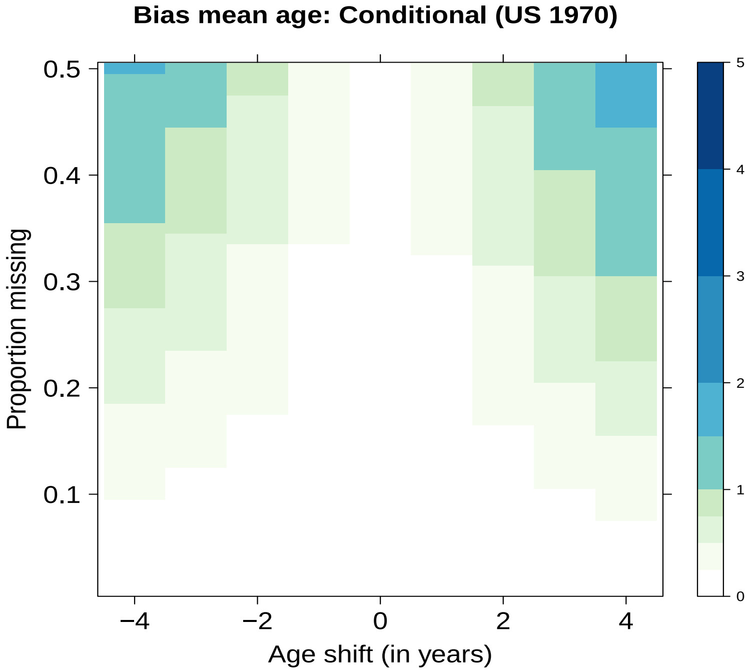

Figure 42: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1970. Source: Own calculations.

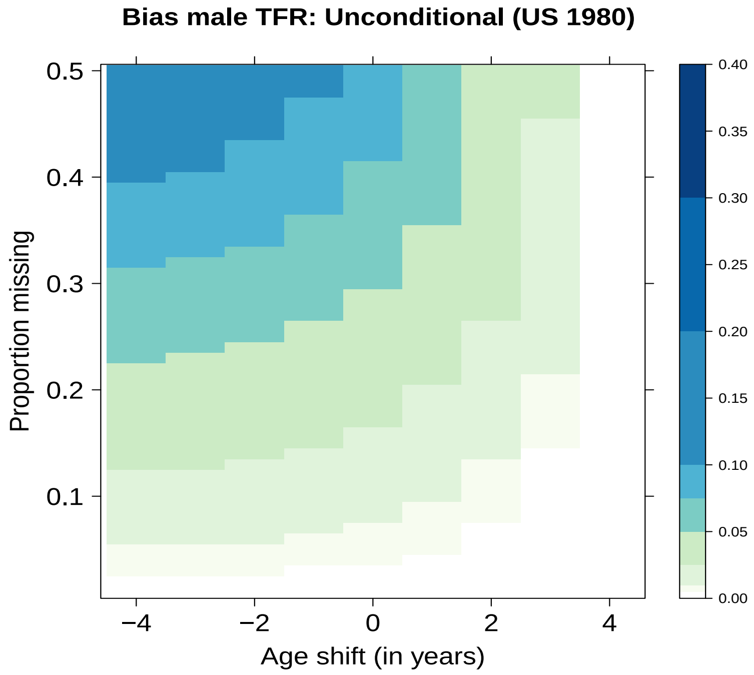

Figure 43: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1980. Source: Own calculations.

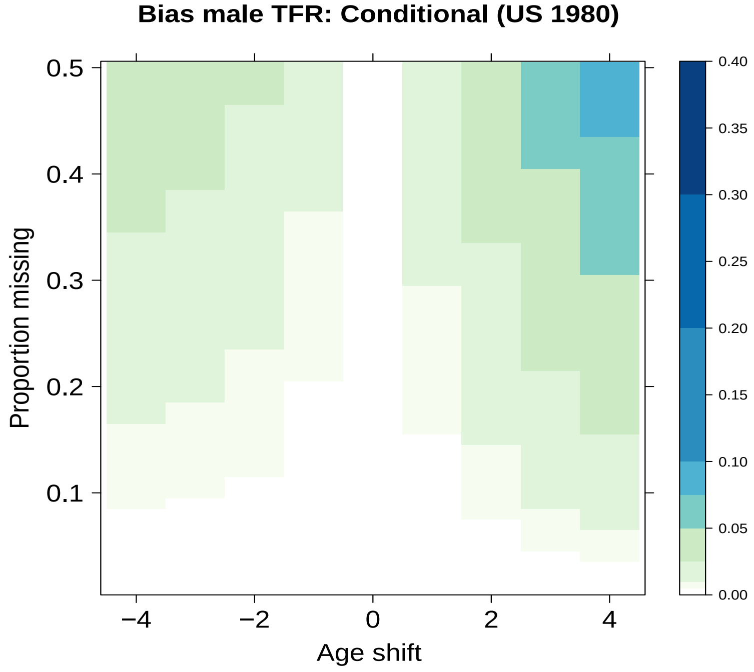

Figure 44: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1980. Source: Own calculations.

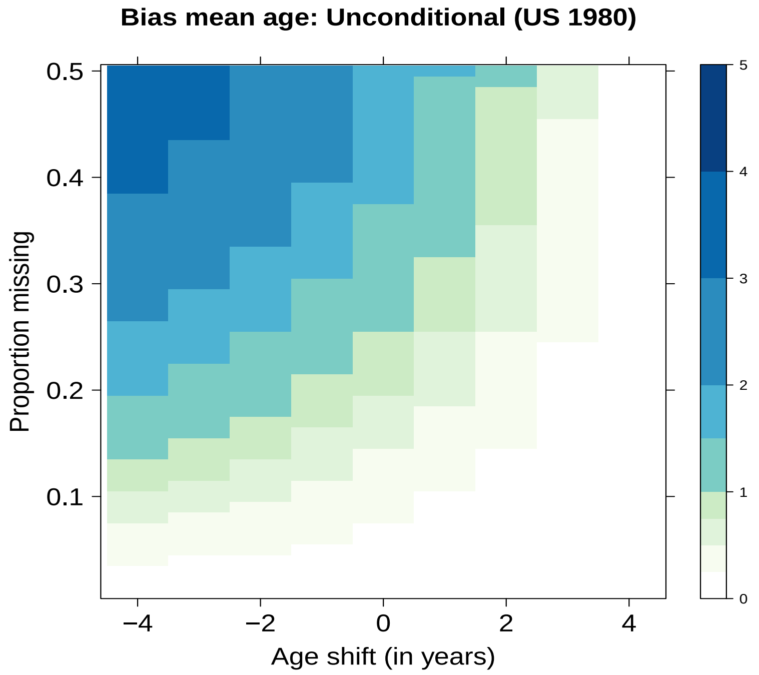

Figure 45: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1980. Source: Own calculations.

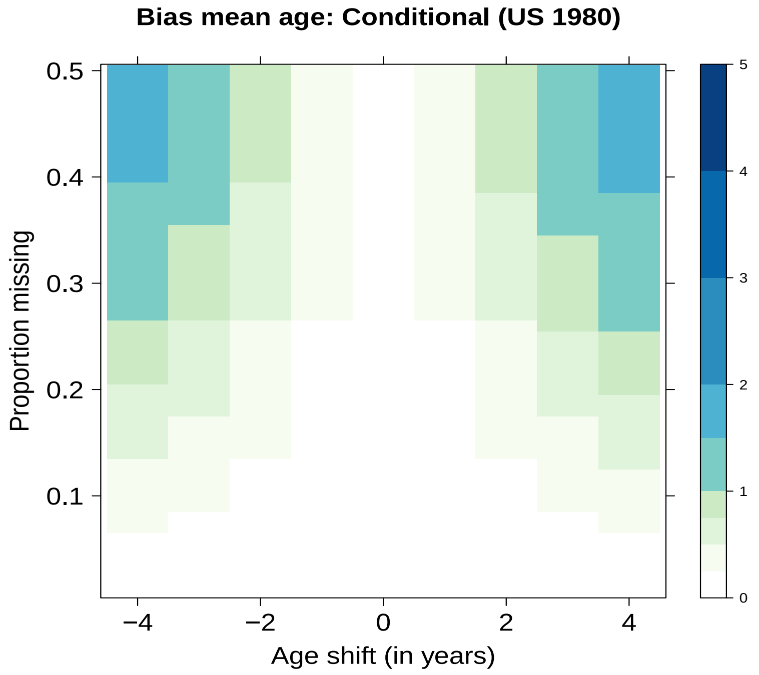

Figure 46: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1980. Source: Own calculations.

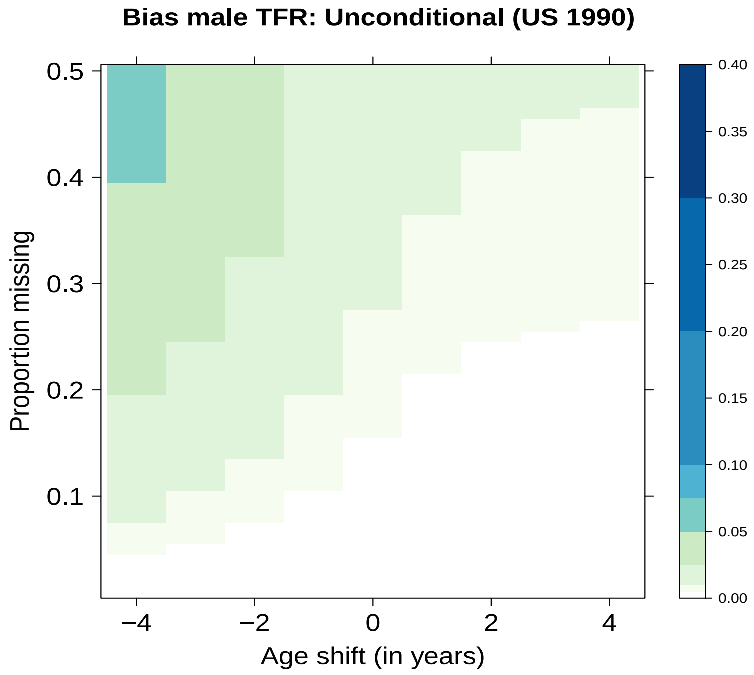

Figure 47: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1990. Source: Own calculations.

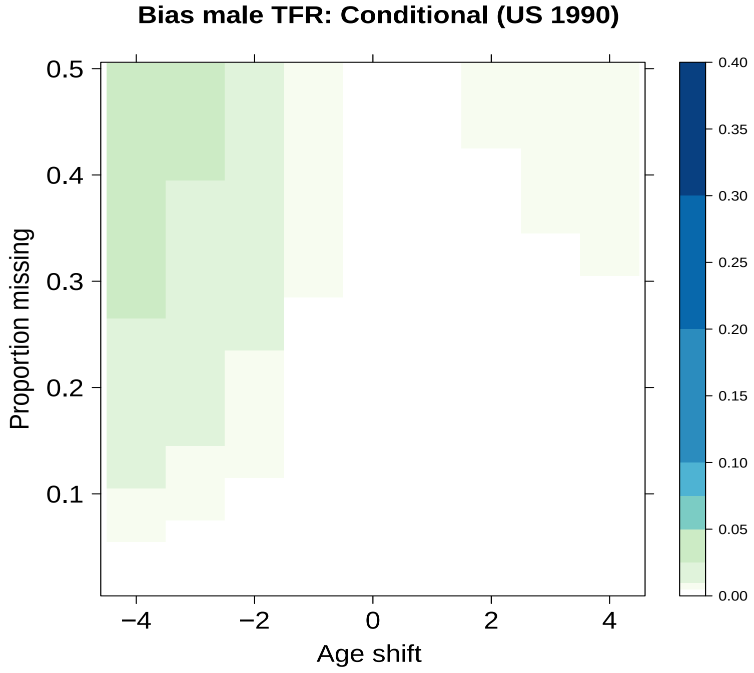

Figure 48: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1990. Source: Own calculations.

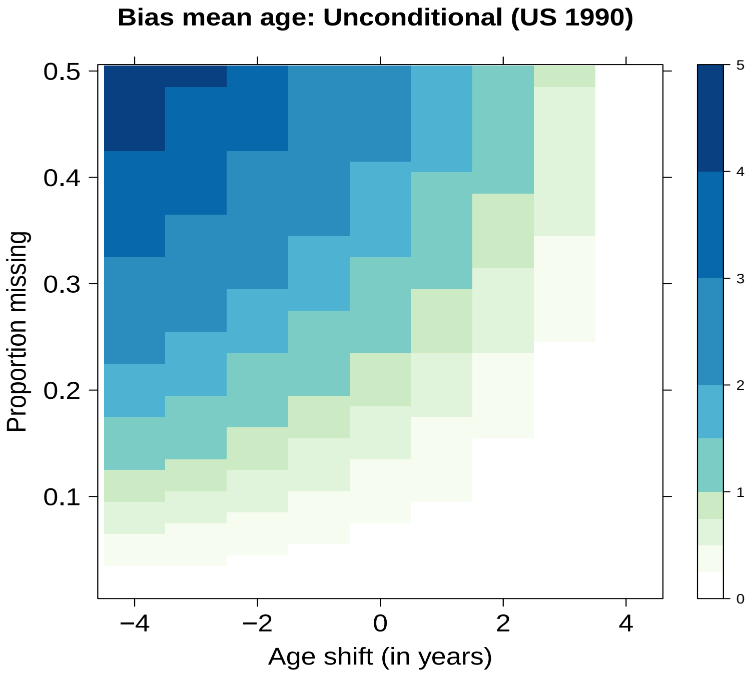

Figure 49: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 1990. Source: Own calculations.

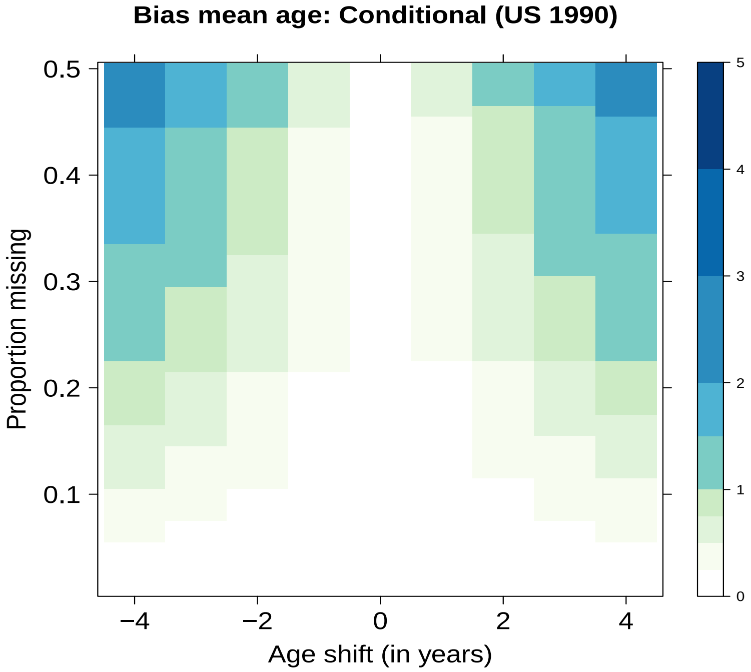

Figure 50: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for the U.S. 1990. Source: Own calculations.

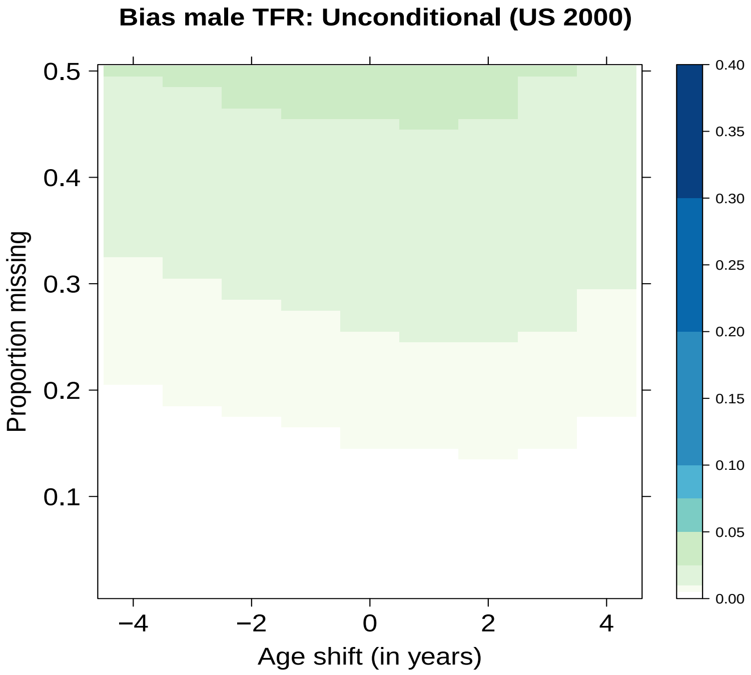

Figure 51: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 2000. Source: Own calculations.

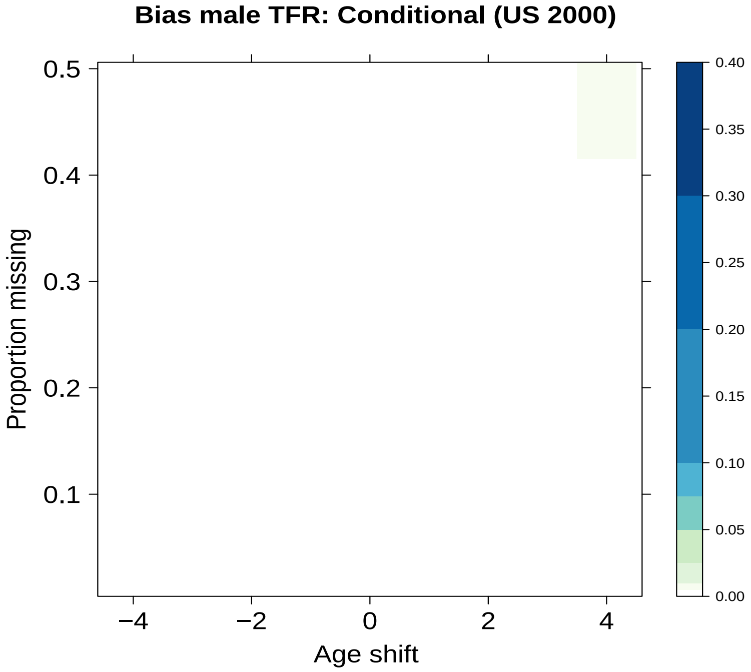

Figure 52: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for the U.S. 2000. Source: Own calculations.

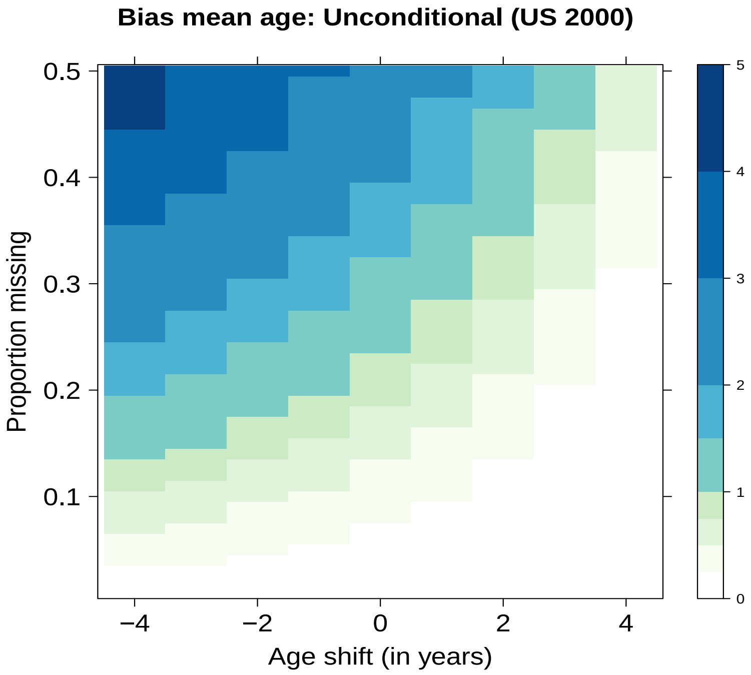

Figure 53: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 2000. Source: Own calculations.

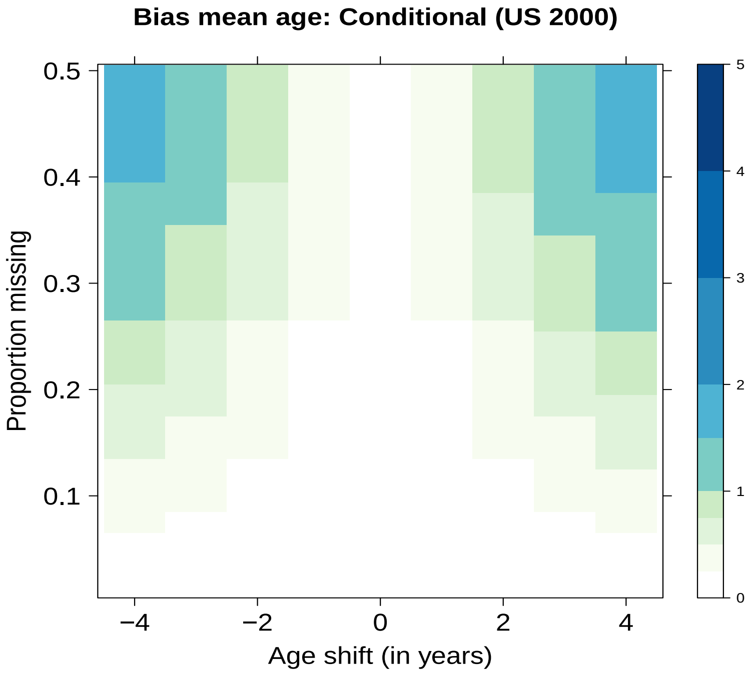

Figure 54: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for the U.S. 2000. Source: Own calculations.

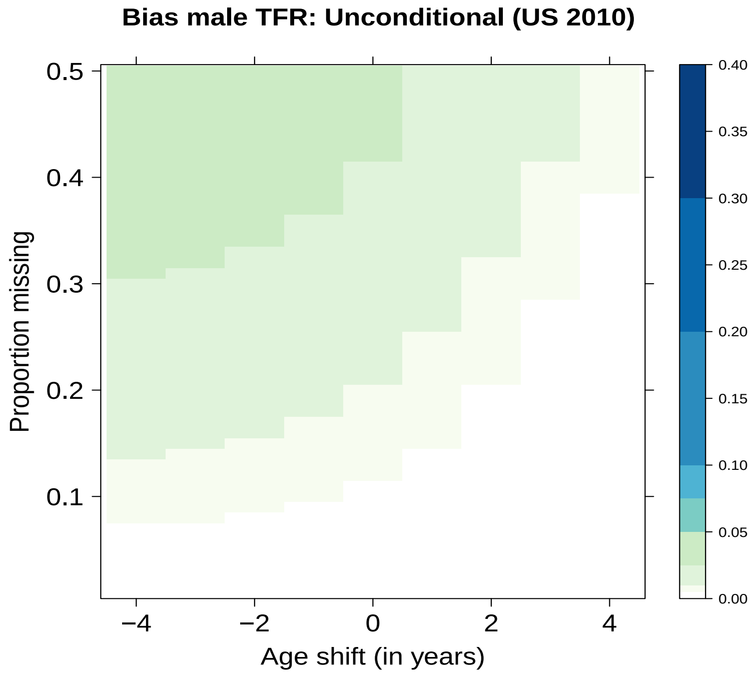

Figure 55: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 2010. Source: Own calculations.

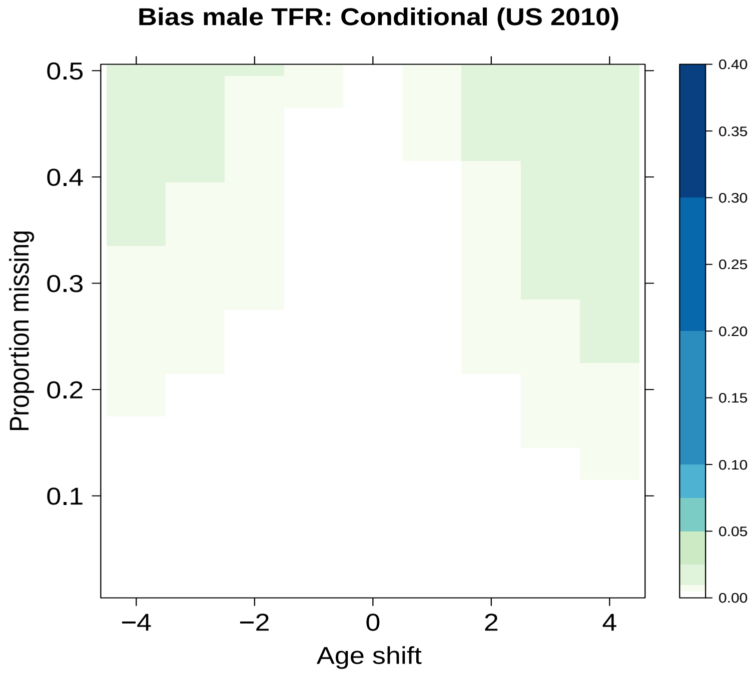

Figure 56: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for the U.S. 2010. Source: Own calculations.

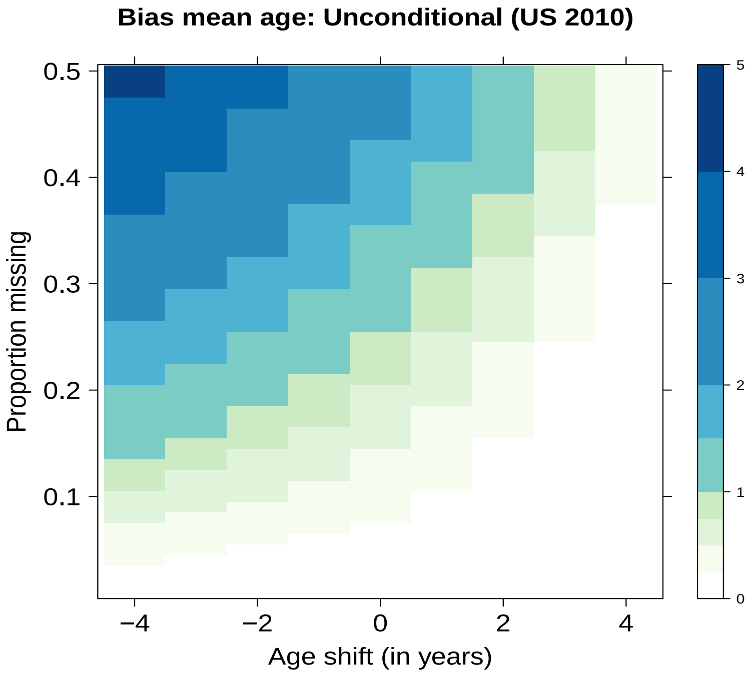

Figure 57: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for the U.S. 2010. Source: Own calculations.

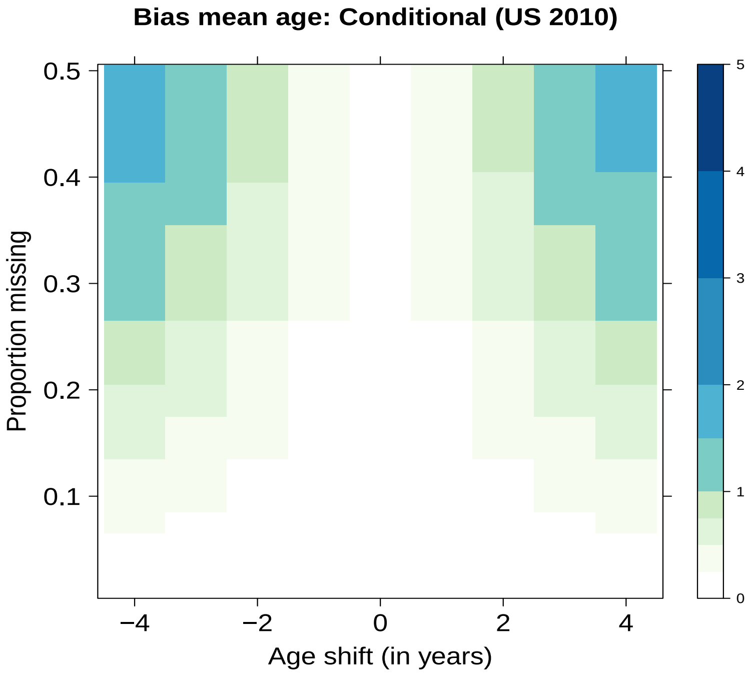

Figure 58: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for the U.S. 2010. Source: Own calculations.

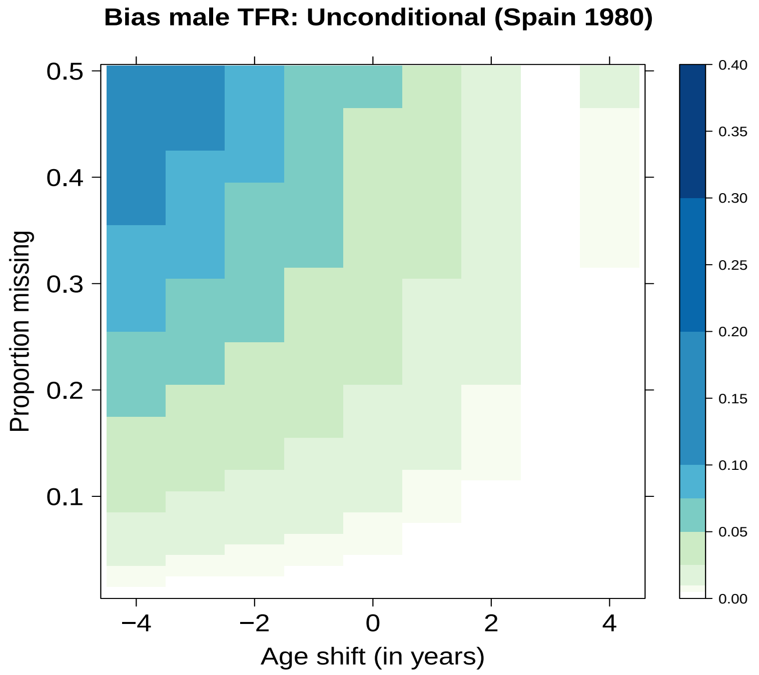

Figure 59: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Spain 1980. Source: Own calculations.

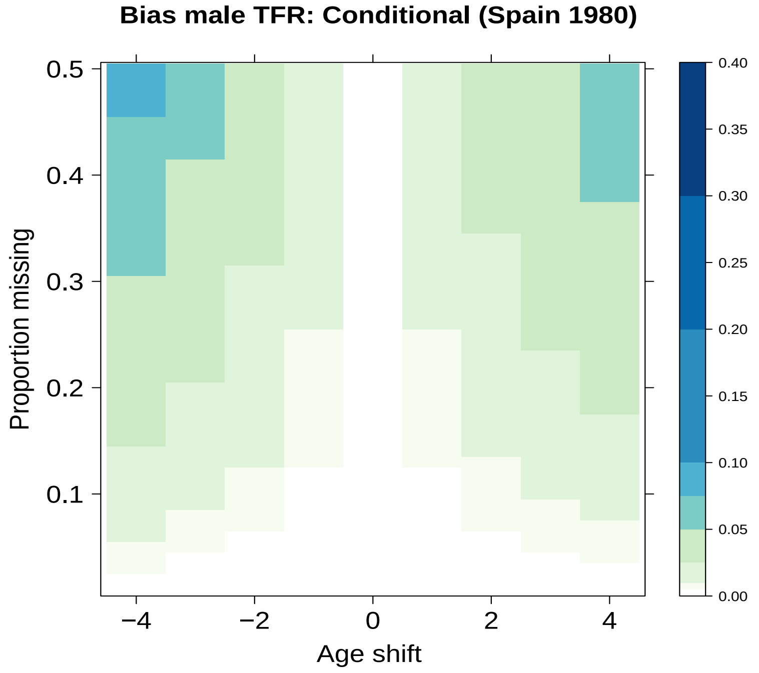

Figure 60: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Spain 1980. Source: Own calculations.

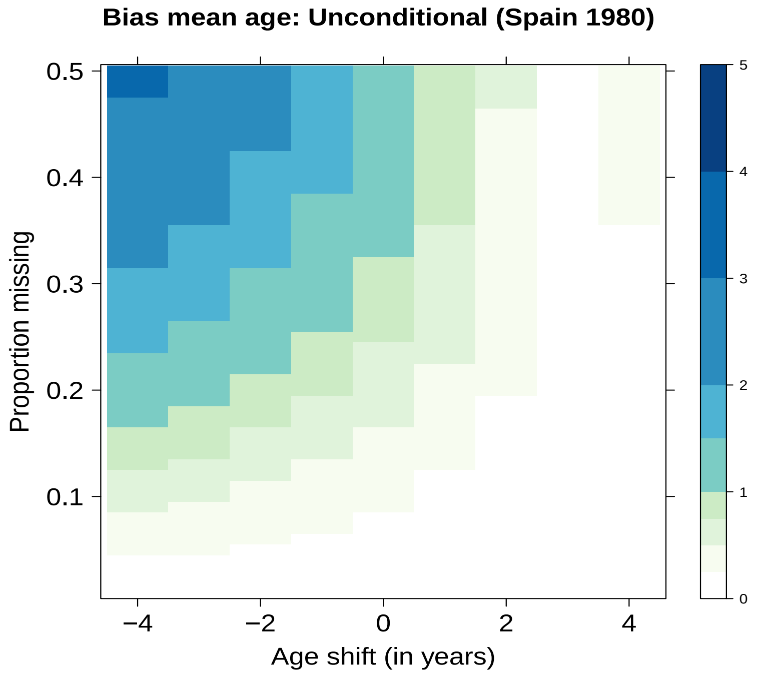

Figure 61: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Spain 1980. Source: Own calculations.

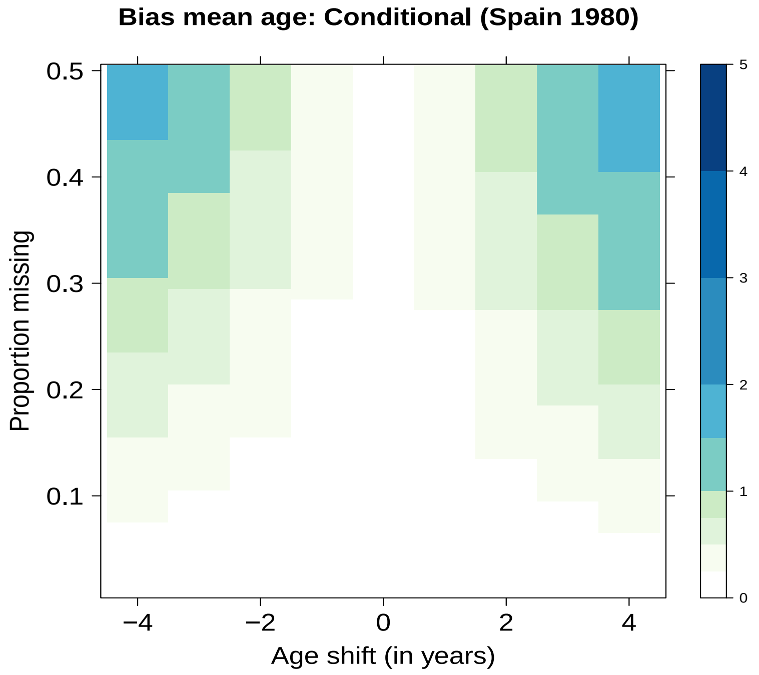

Figure 62: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Spain 1980. Source: Own calculations.

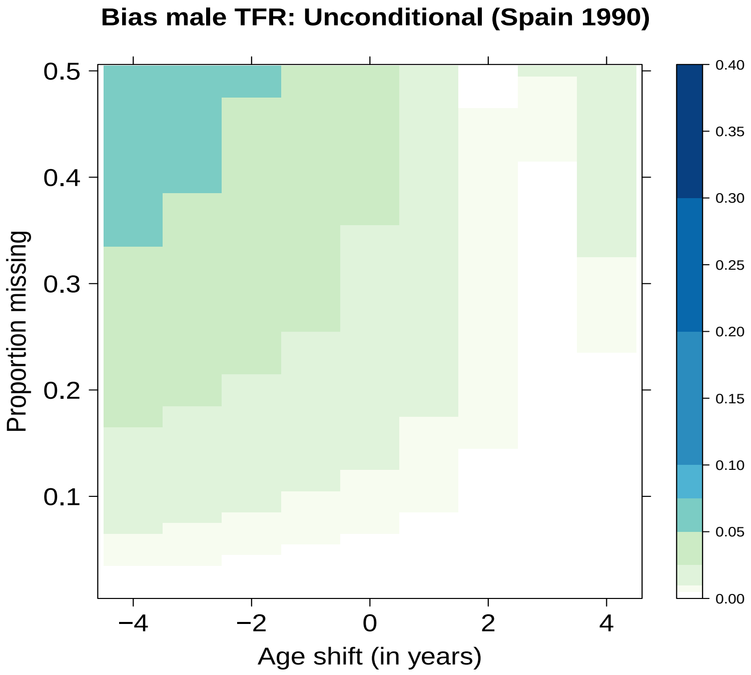

Figure 63: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Spain 1990. Source: Own calculations.

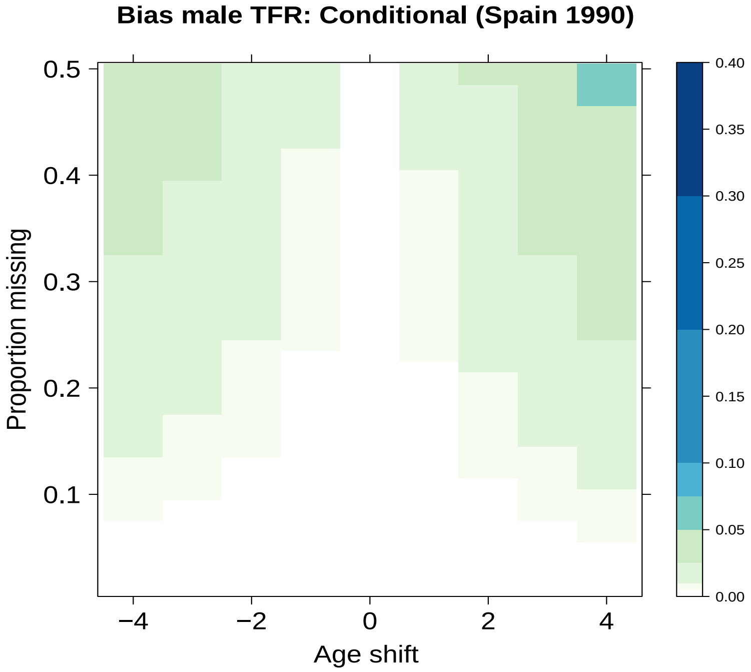

Figure 64: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Spain 1990. Source: Own calculations.

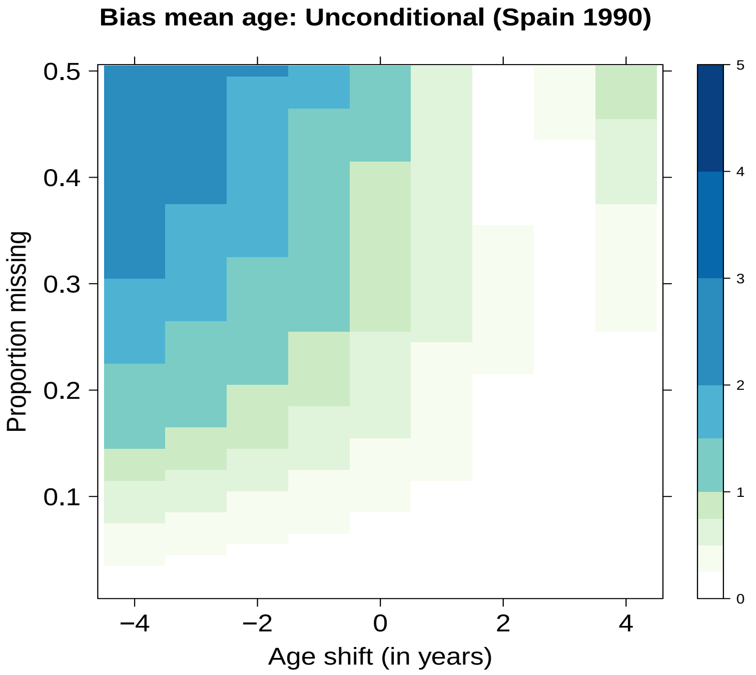

Figure 65: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Spain 1990. Source: Own calculations.

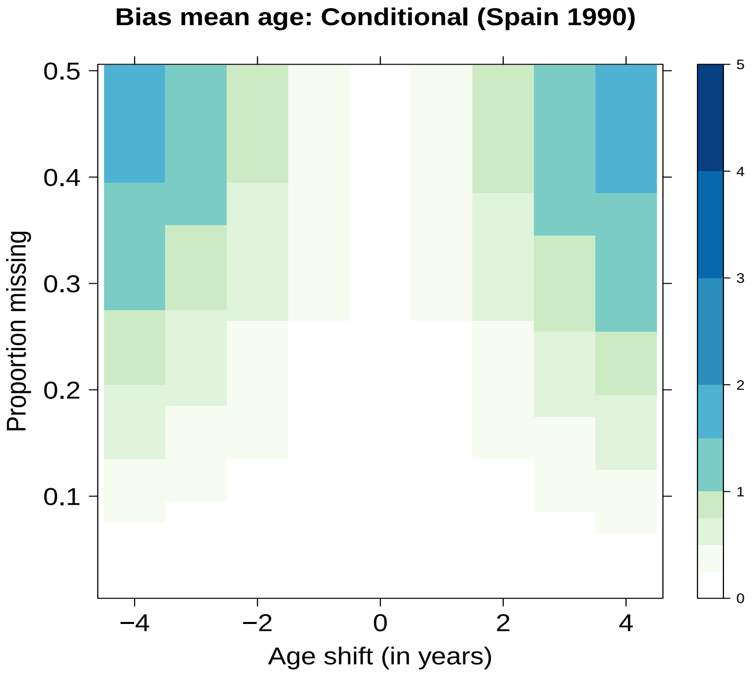

Figure 66: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Spain 1990. Source: Own calculations.

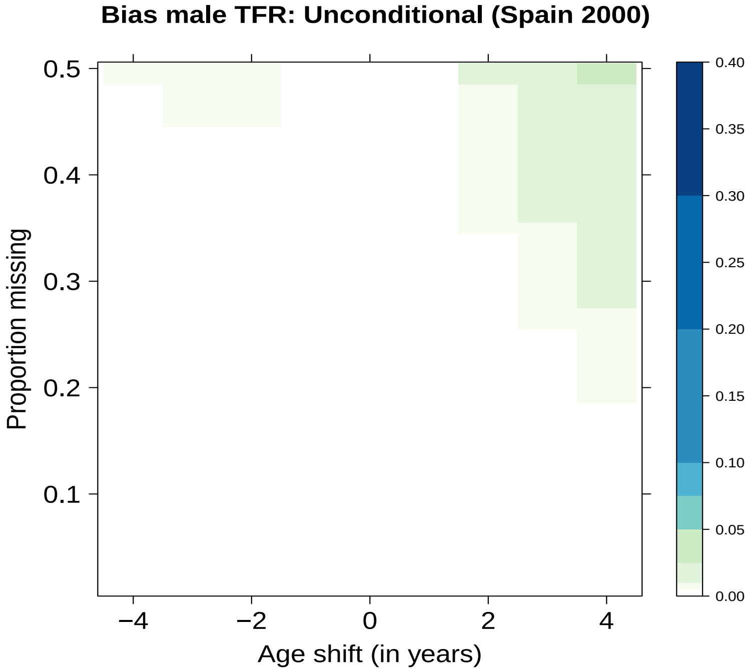

Figure 67: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Spain 2000. Source: Own calculations.

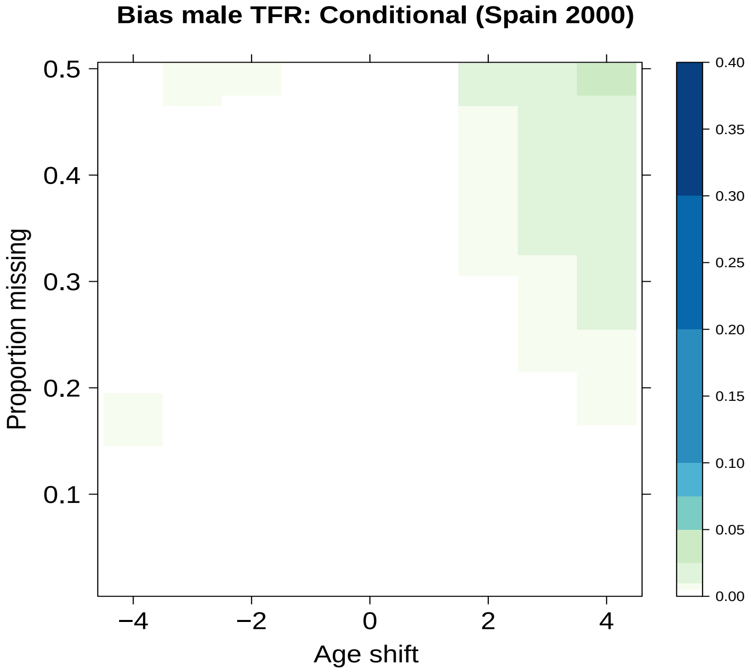

Figure 68: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Spain 2000. Source: Own calculations.

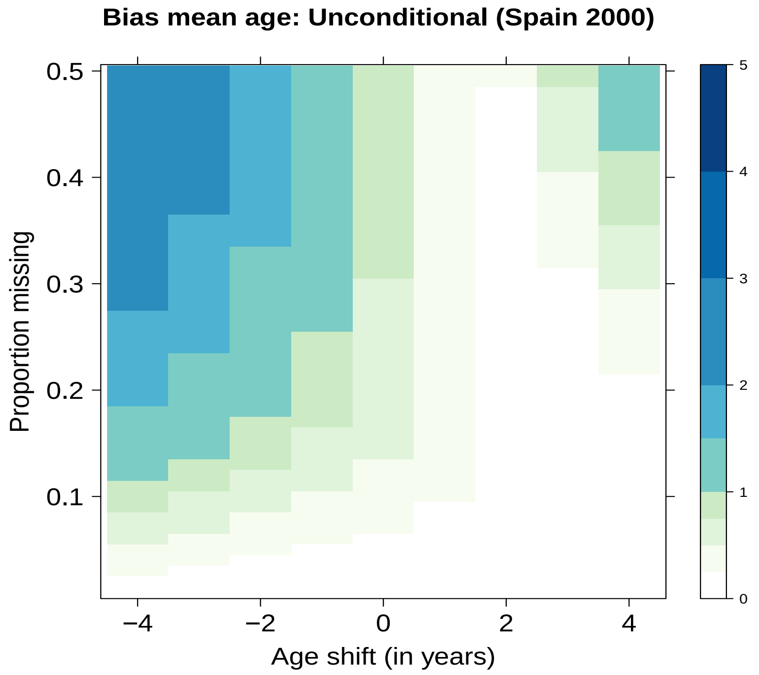

Figure 69: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Spain 2000. Source: Own calculations.

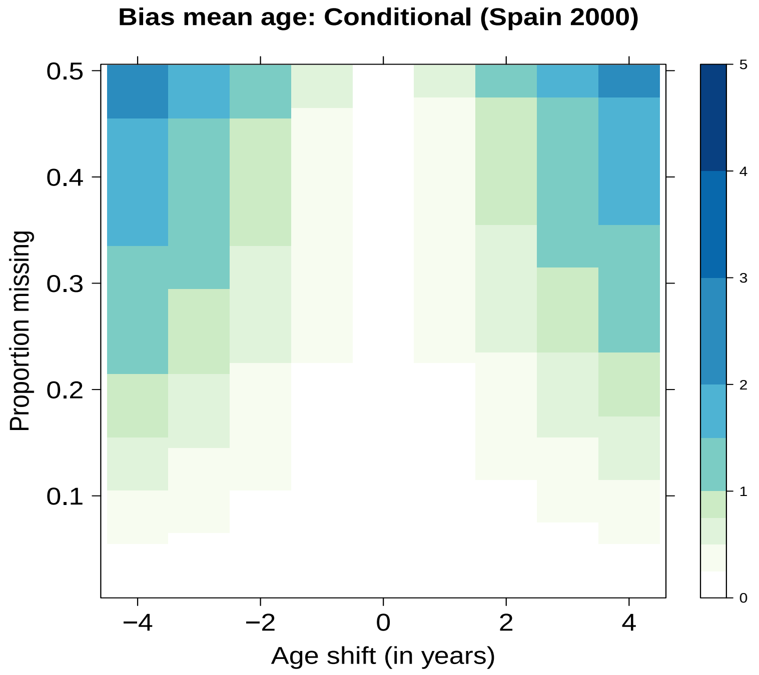

Figure 70: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Spain 2000. Source: Own calculations.

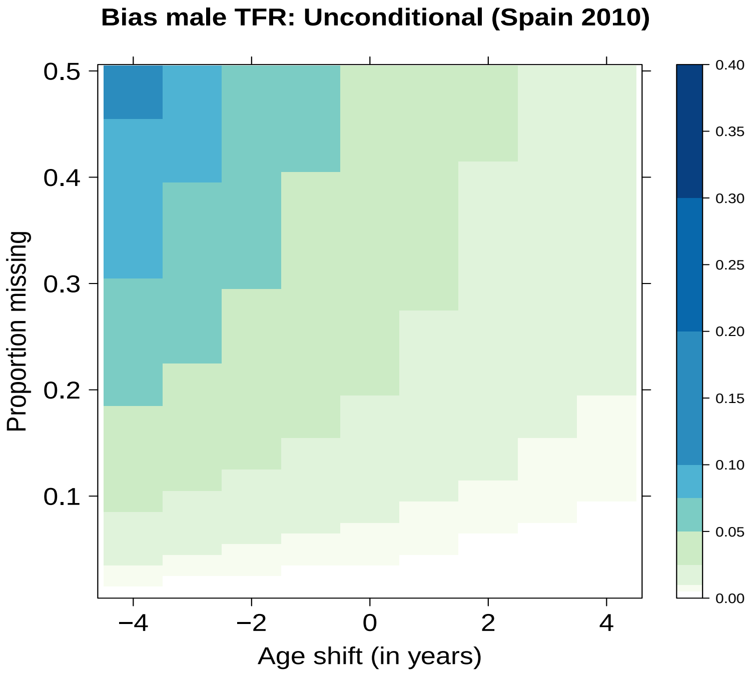

Figure 71: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Spain 2010. Source: Own calculations.

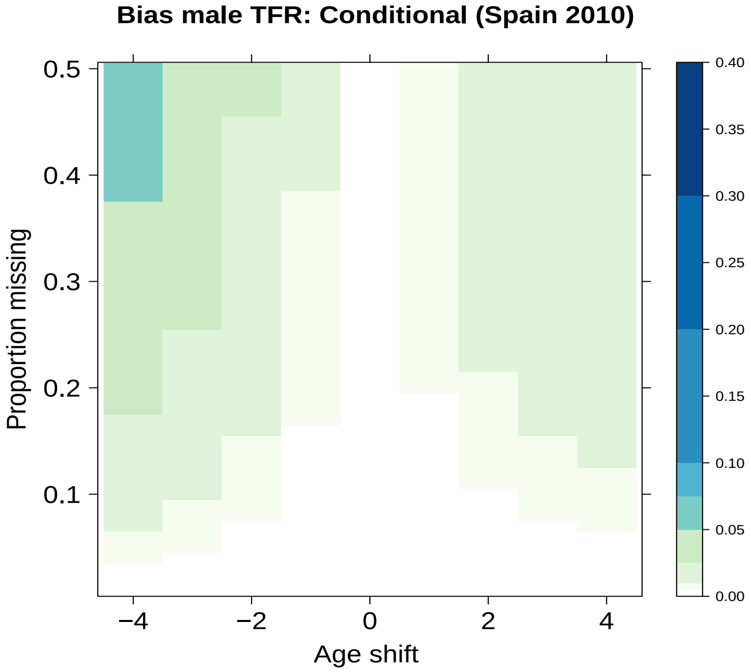

Figure 72: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Spain 2010. Source: Own calculations.

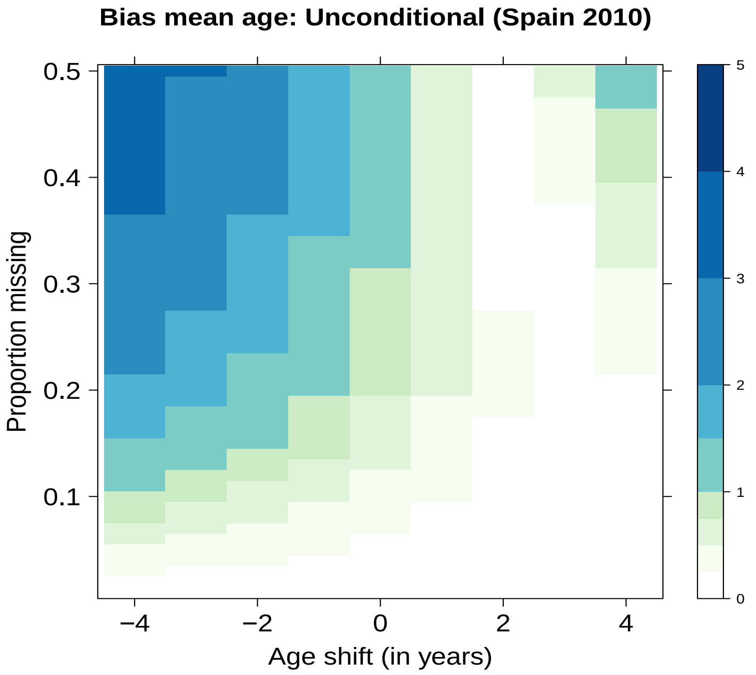

Figure 73: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Spain 2010. Source: Own calculations.

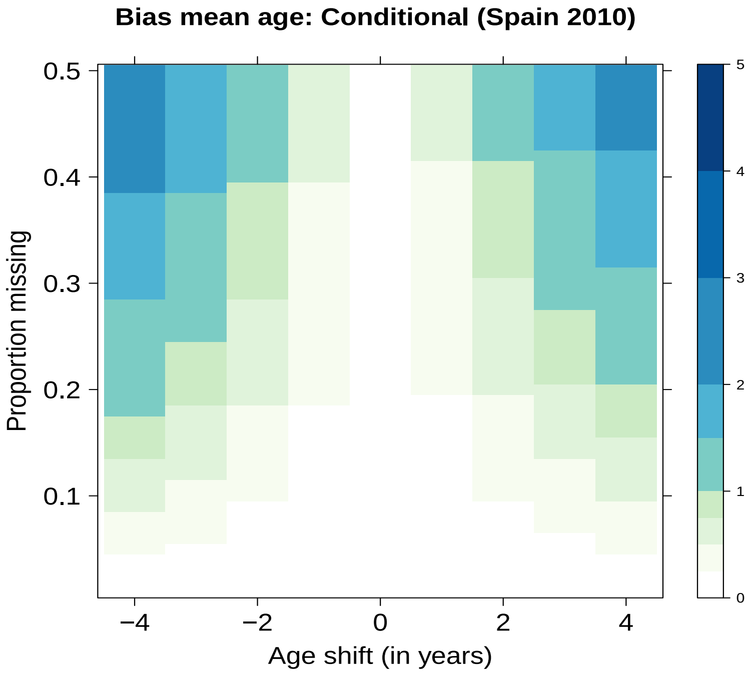

Figure 74: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Spain 2010. Source: Own calculations.

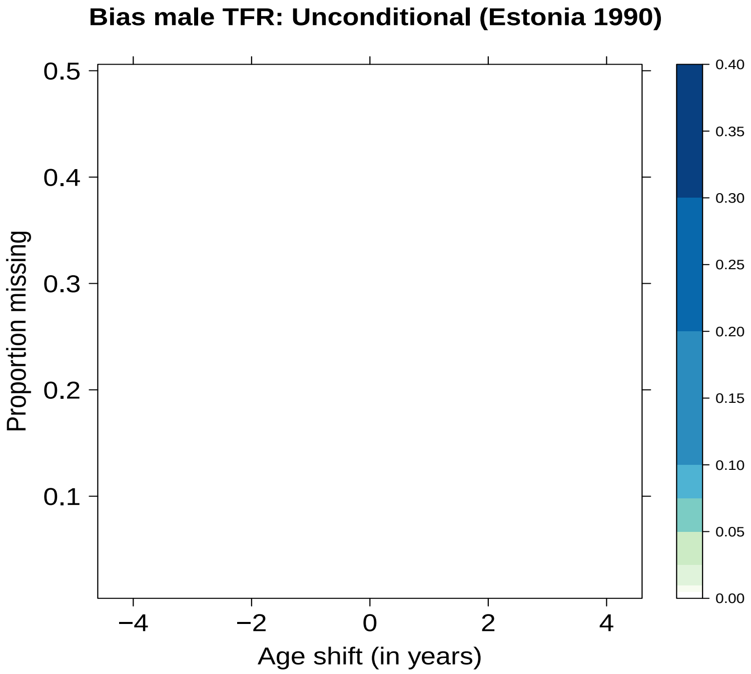

Figure 75: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Estonia 1990. Source: Own calculations.

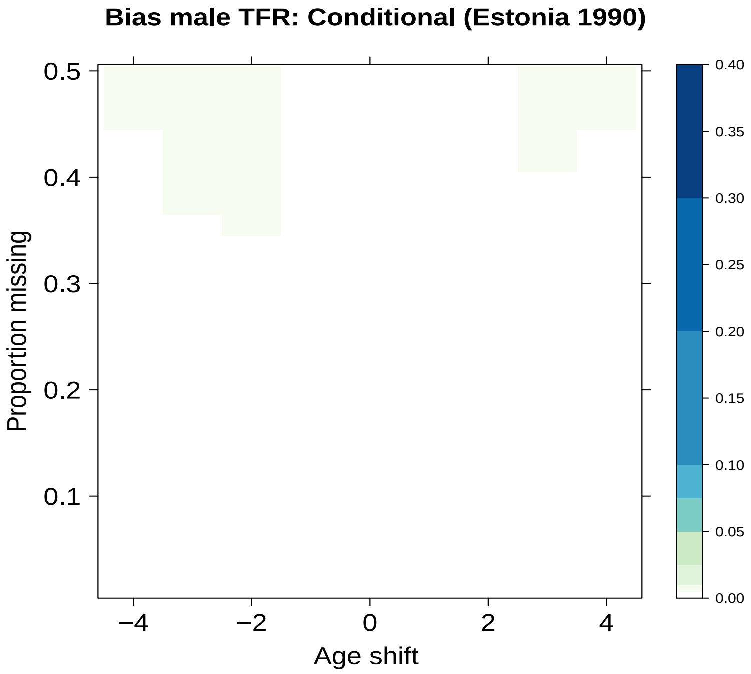

Figure 76: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Estonia 1990. Source: Own calculations.

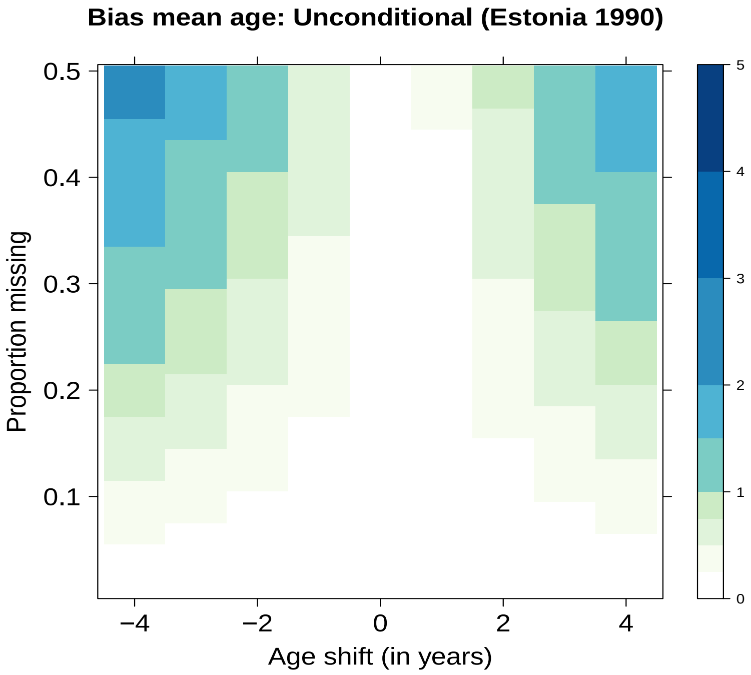

Figure 77: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Estonia 1990. Source: Own calculations.

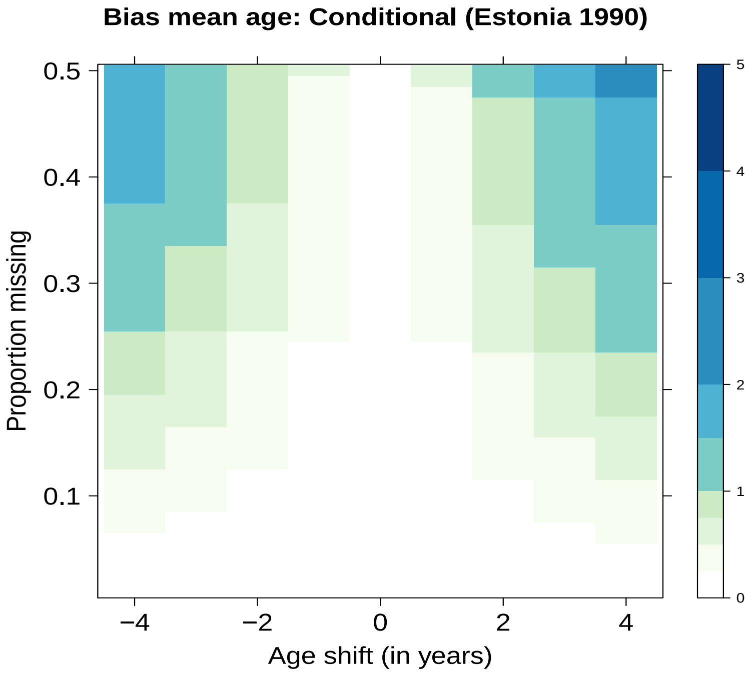

Figure 78: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Estonia 1990. Source: Own calculations.

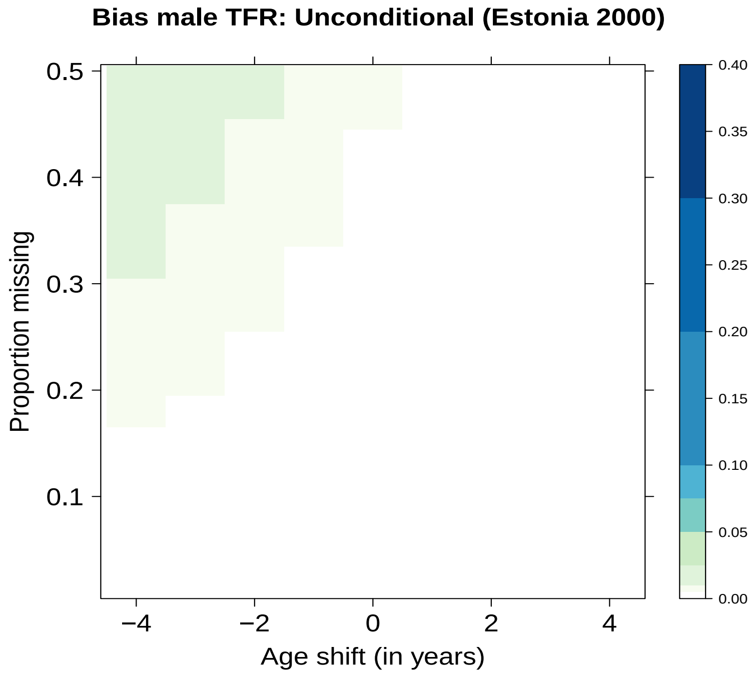

Figure 79: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Estonia 2000. Source: Own calculations.

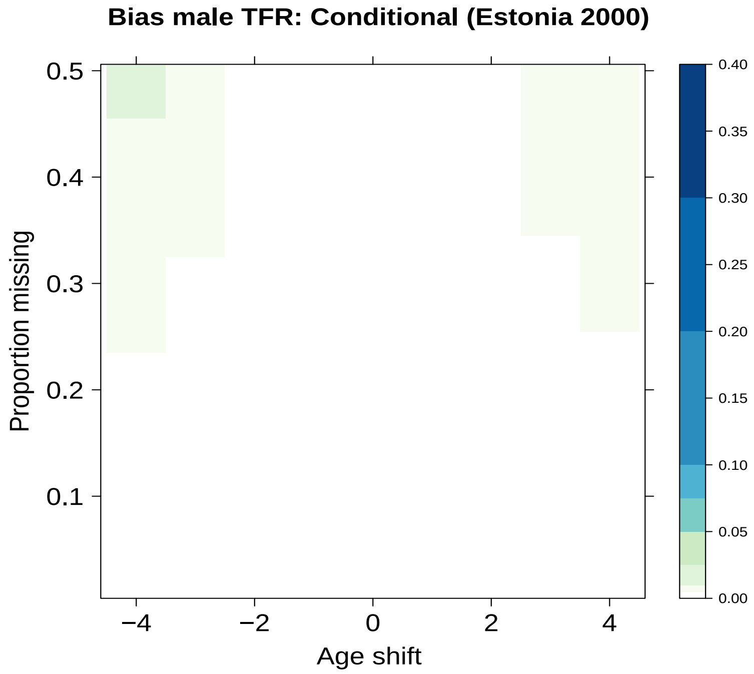

Figure 80: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Estonia 2000. Source: Own calculations.

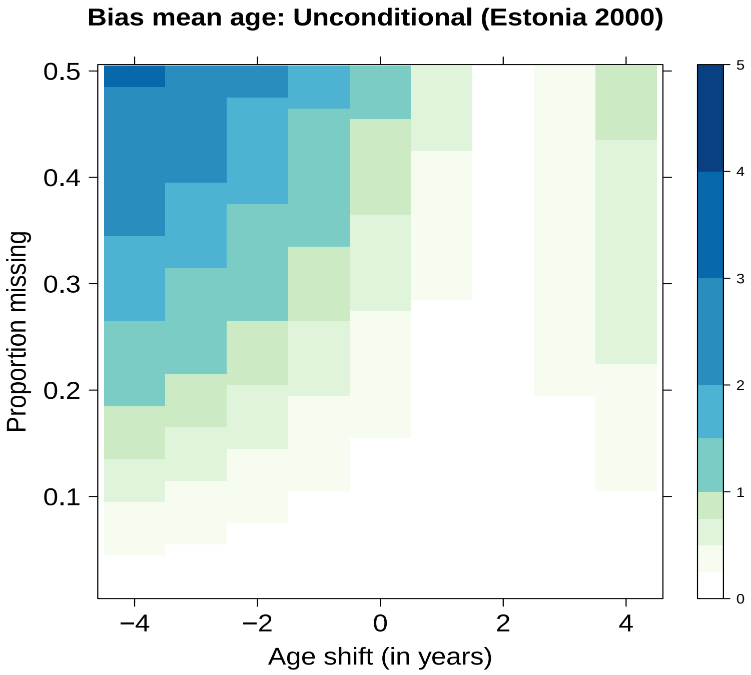

Figure 81: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Estonia 2000. Source: Own calculations.

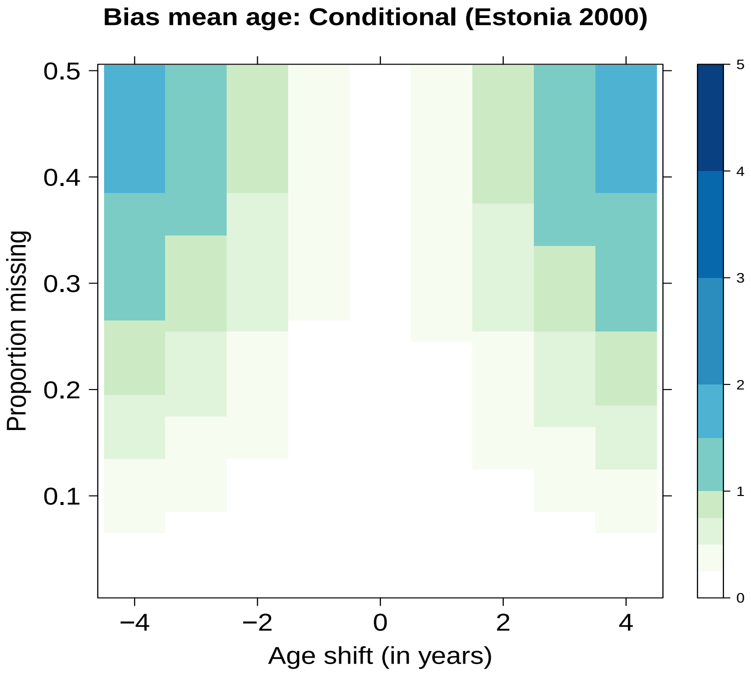

Figure 82: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Estonia 2000. Source: Own calculations.

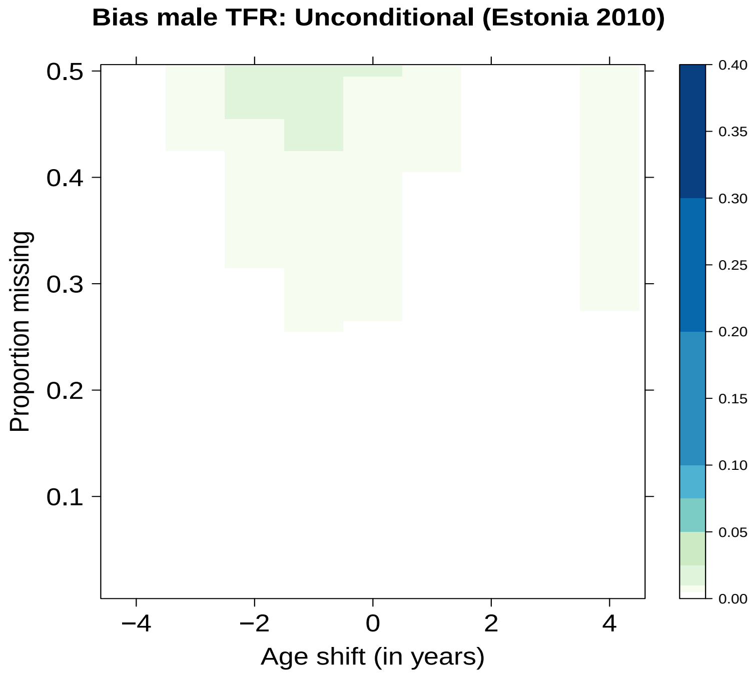

Figure 83: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the unconditional approach by age shift and proportion missing; simulations for Estonia 2010. Source: Own calculations.

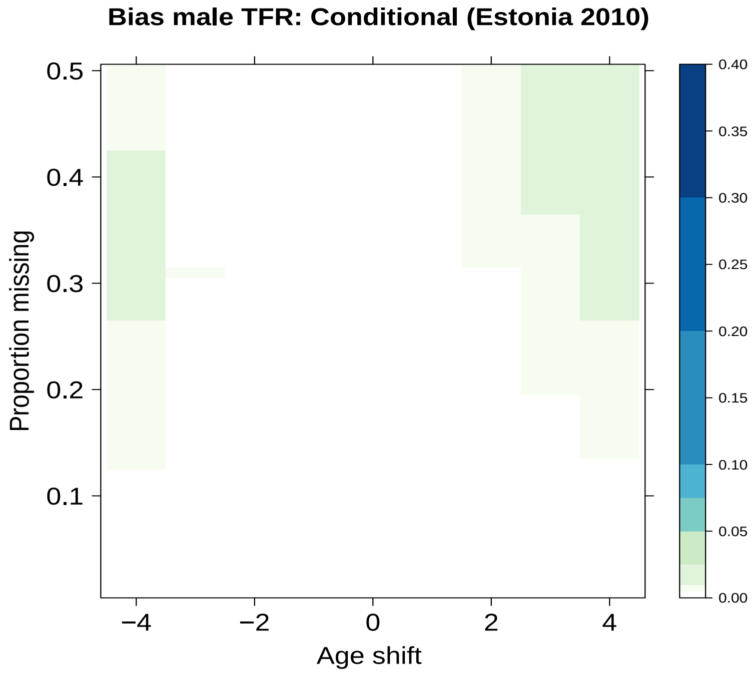

Figure 84: Absolute bias of the male total fertility rate (MTFR) in MTFR points based on the conditional approach by age shift and proportion missing; simulations for Estonia 2010. Source: Own calculations.

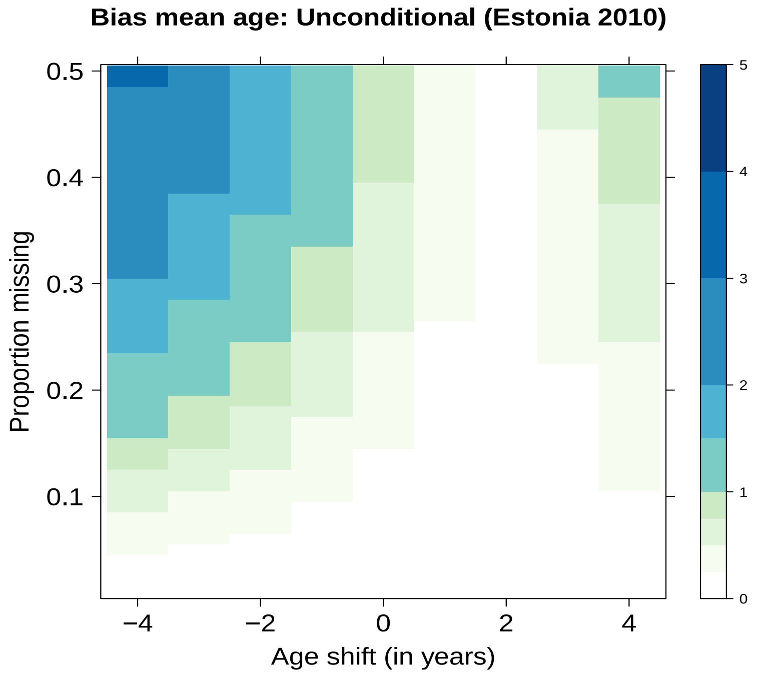

Figure 85: Absolute bias of the paternal mean age at childbearing (PMAC) based on the unconditional approach by age shift and proportion missing; simulations for Estonia 2010. Source: Own calculations.

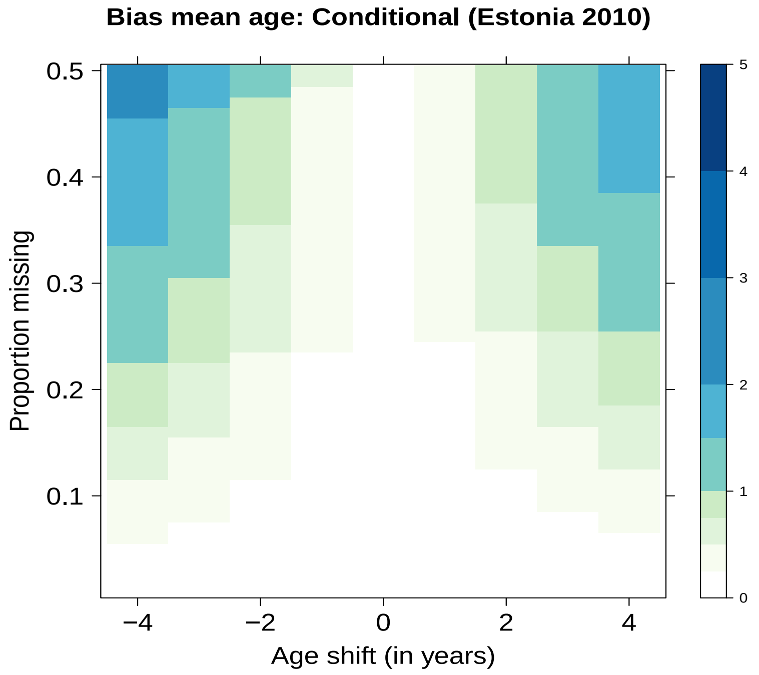

Figure 86: Absolute bias of the paternal mean age at childbearing (PMAC) based on the conditional approach by age shift and proportion missing; simulations for Estonia 2010. Source: Own calculations.

D Additional figures on female fertility

Figures 87 and 88 provide additional information on the contexts in which we apply simulations by displaying trends in female fertility, while Figure 89 presents trends in the proportion of missing values in the birth register data on which we base our simulations.

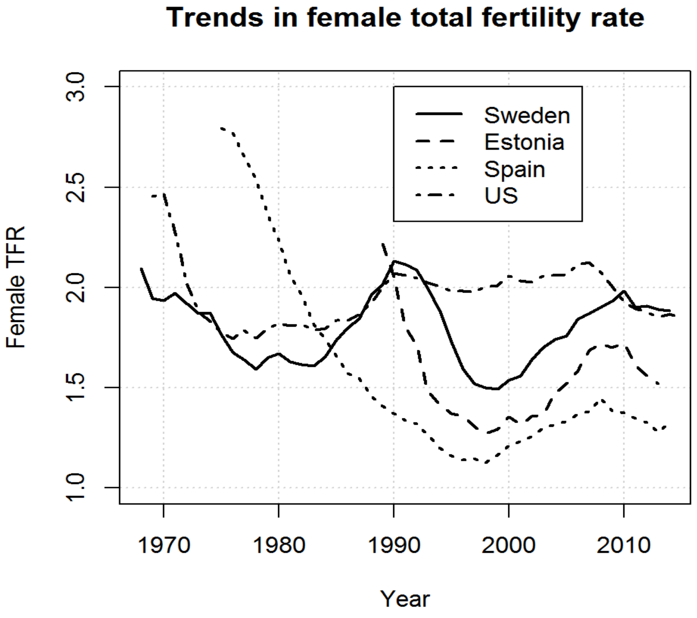

Figure 87: Trends in the total fertility rates of females by country. Source: Statistics Sweden, NBER, Spanish Statistical Office, Statistics Estonia, HMD; own calculations.

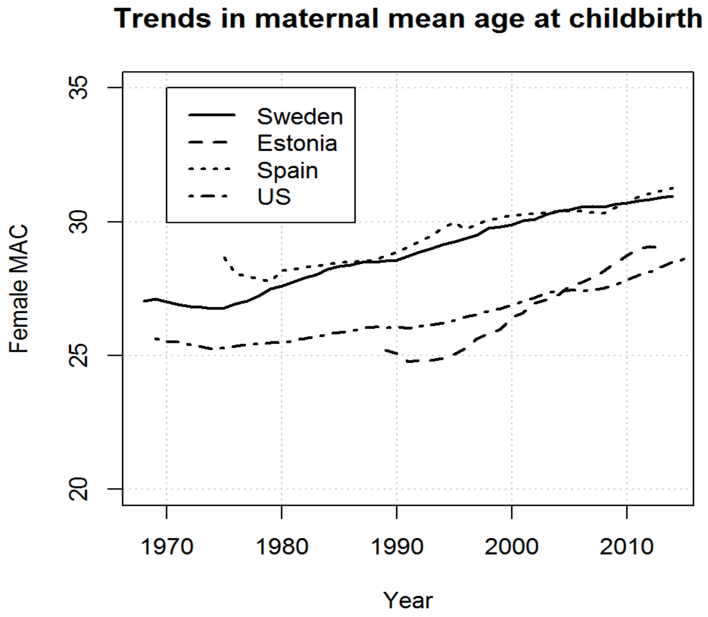

Figure 88: Trends in the female mean age at childbirth by country. Source: Statistics Sweden, NBER, Spanish Statistical Office, Statistics Estonia, HMD; own calculations.

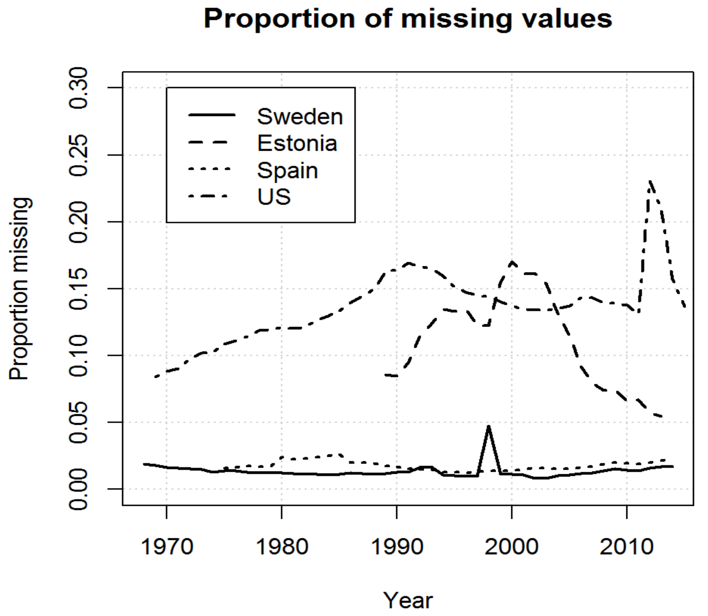

Figure 89: Trends in the proportion of births with unknown paternal age by country. Source: Statistics Sweden, NBER, Spanish Statistical Office, Statistics Estonia, HMD; own calculations.