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 2023-034 l August 2023
Revised August 2023
https://doi.org/10.4054/MPIDR-WP-2023-034

Analyzing biases in genealogies using demographic microsimulation

Liliana P. Calderón-Bernal l calderonbernal@demogr.mpg.de
Diego Alburez-Gutierrez l alburezgutierrez@demogr.mpg.de
Emilio Zagheni l office-zagheni@demogr.mpg.de

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.

Analyzing biases in genealogies using demographic microsimulation

Liliana P. Calderón-Bernal

Max Planck Institute for Demographic Research and Stockholm University

calderonbernal@demogr.mpg.de

Diego Alburez-Gutierrez

Max Planck Institute for Demographic Research

Emilio Zagheni

Max Planck Institute for Demographic Research

August 24, 2023

Abstract

Genealogies are promising sources for addressing many questions in historical and kinship demography. So far, an incomplete understanding of the biases that affect their representativeness has hindered their full exploitation. Here, we report on a series of experiments on synthetic populations aimed at understanding how different sources of bias in ascendant genealogies can affect the accuracy of demographic estimates. We use the SOCSIM demographic microsimulation program and data for Sweden from the Human Fertility Collection (1751-1890), the Human Fertility Database (1891-2022), and the Human Mortality Database (1751-2022). We analyze three sources of bias: selection in direct lineages, incomplete reconstruction of family trees, and missing information on some subpopulations. We evaluate their effect by comparing common demographic measures estimated from 'fully-recorded' and 'bias-infused' synthetic populations. Our results show that including only direct lineages leads to an underestimation of Total Fertility Rate (TFR) (c.a. -39% or 0.61 times lower) before the onset of fertility decline, and an overestimation of life expectancy at birth (e_0) over the first two centuries (c.a. $+42.2\%$). However, after adding selected collateral kin, the accuracy of the estimates improves: TFR is underestimated by only -0.11% during the first century and e_0 is overestimated by only $+1.5\%$ over the whole period.

Keywords: genealogies, microsimulation, biases, historical demography, kinship

Introduction

Long-term analysis of demographic dynamics, especially considering generational and kinship relationships is usually challenging and data demanding. Questions involving inter- (i.e., between) and multi- (i.e., many) generational perspectives often require data on vital events and kinship networks spanning decades or centuries. For instance, examining the familial transmission of demographic outcomes such as longevity requires long historical data series including kinship information, that allow to consider the life-time of multiple generations. These data requirements can limit the scope of inter- and multi-generational studies to specific time periods or geographic areas for which such data sources exist. However, unique opportunities for population research have recently emerged thanks to the availability of novel data sources driven by the Data Revolution (Alburez-Gutierrez et al., 2019; Kashyap, 2021), coupled with the increasing use of computationally-intensive tools such as microsimulation (Zagheni, 2015). Novel data sources, including those resulting from digitization and crowd-sourcing of historical records, can provide opportunities to study long-term dynamics, whose analysis has often been limited by the lack of (good) historical data. Thus, comprehending the potential and constraints of available sources, as well as appropriate methods for their use, can help broaden the scope of research in the fields of historical and kinship demography.

Genealogies hold great promise for this type of analysis, as they could enable us to link human populations over time, sometimes across space, and across generations. However, as they often suffer from problems of coverage and representativeness (Dupaquier, 1993), a deep understanding of their characteristics, quality problems, and biases is essential for informed use in demographic studies. Missing information issues include dates of birth and death and the omission of women, children who died at an early age and people who brought dishonor to the family (Hollingsworth, 1976; Zhao, 2001). Moreover, genealogies are usually records of surviving patrilineal lineages, which often experience better demographic conditions and show a higher sex ratios than the population as a whole. Hence, extinct and matrilineal lineages are often omitted from them (Zhao, 2006). In addition to demographic selectivity, which can lead to underestimation of mortality and overestimation of fertility, individuals with high socioeconomic status are more likely to be included in genealogies (Campbell and Lee, 2002).

Large online genealogical databases have recently emerged through the collaborative efforts of users of genealogical sites, such as Family Search, Geni and WikiTree (Charpentier and Gallic, 2020). These databases have been used to analyze patterns in mortality (Gavrilova and Gavrilov, 2007; Kaplanis et al., 2018; Minardi et al., 2023), morbidity (Rawlik et al., 2019) and fertility (Blanc, 2022a,b; Hsu et al., 2021). Nevertheless, besides the problems of genealogies mentioned above, online databases are non-representative samples of real-world family structures (Chong et al., 2022; Stelter and Alburez-Gutierrez, 2022) and further analysis is needed to improve the accuracy and reliability of the measures derived from these databases.

Genealogies are traditionally divided into ascendant and descendant (Bideau and Poulain, 1984; Jette and Charbonneau, 1984; Oeppen, 1999). In both cases, the starting point is an individual so-called 'ego', but while the former traces their ancestors backwards in time, the latter records their family trees prospectively. Both types of genealogies may or may not include collateral kin (i.e., those relatives who share a common ancestor but are not in a direct line), and have both advantages and limitations. Descendant genealogies have been widely used in historical demography, because they can be used directly in the family reconstitution method. However, they are often limited in number and size, and restricted to small areas with available parish or population registers (Bideau and Poulain, 1984; Jette and Charbonneau, 1984; Dupaquier, 1993). Here, we focus on ascendant genealogies which, despite the biases inherent in their nature, are more likely to be found outside the limited number of countries with high-quality records including kinship ties and have also become more available through online genealogical databases.

Demographic microsimulation has proven useful for investigating long-term kinship patterns (Murphy, 2011) as well as for evaluating historical data and assessing the reliability and bias of genealogies (Oeppen, 1999; Zhao, 1994, 2001, 2006) and family reconstitutions (Ruggles, 1992). Despite some constraints of microsimulation models, such as limitations when considering demographic similarities within the same kin group (Ruggles, 1993), or the dependence of the demographic events (and their timing) on the assumptions and input parameters (Zhao, 2006), they remain a powerful tool for analyzing the effects of selection and under-representation issues in genealogies. For instance, Zhao (2001) showed that the male lineages recorded in Chinese genealogies appear to be quite similar to the surviving patrilineages identified in the outputs of a microsimulation, giving an idea of the patterns of selection in genealogies and the differences with the demographic rates calculated for the whole population.

Since the possibility to infer demographic dynamics from genealogical data is affected by their nature and representativeness, it is essential to assess the size and effect of their biases before drawing conclusions for the general population. In this research, we conduct a series of experiments on synthetic populations, simulated using the SOCSIM demographic microsimulation program (Hammel et al., 1976), and taking Sweden (1751-2022) as a case study. We adopt the perspective of a group of genealogists to replicate the construction of ascending family trees and then evaluate the effect of some typical sources of bias in such trees on demographic measures. More specifically, we aim to understand **how these sources of bias affect the accuracy of fertility and mortality estimates derived from ascendant genealogies**. Our analysis seeks to contribute to a better understanding of the possibilities and limitations of using genealogies for demographic research.

Data and Methods

Demographic microsimulation

We run demographic microsimulations for Sweden (1751-2022) using the SOCSIM microsimulation program to obtain 'fully-recorded' synthetic populations, and thus information on vital events and kinship relationships for every single simulated individual who was ever alive. SOCSIM is an open-source demographic microsimulation program, originally developed at the University of California Berkeley (Hammel et al., 1976), and written in C programming language. It has been used for decades in demographic research to address issues such as kin availability and kin loss (Murphy, 2004, 2011; Verdery and Margolis, 2017; Zagheni, 2011), among others. The microsimulator takes as input an initial population file (with information on each individual's sex and date of birth) and monthly age-specific fertility rates and age-specific probabilities of death that hold over a given period for individuals of a particular sex, group, and marital status (married, single, divorced, widowed). Fertility rates can be parity-specific, but are not in this study. During the simulation, SOCSIM schedules and executes vital demographic events (births, marriages and deaths) for each 'living' simulated individual in the initial population and their descendants.

A brief description of how the microsimulator works is given in Mason (2016) and summarized below. At the beginning of each simulation segment (i.e., when the demographic rates or societal constants change) or month, SOCSIM schedules an event for each living individual to be executed at a future date. Only one event can be scheduled for each individual at any one time. After a person's event has been executed (except in the case of a death) or a change in their marital status or parity, a new event is scheduled for that person. To determine the next event to be scheduled for each individual, SOCSIM generates a random waiting time for each event for which each individual is at risk, considering the sex, age, group, and marital status specific rates. Once all potential events have randomly generated waiting times, the event with the shortest waiting time is selected and scheduled. Hence, the event competition follows a competing risk framework, wherein the probability of experiencing each event for which the individual of a given sex, age, marital status is at risk is independent of all others. All the events scheduled for a given month are executed in random order. SOCSIM then increments the month and repeats the event execution. At the end of the simulation, SOCSIM writes an output population file containing information about each individual who has ever lived and a marriage file containing information about each marriage generated during the simulation.

We run simulations from within R using the 'rsocsim' R-package (Theile et al., 2023) and input rates from the Human Fertility Collection (HFC) (1751-1890), the Human Fertility Database (HFD) (1891-2022) and the Human Mortality Database (HMD) (1751-

2022). The last two are retrieved via the 'HMDHFDplus' R-package (Riffe, 2015). To minimize the effects of microsimulation stochasticity without significantly compromising computational time, we run 10 simulations with the same initial population and input rates but different randomly generated seeds. This allows us to perform the experiments, that are explained in the next subsection, on more than one synthetic population and then average the results. As done in previous studies using SOCSIM, we first run the simulator for 100 years using the age-specific rates for 1751 to produce a stable age structure. This results in populations of about 15,000 individuals in 1751, which are then subjected to the corresponding annual rates for 1751-2022, resulting in synthetic populations of about 100,000 living individuals in 2022. Due to the lack of accurate age-specific marriage rates by sex for the entire period, we use the directive 'marriage after childbirth' in 'rsocsim' to create a marriage event and select a living unmarried spouse whenever a previously unmarried female gives birth. Following (Alburez-Gutierrez et al., 2021), spouses for each woman are chosen from all living single men to minimize the squared difference between the observed distribution of 'groom's age - bride's age' and a normal distribution with a mean of two and a standard deviation of three.

To assess the accuracy of our microsimulations, we estimate period Age-Specific Fertility Rates (ASFR) and Age-Specific Mortality Rates (ASMR), and their corresponding summary measures, Total Fertility Rate (TFR) and life expectancy at birth (e_0), based on the 10 SOCSIM outputs to verify that they are close to the input rates and derived measures. Figure A1 in the Appendix compares the estimates of ASFR, ASMR, TFR and e_0 derived from the simulation inputs (i.e., HFC/ HFD and HMD) and outputs. As expected in a stochastic process, there is still some variation around the reference value (input rates), especially when fertility rates are higher (distant periods) and mortality rates are lower (recent periods, for infant and child mortality). This can be explained by the fact that the initial populations are smaller compared to the final populations, after the simulated populations have grown. Nevertheless, the gaps are reduced when the summary measures (TFR and e_0) and the average of the 10 simulations for each measure are calculated. We chose to run 10 simulations as that represents an appropriate compromise between the computational time needed to run simulations and the level of stochasticity that remains after averaging across simulations.

Experiments on synthetic populations

We carry out a series of experiments on synthetic populations to assess the potential effect on demographic measures of some typical sources of bias in ascendant genealogies. Tracing family trees backwards, as is done in ascendant genealogies, relies mostly on lineage survival, i.e., the descendants of given ancestors must have survived to the time of genealogical reconstruction. This is a structural feature of ascendant genealogies. Therefore, throughout the experiments, we adopt the perspective of a group of hypothetical genealogists to analyze three main sources of bias: 1) the **selection on direct lineages**, 2) the **incomplete reconstruction of family trees**, and 3) the **missing information on some subpopulations**.

To evaluate the three sources of bias, we replicate the process of reconstructing the family trees of a group of individuals alive at the end of each simulation: our population of hypothetical genealogists. From each simulation output, we randomly select a sample of 10% of individuals aged 18 or older who are alive at the end of 2022. These individuals, hereafter called the ‘genealogists’, are the starting point for reconstructing the family trees for all of the experiments. We then merge the family trees of all the genealogists from each simulation, to obtain the ‘genealogical subsets’ from each experiment, i.e., those that replicate each source of bias.

We assess the size and effect of the three sources of bias, by comparing common demographic measures estimated from the ‘fully-recorded’ synthetic population, used here as a benchmark, and the ‘bias-infused’ genealogical subsets. As measures of period fertility and mortality, we include: ASFR, ASMR, TFR and e_0. We compute these demographic measures based on the ‘bias-infused’ genealogical subsets and compare the estimates with those derived from the whole 'fully-recorded' population. We follow the same approach for all sources of bias, which are explained in more detail below; see Table 1 for a summary of the experiments. To minimize the effects of microsimulation stochasticity, we run all the experiments over each of the 10 simulations, calculate the output demographic measures from both the 'bias-infused' and the whole 'fully-recorded' populations and then average the results. As a summary measure of the bias, for each simulation, we calculate the absolute difference between the genealogical subsets and the whole simulation, and then average the results to obtain the absolute and relative means of the differences.

The first source of bias arises from the fact that **tracing only direct lineages involves selection**, since direct ancestors (i.e., those related only through parent-child relationships) must have reproduced, and thus the childless are excluded by definition. In the first experiment, we trace all direct ancestors up to the 9^{th} generation (e.g., parents, grandparents, great-grandparents, ..., 6x great-grandparents) of the genealogists. Since more than one descendant of a lineage may have survived to the time of genealogical reconstruction (in our case, the end of 2022), some common ancestors may be included in more than one genealogy, leading to duplicates when the family trees of multiple genealogists are merged. Since duplicates are a common problem in genealogical data, we compute demographic measures using the subset of only direct ancestors both with and without duplicates and compare them with estimates derived from the whole simulated population, to evaluate the effect not only of demographic selection but also of duplicates.

The second source of bias is related to the fact that **family trees reconstructed by genealogists are often incomplete** due to limited knowledge of relatives or the choice of whom to include. Hence, the extent and complexity of the kinship network considered in genealogies may vary across individuals and societies over time. In the first experiment, we limited the genealogical reconstruction to the (up to 510) direct ancestors of the genealogists. However, some individuals who are not in their direct ancestral line, but are related through collateral kinship relationships (i.e., those who descend from a common ancestor but are not in a direct blood line, such as siblings or aunts/uncles, etc.) may be omitted from a family tree. We also analyze the effect of including in the genealogies **some types of collateral kin** in addition to **direct ancestors**. Starting from the genealogists, we trace their own siblings, aunts/uncles and first cousins, as well as the siblings of their direct ancestors up to the 8^{th} generation (i.e., great-aunts/uncles, (...), 6x-great-aunts/uncles) (see Table 1 for the selected kin types). We gradually add one more kin type to the genealogical subset from each simulation, but for readability only present the results with all selected kin types in the main text. In this experiment, we remove the duplicates created after merging the family trees of multiple genealogists. We compute the defined demographic measures from the subsets including the collateral kin and compare them to the estimates derived from the whole simulated population.

The third source of bias is related to **missing information on some subpopulations** who may be forgotten or omitted when reconstructing a family tree, such as early deceased children and unmarried/childless women. This could lead to underrepresentation of these subpopulations in genealogies compared to their actual proportion in a given population. We examine the effect on demographic estimation of omitting a percentage of a) early deceased children and b) childless women from the most complete genealogical subset of Experiment 2, i.e., including all direct ancestors and collateral kin, hereafter referred to as the 'extended genealogy', (see Table 1 for the types of kin included). We follow a similar approach for both subpopulations. For children (Experiment 3A), we randomly remove, from the extended genealogy, a proportion of those who died before the age of 5, allowing for 25%, 50%, 75%, and 100% omissions over the entire period (1751-2022). For childless women (Experiment 3B), which are the same as unmarried women in our simulation setup, we randomly remove, from the extended genealogy, 25%, 50%, 75%, and 100% of all women who survived to at least reproductive ages (15) and had no children. We remove the duplicates from the trees of multiple genealogists, compute the demographic measures based on the subsets with omitted subpopulations and compare them to the estimates derived from both the whole simulation and the extended genealogy.

Results

Through the series of experiments described above, we evaluate the potential effect of three sources of bias in ascendant genealogies on measures of period fertility and mortality. We run all the experiments independently for each of the 10 simulations, calculate the demographic measures for each subset from each simulation, and then average the results. For readability, we present here the mean measures derived either from the whole simulated populations or from the genealogical subsets created for each of the experiments.

Experiment 1. Selection on direct lineages

In the first experiment, we evaluate the bias of selection in direct lineages by comparing the genealogical subsets of direct ancestors reconstructed up to the 9^{th} generation with the whole simulated populations. For women in Sweden, Figure 1 compares the age-specific fertility and mortality rates in 1900-1905, taken as an example from the middle of the period, (panels a and b), and the evolution of the summary measures (TFR and e_0) over the whole period (panels c and d) from the different subsets. Figures comparing the 1900-1905 age-specific estimates with earlier (1800-1805) and more recent years (2000-2005) are provided in Appendix A2.

Regarding age-specific fertility rates, panel a of Figure 1 suggests that the estimates from the genealogical subsets of direct ancestors (lines with squares and dots) are lower than those from the whole simulation (lines without shapes). Since these subsets include only the direct ancestors of the genealogists, collateral kin and especially those without descent are underrepresented in this type of genealogy. Thus, mothers appear to have had fewer children than they actually gave birth to, since the ancestors' siblings are not included in the tree of a given genealogist. In addition, the estimates for all ages from the genealogical subset with duplicates (lines with squares) are lower than those from the subset without duplicates (lines with dots). This is more remarkable in the most distant periods (see Appendix A2). The number of possible duplicates is likely to increase as we go back in time, since more distant ancestors are likely to be duplicated in more genealogies than more recent ancestors. Thus, although the birth of a given ancestor may be counted in more than one genealogy - creating duplicates in the numerator - female ancestors of reproductive age in the same year of birth of such an ancestor may be overcounted more times, resulting in a greater number of duplicates in the denominator. The earliest birth dates of females from earlier generations could increase the likelihood of being included in multiple genealogies, which could explain the greatest underestimation of fertility in the genealogical subset with duplicates, especially as we go back in time. In terms of timing, the age distribution is close to that from the whole simulation, yet fertility peaks at lower ages (25-30) in the genealogical subsets with duplicates.

Looking at the fertility summary measure (i.e., TFR), panel c of Figure 1 shows a potential underestimation of fertility from genealogies of only direct ancestors (blue line with squares and olive line with dots) throughout most of the period. Estimates from the genealogical subset with duplicates are only slightly higher (1 to 10%) than the whole simulation for a few years in the 1930s, when fertility already reached a very low level. However, the difference between the two subsets of direct ancestors and the whole simulation changes over time, with the overall gap being larger before the 20^{th} century, when fertility was at its highest level (a fluctuating TFR but always above 4 children per women). During that period, the TFR is underestimated by between 1.85 and 3.96 children per woman in the genealogical subset with duplicates and by between 1.23 and 2.53 in the genealogical subset without duplicates. Before the fertility decline, including only direct ancestors seems to have a larger effect on the summary measure (TFR): on average, -70.5% and -39% of the whole simulation value in the genealogical subsets with and without duplicates, respectively. When considering high fertility periods, the likely underreporting of the actual number of births in the genealogies of direct ancestors, due to the exclusion of collateral kin and without-descent ancestors, seems to be more pronounced as the real number of births per woman was also higher. Otherwise, births to mothers belonging to lineages that have not survived to the present will not be included in this or subsequent experiments, which could also lead to an underestimation of fertility in ascendant genealogies.

As for age-specific mortality, the estimates derived from both genealogical subsets of direct ancestors, with and without duplicates, are relatively close to each other (lines with squares and dots in panel b of Figure 1). Thus, the inclusion of duplicates in the data does not seem to have much effect on the age distribution of deaths. In addition, estimates from genealogies of direct ancestors are very close to the estimates for the whole simulation after age 35. This pattern also holds for earlier and more recent years (see Appendix A2), although the age at which the gap begins to close appears to decrease over time. This suggests that, for adult and old age, the shape of the mortality curve is not strongly biased by deriving the measures from direct ascendant genealogies. Nevertheless, infant, child and young adult mortality cannot be accurately estimated from these genealogical subsets, since there are no deaths before age 15 and those between ages 15 and 35 are underestimated. In the last century, there are no deaths before age 20. This can be explained by the fact that direct ancestors must have survived to reproductive age to be the ancestors of some of the genealogists.

Considering now the effect of omitting early deaths on the summary measure of mortality (i.e., e_0), panel d of Figure 1 suggests a clear overestimation of life expectancy at birth until the middle of the last century (here, 1948), ranging for females from almost no bias (0.6 and 0.33 years) to 46.75 and 44.75 years higher in 1773 for the subsets with and without duplicates, respectively. This corresponds on average to +44.5% and +42.2% with respect to the e_0 estimated from the whole simulation in the genealogical subsets with and without duplicates, respectively. However, it shows a slight underestimation from then on, reaching -2.64 and -2.48 years in 1992 for the subsets with and without duplicates, with respect to the estimates from the whole simulation. This change in the direction of the bias from overestimation to underestimation may be related to the fact that the burden of infant and child mortality - which is not captured in direct ascendant genealogies - had a greater positive impact on life expectancy estimates in past centuries than in more recent periods, when improvements in mortality are mostly associated with old-age mortality.

Experiment 2. Incomplete reconstruction of family trees

In our second experiment, we examine the bias resulting from the incomplete reconstruction of family trees, by examining the inclusion of selected collateral kin in a genealogical subset of direct ancestors. Although we include some relatives from the same generation as the genealogists', such as siblings and cousins, the majority of the family tree belongs to past generations. Therefore, demographic events corresponding to more recent years are still partially covered by this extension of an ascendant genealogy. For the sake of readability, we compare here the estimates from the genealogical subsets of only direct ancestors, direct ancestors with all selected collateral kin, and the whole simulated population. For women in Sweden, Figure 2 compares age-specific fertility and mortality rates in 1900-1905, taken as a mid-point example, and the evolution of the summary measures (TFR and e_0) over the whole period. Figures comparing the estimates derived from the subsets that gradually add one kin type (e.g., direct ancestors plus siblings or direct ancestors plus siblings and aunts/uncles) are provided in Appendix A3 and A4.

With respect to age-specific fertility, panel a of Figure 2 shows that the estimates from the genealogical subsets with only direct ancestors and those together with collateral kin (lines with dots and diamonds) are lower than the estimates from the whole simulation (lines without shapes). However, after including all selected types of collateral kin (see Table 1 for details), the estimates become closer to those derived from the whole simulation, both in terms of level and timing. The gap appears to narrow as we include more collateral ancestors and go back in time (see Appendix A3). The remaining underestimation of fertility at all ages, especially for the most recent years, may be due to the types of kin included in an ascendant genealogy. Here only the births of the genealogists and their siblings or first cousins and, for earlier generations, of the genealogists' direct ancestors and their siblings, but not of the latter descendants, are included in the genealogical subset.

This trend could also be observed over time, see panel c of Figure 2. Estimates of TFR based on genealogies including collateral kin (purple line with diamonds) are now closer to the estimates from the whole simulation during the first century of analysis (i.e., 5 to 8 generations backward), implying only an underestimation of -0.11 children per woman (or a reduction of -0.11% in TFR). After that, the gap with the whole simulation starts to grow progressively until it reaches -1.43 children per woman in 2022 (or a reduction of -0.92%) compared to the whole simulation. TFR from genealogies that include selected collateral kin are likely to be more accurate for the earliest periods, since the number of direct ancestors increases with each generation backward, and the number of their siblings (x-great-uncles) is also likely to increase due to high fertility levels in past centuries. However, the underestimation of fertility progressively increases in recent years, because the types of relatives that contribute to the numerator of the fertility rates (collateral and descendant kin from recent generations, such as children, nieces and nephews) cannot be captured under this scope of an ascendant genealogy. The differences in the accuracy of the estimates as an additional type of kin is progressively included are illustrated in Figure A4 in the Appendix.

Regarding age-specific mortality, the estimates for adult and old-age mortality from the genealogical subsets with only direct ancestors and those together with collateral kin (lines with dots and diamonds) are quite similar, see panel b of Figure 2. This also holds for earlier and more recent periods (see Figure A3 in the Appendix). However, the estimates from the genealogical subset that includes collateral kin are now much closer in terms of level and timing to the estimates from the whole simulation, even for infant and young age mortality which are nonexistent in the subset of direct ancestors. Although the estimates are slightly lower at early ages, the inclusion of collateral kin in the genealogical reconstruction improves the estimation of early deaths.

The inclusion of collateral kin also affects the summary measure of mortality (e_0). As shown in panel d of Figure 2, estimates of life expectancy at birth from the genealogical subset with selected collateral kin, become very close to those derived from the whole simulation over the entire period, although the former are still slightly higher, resulting in an overall overestimation of 0.7 years or +1.5%. Therefore, the accuracy of demographic estimates based on genealogies improves significantly after the inclusion of all selected collateral kin, as each generation backward provides progressively more information about the demographic events of each period when the ancestors' siblings are included. A comparison of the estimates that progressively include an additional type of relative from each generation backward is illustrated in Figure A4 in the Appendix.

Experiment 3. Missing information on some subpopulations

In our third experiment, we examine the bias associated with missing information on two types of relatives who are likely to be omitted from genealogies: children who died at an early age and childless women. We analyze the effects of omitting both subpopulations separately using a similar approach, but do not consider their combination. Figures 3 and 4 compare, for women in Sweden, age-specific fertility and mortality rates in 1900-1905, taken as mid-point example, and the evolution of the summary measures (TFR and e_0) over the whole period, derived after omitting different proportions of children who died before age 5 or childless women, respectively. For readability, we show only the estimates with 25% and 100% of omission. We compare the estimates with omission to those derived from both the whole simulation (used as a benchmark) and the 'extended genealogy' with all direct ancestors and collateral kin, which is still slightly biased as explained in the second experiment. For both subpopulations, removing information biases the estimates in the same direction, but the magnitude is significantly larger when omitting early deceased children. For the latter, we also explore using a threshold of age 1, but the results (not included here) show very similar patterns to those obtained using the threshold of age 5, except for the age-specific mortality rates below the age of 1 or 5. Figures comparing the age-specific estimates from 1900-1905 with earlier (1800-1805) and more recent years (2000-2005) for each subpopulation are provided in Appendix A5 and A6.

On the one hand, Figure 3 compares the estimates derived from omitting children who died before age 5. Regarding fertility, panel a of Figure 3 shows that the age-specific rates from the genealogical subsets with omitted early deceased children are lower than those from the extended genealogy and the whole simulation (lines with diamonds and no shapes), with the bias increasing as the percentage of omission increases, and that the age distribution is almost unaffected. This is true during periods of high fertility and mortality, although the gap with the extended genealogy is almost imperceptible when infant and child mortality is very low (more recent years).

Looking at the changes in the summary measure over time, as shown in panel c of Figure 3, the further back in time, the greater the effect of omitting early deceased children on the underestimation of fertility, which increases proportionally with to the percentage of omission, especially before the 20^th century, when both fertility and under-5 mortality were high in Sweden. Before that, a 25% omission of early-deceased children leads on average to an underestimation of -0.41 children per woman (or a reduction of -9.2% in the TFR), while a 100% of omission leads on average to an underestimation of -1.25 children per woman (or a reduction of -27.76% in the TFR). Since under-five mortality was particularly high in earlier centuries, and in the case of Sweden began to decline from the 18^th century onward, while going back in time, a greater number of children ever born would be missing if those who died at an early age were omitted from the genealogies. From the 20^th century on, the gap to the extended genealogy begins to close, being minimal in the last decades.

For mortality, as expected, the age-specific estimates are lower only for ages 0-1 and 1-5, being non-existent with 100% omission (see panel b of figure 3). This also holds for earlier and more recent years (see Appendix A5). Nevertheless, the omission of these early deceased children actually has a large effect on the overestimation of life expectancy at birth (e_0) (see panel d of Figure 3). Again, the effect of omitting children who died early on the overestimation of e_0 increases significantly going back in time, and also increases with the proportion of omission. A 25% omission of early-deceased children can lead to an overestimation of e_0 by up to 4.74 years (i.e., +20.52%), while a 100% omission to an overestimation of e_0 by up to 21.5 years (i.e., +95.5%).

On the other hand, Figure 4 compares the estimates obtained by omitting childless women. For fertility, the effect of the omission on age-specific rates (panel a of Figure 4) is quite similar to that of early deceased children, with estimates lower than those from the extended genealogy and the whole simulation in past centuries, though the gap with the extended genealogy is almost nonexistent in most recent years. Again, the age distribution is unaffected.

However, the effect on the fertility summary measure over time shows a slightly different trend than for the early-deceased children (see panel c of figure 4). Compared to the extended genealogy, omitting a larger proportion of childless women (lighter colors) led to a slightly larger underestimation of fertility until the 1910s (when the TFR was above 3 children per women) and a slightly larger overestimation for some decades afterwards. For recent decades, the omission has almost no effect, but as discussed above, fertility estimates are already lower in those years, due to the the types of relatives included in the extended genealogy. However, the estimates with omitted childless women are always lower than the whole simulation, leading on average to -0.33 children per woman with 25% of omission (or a reduction of -14% in the TFR) and -0.34 children per woman with 100% of omission (or a reduction of -13.8% in the TFR).

As for mortality, the age-specific estimates are lower than those from the whole simulation and the extended genealogy only for ages 15-35 (see panel b of figure 4). This holds not only for 1900-1905, but also for previous and more recent centuries (see Appendix A6). Nevertheless, the omission of childless women has relatively little effect on the overestimation of life expectancy at birth (e_0), which increases slightly as larger proportions of childless women are omitted (see panel d of figure 4). Over the whole period, a 25% of omission of childless women leads to an overestimation of e_0 by up to 2.15 years (i.e., +6.79%), while a 100% omission to an overestimation of e_0 by up to 4.2 years (i.e., +9.57%).

Discussion

Genealogies are promising sources for research on historical and kinship demography. However, these data have not been leveraged to their full potential as we have not fully understood the biases that affect their representativeness. Here we reported on a series of experiments on synthetic populations aimed at understanding how three main sources of bias in ascendant genealogies can affect the accuracy of demographic estimates. Using the SOCSIM demographic microsimulation program and taking Sweden (1751-2022) as a case study, we generate fully recorded synthetic populations that allow us to conduct some experiments on ascendant genealogies. We then adopt the perspective of a group of hypothetical genealogists, trace their family trees and extract different genealogical subsets that reproduce the three sources of bias. Hence, based on lineages that survived to the present, we explore three sources of bias in genealogies: the selection in direct lineages leading to the exclusion of the childless, the incomplete reconstruction of family trees involving the inclusion/exclusion of given types of kin, and missing information on some subpopulations, such as early deceased children and unmarried/childless women, who are often underrepresented in genealogies.

Our analysis highlights three key points. On the one hand, besides the exclusion of extinct lineages in the genealogies of survivors, the extent and completeness of the family tree, approximated here by the types of direct and collateral kin included, seems to affect the accuracy of demographic estimates based on ascendant genealogies. Their accuracy improves as more types of relatives from distant generations are included in the family tree. On the other hand, such effects do not appear to be linear or unidirectional, as their magnitude varies over time, particularly between the periods characterized by high and low fertility and mortality levels. Finally, the omission of subpopulations who are normally underrepresented in genealogies seems to follow a similar pattern of variation, although it is more pronounced for children. Such trends may suggest the existence of a potential relationship between changes in the size and effect of the selected biases in genealogies over time and changes in the quantum and timing of fertility and mortality resulting from the demographic transition process. However, such an observation demands further investigation.

According to previous research, genealogical data are biased toward higher fertility and lower mortality. We found lower mortality in the bias-infused genealogies, leading to an overestimation of life expectancy at birth, especially in experiments 1 and 3, which included only direct ancestors (c.a. +42.2% until the mid-20^th century) or omitted subpopulations. However, we did not find the expected higher fertility in any of the three experiments. The estimates derived from our genealogical subsets are generally lower than those derived from the whole simulation, especially in Experiment 1 which considers only direct ancestors and thus likely underestimates the number of births a woman might have given (c.a. −39% during the first 150 years, i.e., before the onset of the fertility decline). After including all selected collateral kin in Experiment 2 (i.e., the siblings of direct ancestors), the gap narrows considerably, especially for periods of high fertility (c.a. −0.11% during the first century). This also true for life expectancy at birth, which is overestimated by only +1.5% over the whole period in Experiment 2. Both changes in the estimates suggest that the expansion of the kinship network in the family tree is essential for the accuracy of demographic estimates based on genealogies.

The results presented here have limitations that we would like to acknowledge. First, our analysis is based on synthetic populations, which are not the same as real populations and therefore cannot reproduce the whole complexity of their dynamics and structures. For instance, we do not consider familial transmission of fertility and mortality behavior, as the input data are only disaggregated by sex and age.[1] Thus, beyond the individual stochasticity resulting from the microsimulation, there is no predefined clustering of families with better or worse demographic conditions. Second, in our research design, genealogists are randomly selected from the individuals aged 18 and older at the end of the simulation. Therefore, based on our synthetic populations, we can only reproduce the selection resulting from the survival of some lineages to the time of genealogical reconstruction, but not that resulting from other factors such as better demographic or socioeconomic conditions. Third, our experiments are based on an ideal scenario of building extensive and fully recorded family trees, where genealogists can track demographic information for all ancestors and collateral kin 8 generations backwards without any restrictions. However, it may be more difficult for actual genealogists to obtain complete and reliable information for distant generations, as data for the earliest periods are more likely to be imprecise, incomplete, unavailable, or more available for larger or wealthier families in the past. Therefore, apart from testing the omission of early deceased children and childless women (Experiment 3), we do not assess the bias resulting from imprecise or incomplete data in genealogies, which may also limit the scope of our results. Fourth, the definition and implementation of an ascendant genealogy in this research may also affect the results, as demographic information may be incomplete due to our choice of kin types, especially for the most recent periods for which descendants, such as children,

[1] In the current version of rsocsim, a heterogeneous fertility option can be enabled to allow for its heritability through the maternal line. However, the default option leads to a significant underestimation of fertility, especially for the periods when it was at high levels, which would require significant calibration of the input rates.

nieces, nephews or grandchildren, could provide some information. We also exclude affinal and in-law relatives so as to limit the genealogical reconstruction to consanguinity. Finally, real-world genealogies may be affected by one or more sources of bias simultaneously, and such a potential combination of biases may also vary over time and across generations. However, we consider it is important to analyze the sources of bias one at a time, before adding the complexity of variations in bias over time and across generations.

Our study of fully-recorded synthetic genealogies provided important insights for researchers using genealogical datasets for historical demographic research. We showed that deriving demographic estimates from direct lineages exclusively produces unrealistic results. Including collateral kin to these ancestors-only genealogies (especially siblings of direct ancestors) significantly improved the accuracy of our estimates, particularly for remote periods. This shows that the completeness of family trees within ascendant genealogies is crucial. Researchers working with this data should always compare genealogy-based estimates with those obtained from other traditional sources to get an idea of the magnitude of the biases in the data and their direction. The underestimation of fertility may suggest the exclusion of some types of collateral kin in a genealogy, particularly for high-fertility periods. A significant overestimation of life expectancy at birth in past centuries is likely to suggest an underestimation of infant and child mortality, though estimates can become more accurate if they are conditional upon survival to age 5, 10, etc. Finally, ascendant genealogies can hardly account for contemporary demographic events. For this reason, researchers should carefully consider the period for which there is sufficient high-quality data in the genealogies.

For future studies, we identify three main lines of research. First, studies can examine the extent to which changes in the size and effect of the sources of bias are related to changes in fertility and mortality levels resulting from the process of the demographic transition. This could include comparisons with countries experiencing different patterns. Second, studies can replicate our analysis focusing on other demographic measures, such as parity distributions by cohort and survival thresholds at ages other than 0. Third, studies can evaluate the size and effect of the sources of bias by considering a cohort perspective.

Acknowledgments

We thank Jim Oeppen for relevant exchanges during the early stages of the research. We thank Martin Kolk and David Hacker for useful comments on a preliminary version of the manuscript, presented at the 2023 Population Association of America Annual Meeting. We thank Tom Theile for assistance with programming, particularly with rsocsim. The paper was discussed at the Work-in-Progress Workshop of the Department of Digital and Computational Demography at the MPIDR, where it benefited from several useful comments. Liliana Calderón-Bernal gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS).

Data availability statement

The code to retrieve the data, run the microsimulations and reproduce the results is available online: https://github.com/liliana-calderon/SOCSIM_Genealogies

References

Alburez-Gutierrez, D., Aref, S., Gil-Clavel, S., Grow, A., Negraia, D. V., and Zagheni, E. (2019). Demography in the Digital Era: New Data Sources for Population Research. In *Proceedings of the 2019 Conference of the Italian Statistical Society*.

Alburez-Gutierrez, D., Mason, C., and Zagheni, E. (2021). The “Sandwich Generation” Revisited: Global Demographic Drivers of Care Time Demands. *Population and Development Review*, 47(4):997–1023.

Bideau, A. and Poulain, M. (1984). De la généalogie à la démographie historique : généalogie ascendante et analyse démographique. *Annales de démographie historique*, 1984(1):55–69.

Blanc, G. (2022a). The Cultural Origins of the Demographic Transition in France. Job Market Paper 2.

Blanc, G. (2022b). Demographic Change and Development from Crowdsourced Genealogies in Early Modern Europe. ⟨hal-02922398v2⟩.

Campbell, C. and Lee, J. (2002). State Views and Local Views of Population: Linking and Comparing Genealogies and Household Registers in Liaoning, 1749–1909. *History and Computing*, 14(1-2):9–29.

Charpentier, A. and Gallic, E. (2020). Can Historical Demography Benefit from the Collaborative Data of Genealogy Websites? *Population*, 75(2):379–408.

Chong, M., Alburez-Gutierrez, D., Del Fava, E., Alexander, M., and Zagheni, E. (2022). Identifying and correcting bias in big crowd-sourced online genealogies. Technical Report WP-2022-005, Max Planck Institute for Demographic Research, Rostock. Edition: 0.

Dupaquier, J. (1993). Généalogie et démographie historique. *Annales de démographie historique*, pages 391–395. Publisher: Editions Belin.

Gavrilova, N. S. and Gavrilov, L. A. (2007). Search for Predictors of Exceptional Human Longevity: Using Computerized Genealogies and Internet Resources for Human Longevity Studies. *North American Actuarial Journal*, 11(1):49–67.

Hammel, E. A., Hutchinson, D. W., Wachter, K. W., Lundy, R. T., and Deuel, R. Z. (1976). *The SOCSIM demographic-sociological microsimulation program: operating manual*. Number 27 in Research series. Institute of International Studies. University of California, Berkeley. OCLC: 2704303.

Hollingsworth, T. H. (1976). Genealogy and historical demography. *Annales de démographie historique*, 1976(1):167–170.

Hsu, C.-H., Posegga, O., Fischbach, K., and Engelhardt, H. (2021). Examining the trade-offs between human fertility and longevity over three centuries using crowdsourced genealogy data. *PLOS ONE*, 16(8):e0255528. Publisher: Public Library of Science.

Human Fertility Collection (2023). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.fertilitydata.org.

Human Fertility Database (2023). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.humanfertility.org.

Human Mortality Database. HMD (2023). Max Planck Institute for Demographic Research (Germany) and University of California, Berkeley (USA) and French Institute for Demographic Studies (France) Available at www.mortality.org.

Jette, R. and Charbonneau, H. (1984). Généalogies descendantes et analyse démographique. *Annales de démographie historique*, 1984(1):45–54.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., and Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175. Publisher: American Association for the Advancement of Science Section: Research Article.

Kashyap, R. (2021). Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. *Population Studies*, 75(sup1):47–75. Publisher: Routledge _eprint: https://doi.org/10.1080/00324728.2021.1969031.

Mason, C. (2016). Socsim oversimplified. berkeley: Demography lab, university of california.

Minardi, S., Corti, G., and Barban, N. (2023). Historical Patterns in the Intergenerational Transmission of Lifespan and Longevity: Evidence from the United States, 1700-1900.

Murphy, M. (2004). Tracing very long-term kinship networks using SOCSIM. *Demographic Research*, 10:171–196.

Murphy, M. (2011). Long-Term Effects of the Demographic Transition on Family and Kinship Networks in Britain. *Population and Development Review*, 37:55–80.

Oeppen, J. (1999). Genealogies as a source for demographic studies: some estimates of bias. In *Workshop on Genes, Genealogies and Longevity, Rostock*, Rostock.

Rawlik, K., Canela-Xandri, O., and Tenesa, A. (2019). Indirect assortative mating for human disease and longevity. *Heredity*, 123(2):106–116. Number: 2 Publisher: Nature Publishing Group.

Riffe, T. (2015). Reading human fertility database and human mortality database data into r. *Rostock: Max Planck Institute for Demographic Research (MPIDR Technical Report TR-2015-004)*.

Ruggles, S. (1992). Migration, Marriage, and Mortality: Correcting Sources of Bias in English Family Reconstitutions. *Population Studies*, 46(3):507–522.

Ruggles, S. (1993). Confessions of a Microsimulator: Problems in Modeling the Demography of Kinship. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 26(4):161–169. Publisher: Taylor & Francis Group.

Stelter, R. and Alburez-Gutierrez, D. (2022). Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 119(10):e2120455119. Supplementary material https://osf.io/9gkmz/.

Theile, T., Alburez-Gutierrez, D., Calderón-Bernal, L. P., Snyder, M., and Zagheni, E. (2023). *rsocsim: SOCSIM with R*. https://github.com/MPIDR/rsocsim, https://mpidr.github.io/rsocsim/.

Verdery, A. M. and Margolis, R. (2017). Projections of white and black older adults without living kin in the United States, 2015 to 2060. *Proceedings of the National Academy of Sciences*, 114(42):11109–11114.

Zagheni, E. (2011). The Impact of the HIV/AIDS Epidemic on Kinship Resources for Orphans in Zimbabwe. *Population and Development Review*, 37(4):761–783. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1728-4457.2011.00456.x.

Zagheni, E. (2015). Microsimulation in Demographic Research. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Elsevier, Oxford.

Zhao, Z. (1994). Demographic Conditions and Multi-generation Households in Chinese History. Results from Genealogical Research and Microsimulation. *Population Studies*, 48(3):413–425. Publisher: Routledge _eprint: https://doi.org/10.1080/0032472031000147946.

Zhao, Z. (2001). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, 55(2):181–193.

Zhao, Z. (2006). Computer microsimulation and historical study of social structure: A comparative review of SOCSIM and CAMSIM. *Revista de Demografia Historica*, XXIV(II):59–88.

Tables

Table 1 Experiments to assess the effect of three sources of bias in ascendant genealogies: genealogical subsets and kin types included in the family tree

Experiment	1	2	3
Source of bias	Selection in direct lineages	Incomplete reconstruction of family trees	Missing information on some subpopulations
Effect on genealogies	Exclusion of the childless	Inclusion or exclusion of collateral kin	Under-representation of subpopulations
Population of genealogists	10% sample of individuals aged 18+ alive by 31.12.2022		
Genealogical subsets:			
Direct ancestors	All	All	All
Collateral kin and nuclear	No	All	All
Omission of children dead before age 1 or 5	No	No	25%, 50%, 75%, 100% removed
Omission of childless women	No	No	25%, 50%, 75%, 100% removed
Kin types in genealogies			
Genealogist (ego)	All	All	All
Parents	All	All	All
Grandparents	All	All	All
Great-grandparents	All	All	All
2x-Great-grandparents	All	All	All
3x-Great-grandparents	All	All	All
4x-Great-grandparents	All	All	All
5x-Great-grandparents	All	All	All
6x-Great-grandparents	All	All	All
Siblings	No	Gradually/All	All
Aunts/uncles	No	Gradually/All	All
First cousins	No	Gradually/All	All
Great-aunts/uncles	No	Gradually/All	All
2x-Great-aunts/uncles	No	Gradually/All	All
3x-Great-aunts/uncles	No	Gradually/All	All
4x-Great-aunts/uncles	No	Gradually/All	All
5x-Great-aunts/uncles	No	Gradually/All	All
6x-Great-aunts/uncles	No	Gradually/All	All

Figures

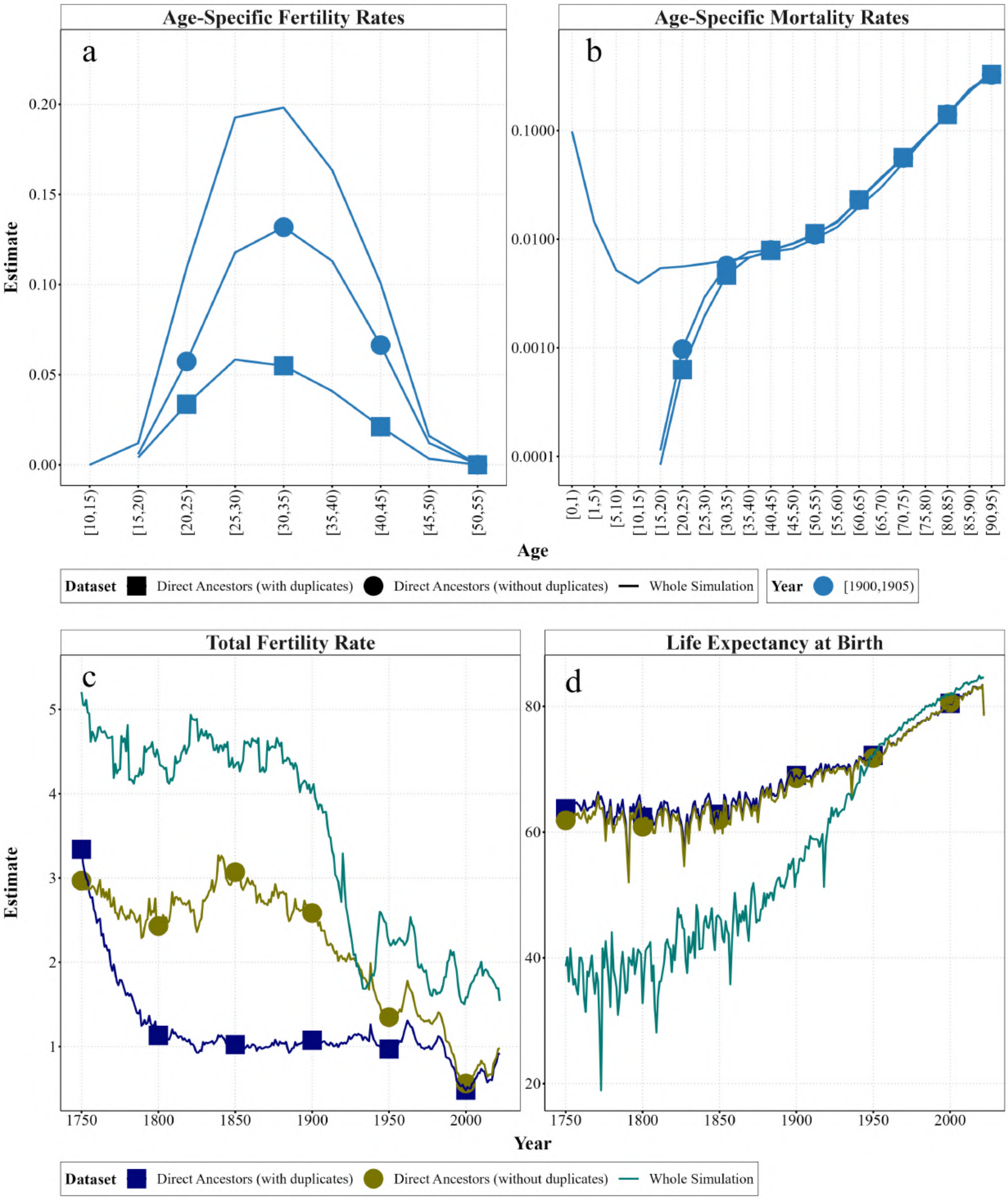

Figure 1: Experiment 1: Age-specific and summary demographic measures derived from genealogical subsets of only direct ancestors (with and without duplicates) versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset. At young (< 15) and very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown.

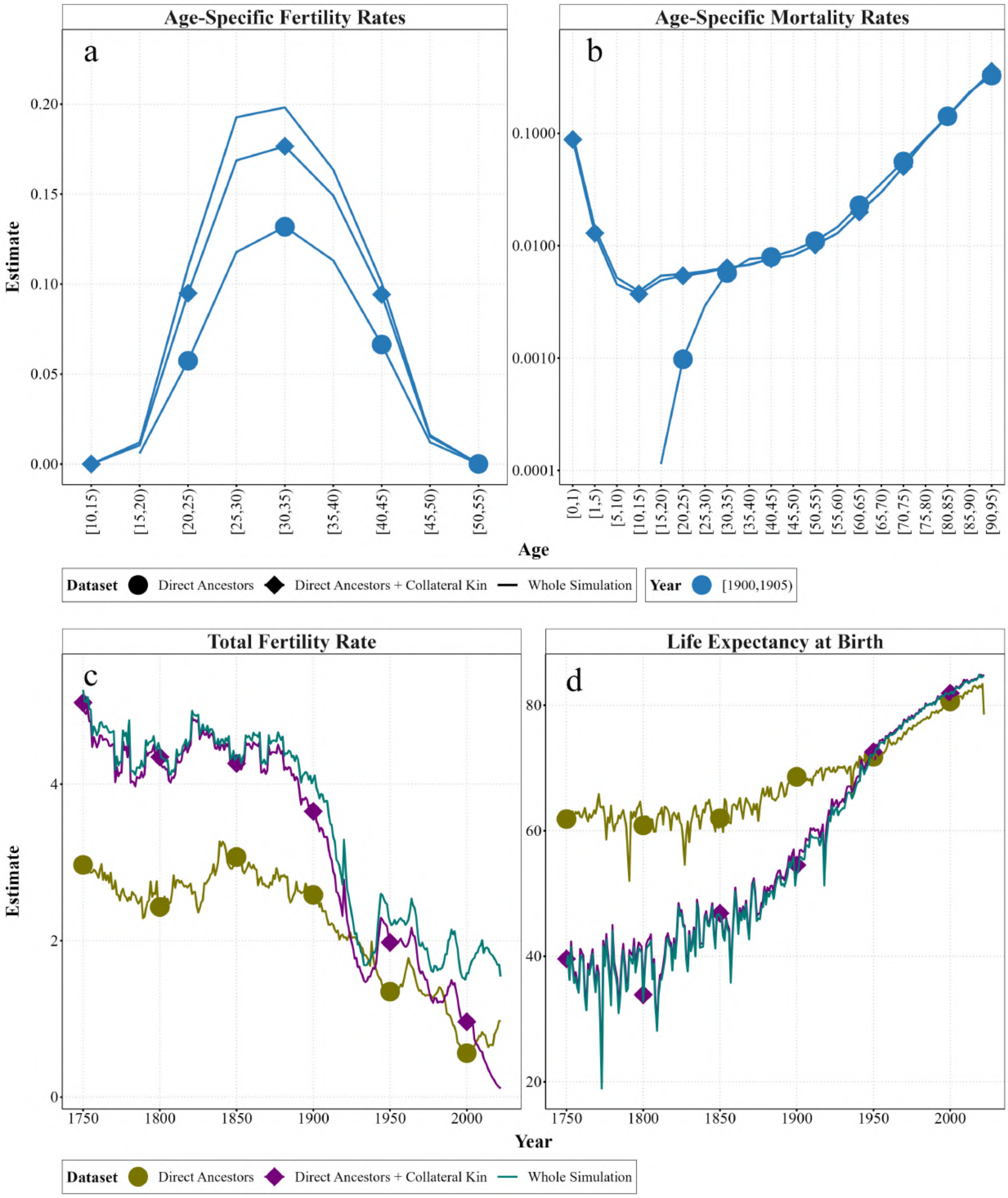

Figure 2: Experiment 2: Age-specific and summary demographic measures derived from genealogical subsets of only direct ancestors and together with all selected collateral kin versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset. At young (< 15) and very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown.

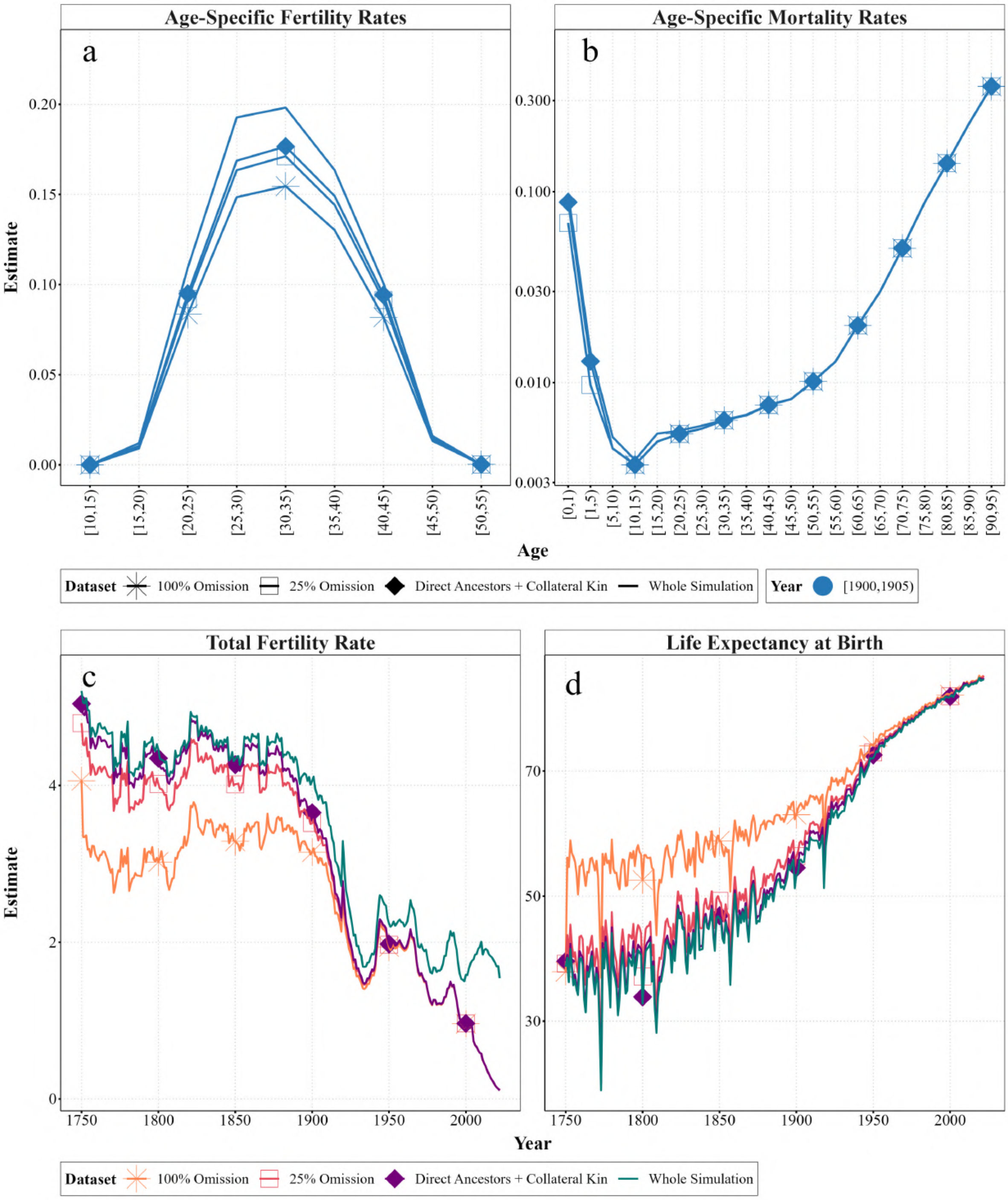

Figure 3: Experiment 3A: Age-specific and summary demographic measures derived from genealogical subsets omitting different proportions of children who died before the age of 5 versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset. At young (< 10) and very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown. For the dataset with 100% omission all mortality rates below age 10 are 0.

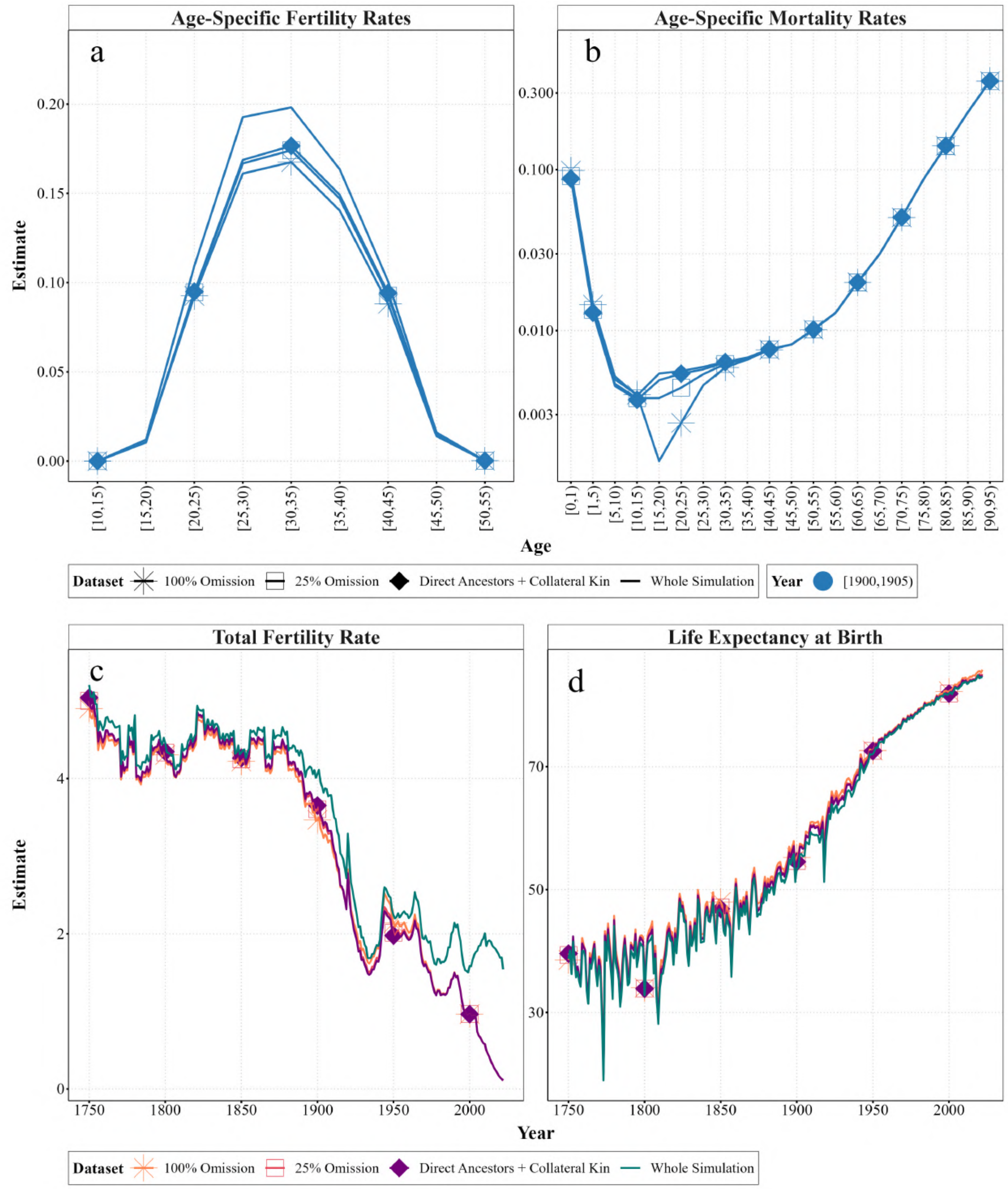

Figure 4: Experiment 3B: Age-specific and summary demographic measures derived from genealogical subsets omitting different proportions of childless women versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset. At very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown.

Appendix

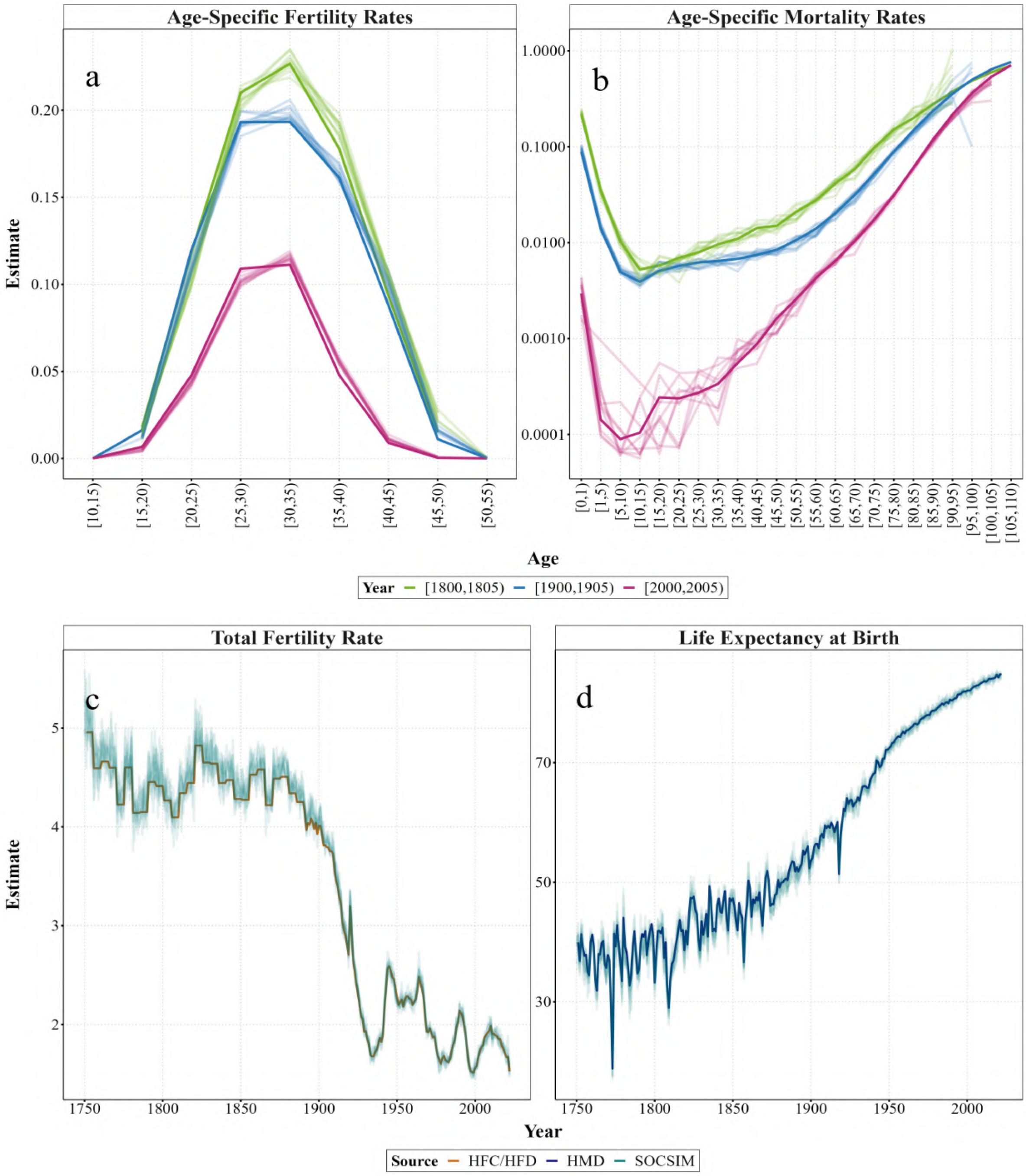

Figure A1 : Age-specific and summary demographic measures, retrieved from the Human Fertility Collection (HFC), the Human Fertility Database (HFD), the Human Mortality Database (HMD), and 10 SOCSIM outputs. The bold lines correspond to HFC/HFD and HMD estimates and the transparent lines to the 10 SOCSIM simulations. At young and very old ages, mortality rates from the simulations can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown.

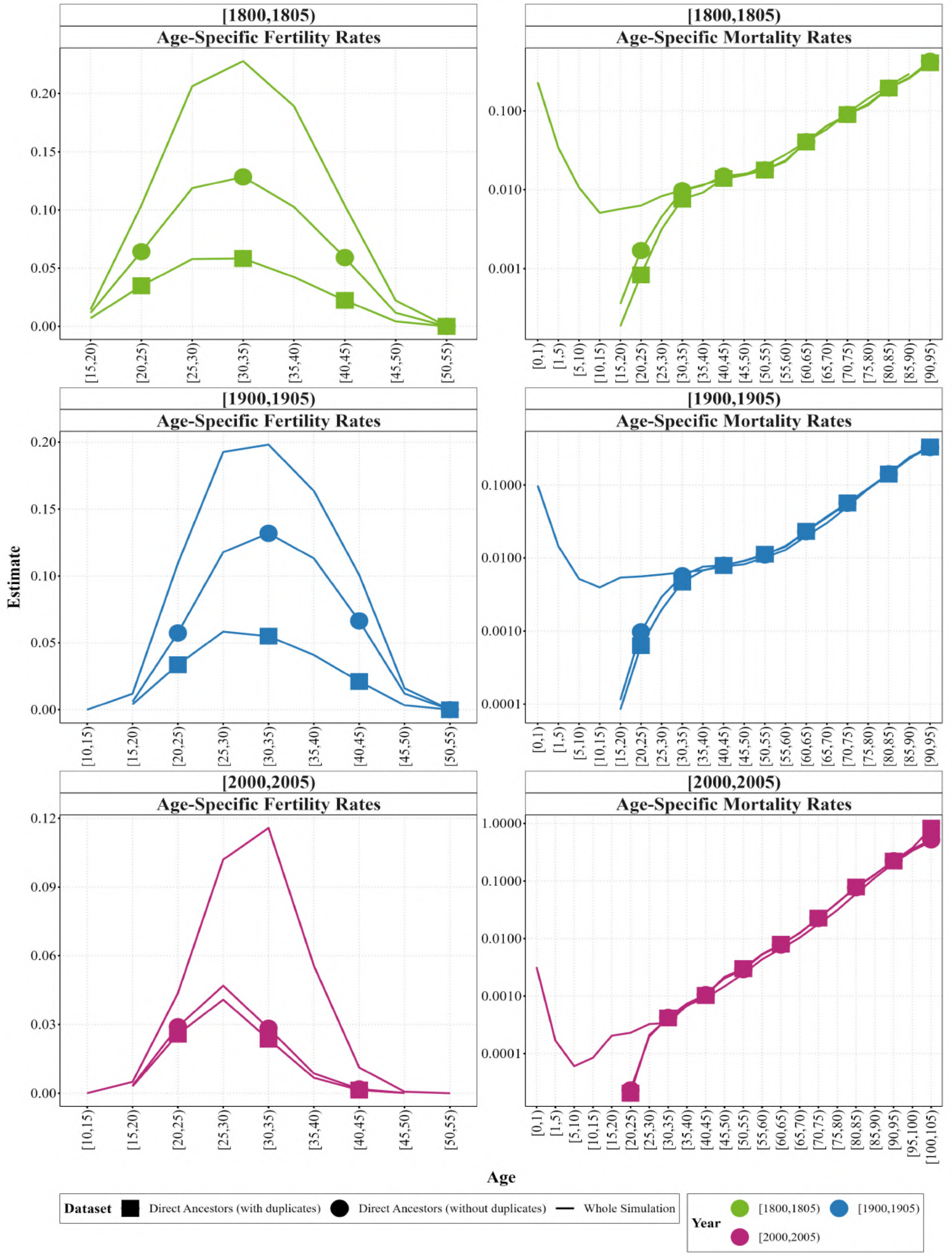

Figure A2 : Experiment 1: Age-specific demographic measures derived from genealogical subsets with only direct ancestors (with and without duplicates) versus the whole SOCSIM simulations in three selected periods. In each panel, the figure represents the means for each dataset. At young (< 20) and very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Therefore, these values are not shown.

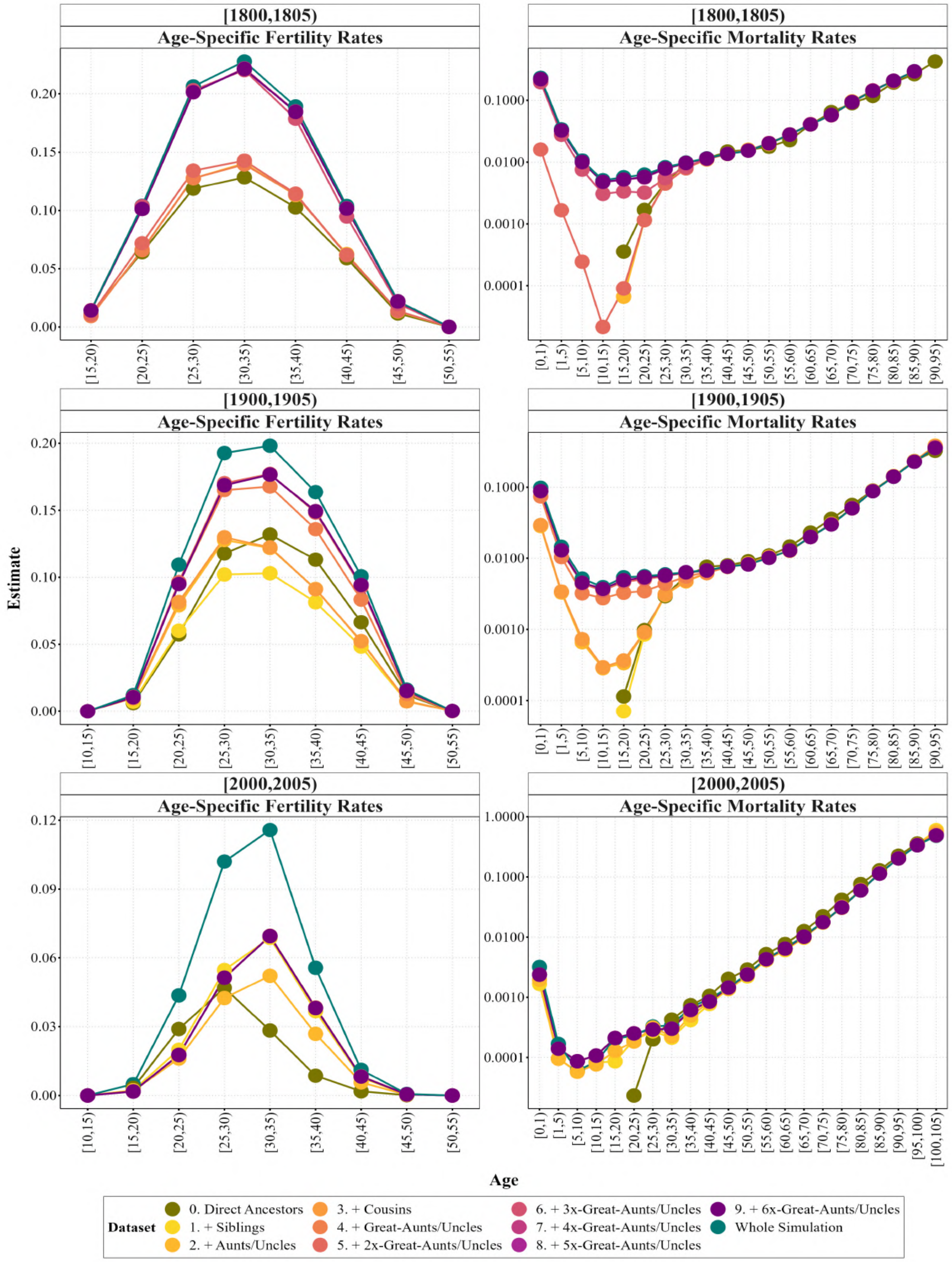

Figure A3 : Experiment 2: Age-specific demographic measures derived from genealogical subsets with only direct ancestors and with progressive inclusion of collateral kin versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset. At young (< 20) and very old (> 95) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown. For the dataset with 100% omission of early-deceased children all rates below age 10 are 0.

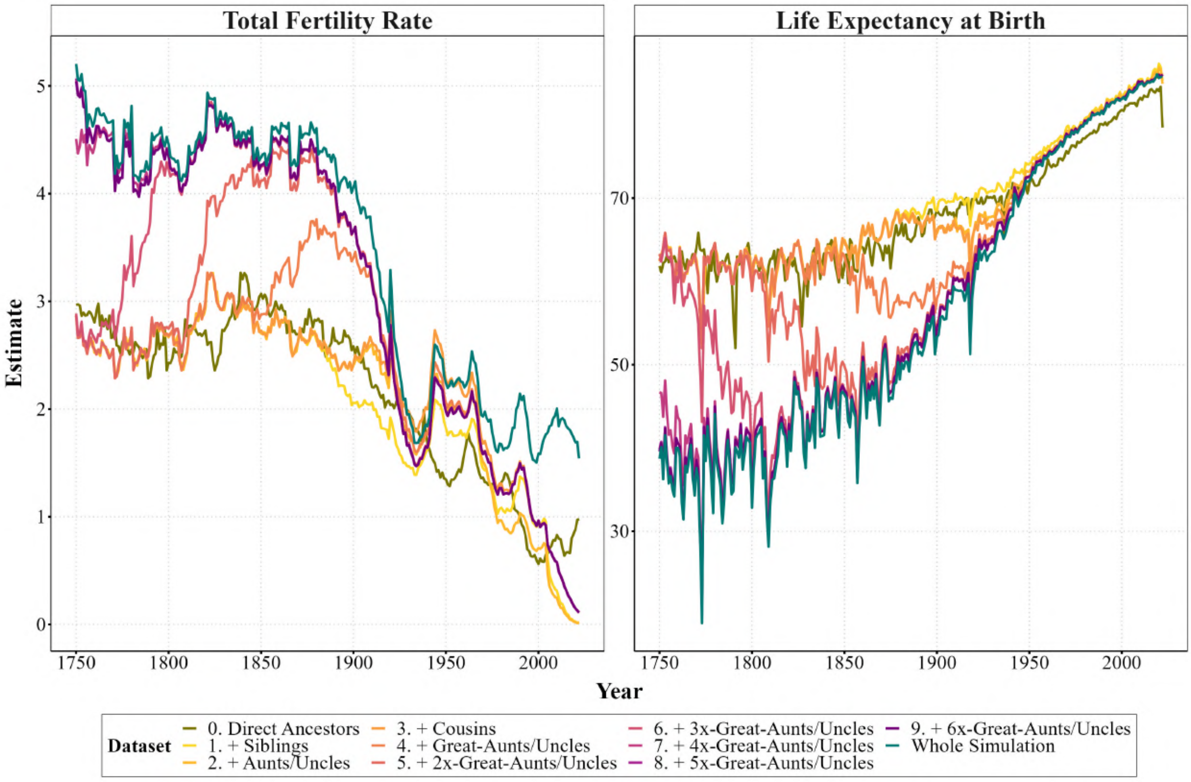

Figure A4 : Experiment 2: Summary demographic measures derived from genealogical subsets with only direct ancestors and with progressive inclusion of collateral kin versus the whole SOCSIM simulations. In each panel, the figure represents the means for each dataset.

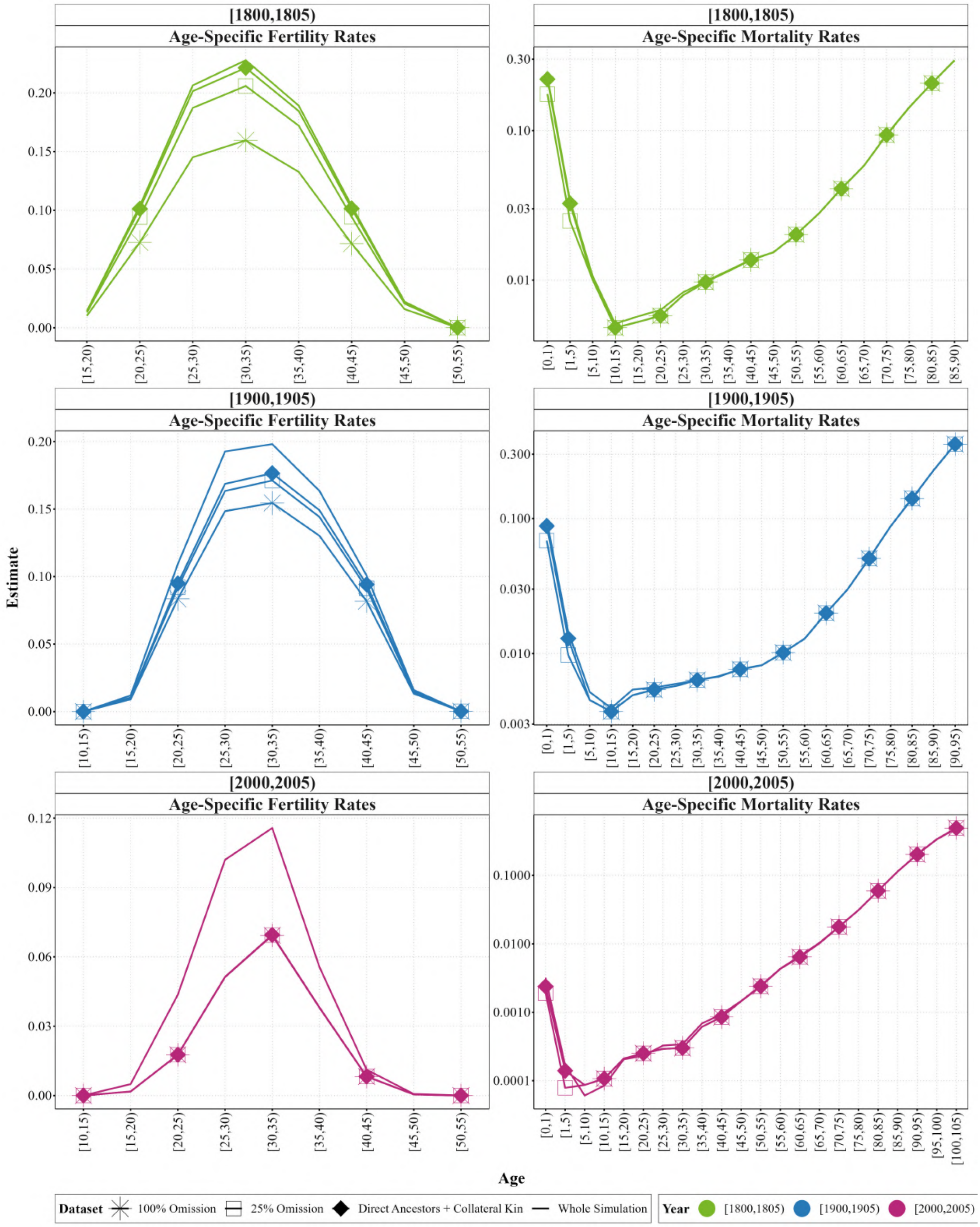

Figure A5 : Experiment 3A: Age-specific demographic measures derived from genealogical subsets omitting different proportions of children who died before the age of 5 versus the whole SOCSIM simulations in three selected periods. In each panel, the figure represents the means for each dataset. At young (< 10) and very old (> 90) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown. For the dataset with 100% omission all mortality rates below age 10 are 0.

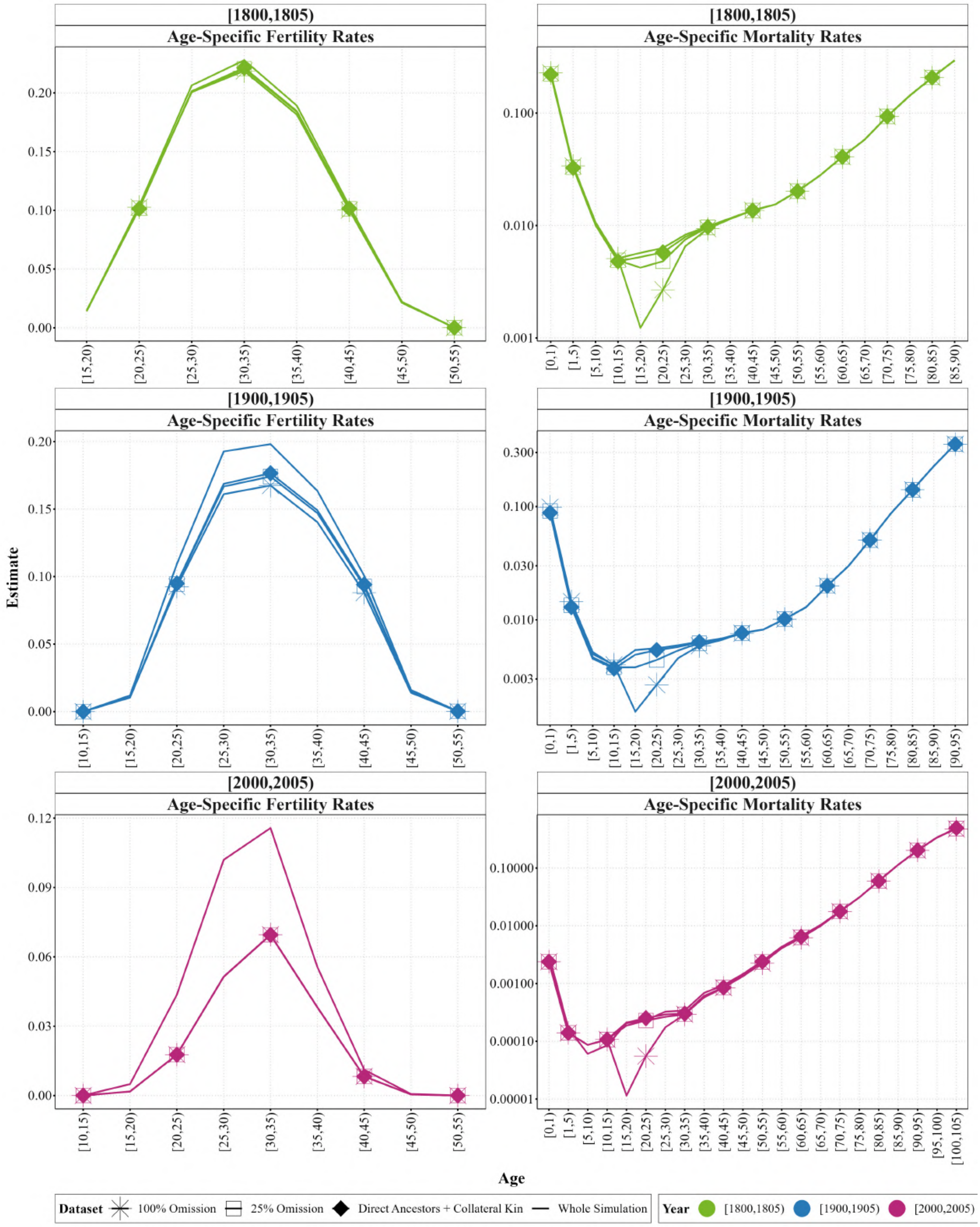

Figure A6 : Experiment 3B: Age-specific demographic measures derived from genealogical subsets omitting different proportions of childless women versus the whole SOCSIM simulations in three selected periods. In each panel, the figure represents the means for each dataset. At very old (> 90) ages, mortality rates from the experiment can be 0, which introduces infinite values into the log scale used in the figure, as well as infinite (x/0) or NaN (0/0). Hence, these values are not shown.