– Supplementary material for 35–53 – Estimating male fertility in eastern and western Germany since 1991: A new lowest low?

Christian Dudel & Sebastian Klüsener

1 Adjusted population denominators

The denominators of the fertility rates are based on official population counts. Here we faced the challenge that German population estimates published in the 1990s and the 2000s are substantially biased. This is related to the rather chaotic conditions surrounding German reunification in 1990, and the long gaps between the last censuses in East and West Germany (1981/1987) and the first census after reunification in 2011. As there will be no official backward adjustment, we decided to use adjusted numbers based on an approach developed by Klüsener et al. (2016). In a first step, it accounts for substantial official corrections of, in part long-standing, errors in the population estimates implemented in the years prior to the 2011 census, and adds these to the difference between the population estimates based on the old censuses and the new census. In a second step, the difference between the old and new population estimates by cohort, sex, and region is assumed to have accumulated uniformly over the inter-censal period (for details see Klüsener et al. 2016). The population denominator is obtained by taking the mean of the adjusted population by single age at the beginning and end of the year.

2 Imputation methods: Matching and regression approach

We apply four approaches to impute the missing age of the father. The first approach and the second approach are described in the main text. The third imputation approach is based on linear regression. The maternal age, the maternal age squared, the maternal employment status and the federal state in which the mother has her place of residence are used as explanatory variables. For this we specified the

following regression:

$$y_i = \alpha + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 e_i + \sum_{j=2}^{16} \gamma_j s_{ij} + e_i,$$

where y_i denotes the age of the father for birth i, a_i the age of the mother, e_i the employment status of the mother (1=employed), and s a set of dummy variables capturing in which of the 16 federal states j the mother resided while giving birth. The regression equation is estimated separately for each year and region using all nonmarital births with information on the paternal age. Coefficient estimates are then used to impute the (rounded) age of the father for nonmarital births for which this information is missing.

The fourth approach uses a non-parametric one-to-one nearest-neighbor matching approach, controlling for the same variables as the regression approach. For each nonmarital birth without information on the paternal age, the most similar nonmarital birth with paternal age information is selected and used for imputing the paternal age. To assess which observation is the most similar, we chose the Gower distance (e.g., Dettmann, Becker, and Schmeißer 2011),

$$D_{ik} = \sum_{v=1}^{n} d_{v,ik},$$

where D_{ik} is the distance between observations i and k, n is the number of variables, and $d_{v,ik}$ is the distance between observations i and k for variable v. The latter is defined as

$$d_{v,ik} = \begin{cases} |x_{vi} - x_{vk}| / \max[|x_{va} - x_{vb}|] & \text{if } x_v\text{is metrical} \\ 1 - \mathbb{I}(x_{vi} = x_{vk}) & \text{if } x_v\text{is nominal} \end{cases},$$

where x_{vi} is the value of variable v for observation i, $\max[|x_{va} - x_{vb}|]$ is the maximum difference occurring for variable x_v, and $\mathbb{I}(x_{vi} = x_{vk})$ is an indicator variable which equals 1 if x_{vi} equals x_{vk} and 0 otherwise. This approach is also applied separately for each year and region.

3 Implementation of the sensitivity analyses

Our sensitivity analyses regarding estimates of paternal ages for cases in which this information is missing are based on two scenarios. In the first scenario the average age difference between fathers and mothers of three years is roughly doubled, while in the second scenario it is set to a negative value. To generate the scenarios, we

shift the distribution of the age of the father conditional on the age of the mother by either plus or minus four years. That is, if $p(x|y)$ is the proportion of fathers aged x conditional on the mother being aged y, then the shifted distribution $p_S(x|y)$ is calculated as either $p_S(x+4|y) = p(x|y)$ or $p_S(x-4|y) = p(x|y)$. If $x \pm 4$ is outside the age range 17-59, it is replaced with either 17 or 59. For example, in 2010 the average age of the father conditional on the mother being age 30 was 33.5 years; i.e., the average age difference was 3.5 years. In our sensitivity analysis the average age of the father is either 29.5 or 37.5, resulting in age differences of -0.5 and 7.5 years, respectively.

References

Dettmann, E., Becker, C., and Schmeißer, C. (2011). Distance functions for matching in small samples. *Computational Statistics and Data Analysis* 55: 1942–1960.

Klüsener, S., Grigoriev, P., Scholz, R.D., and Jdanov, D.A. (2016). Adjustment of inter-censal population estimates for Germany 1987–2011: The implementation for the Human Mortality Database. Rostock: Max Planck Institute for Demographic Research (MPIDR working paper; forthcoming).