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 I August 2023
Revised January 2025
https://doi.org/10.4054/MPIDR-WP-2023-034

Analysing biases in genealogies using demographic microsimulation

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

Analysing 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

January 6, 2025

Abstract

An incomplete understanding of biases affecting the representativeness of genealogies has hindered their full exploitation. We report on a series of experiments on synthetic populations designed to assess how different biases in ascendant genealogies can affect the accuracy of demographic estimates. Using the SOCSIM microsimulation programme and Swedish fertility and mortality data (1751-2022), we analyse three sources of bias: selection in direct lineages, incomplete reconstruction of family trees, and missing information on subpopulations. Comparing demographic measures derived from 'fully-recorded' and 'bias-infused' synthetic populations, we find that including only direct ancestors leads to an underestimation of total fertility rate (TFR) (c.a. -42%) and an overestimation of life expectancy at birth (e_0) (c.a. $+33\%$) over the whole period. However, after including direct ancestors' offspring, TFR became overestimated (c.a. $+21\%$) while e_0 overestimation was limited to $+1.8\%$. Our study shows that the completeness of family trees is essential for obtaining accurate demographic estimates.

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 allows to consider the lifetime of multiple generations. These data requirements can limit the scope of inter- and multi-generational studies to specific 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 digitisation 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 and appropriate methods for their use can help broaden the scope of research in historical and kinship demography.

Genealogies hold great promise for this type of analysis, as they could enable us to link human populations over time, across space and 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 dishonour to the family (Hollingsworth, 1976; Zhao,

2001). Moreover, genealogies are usually records of surviving patrilineal lineages, which often experience better demographic conditions and show 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 underestimating mortality and overestimating fertility, individuals with high socioeconomic status are more likely to be included in genealogies (Campbell and Lee, 2002).

Large online genealogical databases have recently become more available through the collaborative efforts of users of genealogical sites, such as Family Search, Geni and WikiTree (Charpentier and Gallic, 2020). These have been used to analyse patterns in mortality (Gavrilova and Gavrilov, 2007; Kaplanis et al., 2018; Minardi et al., 2024), morbidity (Rawlik et al., 2019) and fertility (Blanc, 2022a,b; Hsu et al., 2021). However, 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; Colasurdo and Omenti, 2024). Previous research has explored some of the problems in both offline and online genealogies, but further analysis is needed to improve the accuracy and reliability of the measures derived from these databases. This research seeks to understand the potential and limitations of using genealogical data for conducting demographic research by shedding light on the size and effect of some of their biases.

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 lateral kin i.e., those relatives who share a common ancestor but are not in a direct line. For historical demography, descen-

dant genealogies have often been considered more appropriate, as they can be used directly with the family reconstitution method (Bideau and Poulain, 1984; Jette and Charbonneau, 1984; Dupaquier, 1993), and are potentially unbiased when they are complete (Oeppen, 2021). However, they are often limited in number and size and restricted to specific areas with available parish records, such as the ones used by Henry (1956); Hollingsworth (1964, 1977), population registers in Belgium or Sweden, or the genealogical records for China. In this research, 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. This type of genealogy has also become more accessible 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 analysing the effects of selection and under-representation issues in genealogies. For instance, Zhao (2001) used microsimulation to compare the demographic conditions of the members of simulated male surviving patrilineages — considered to be very similar to the male lineages recorded in Chinese genealogies — with those of the members of all simulated lineages, taken as a representative sample of the entire population. The results showed that, for the first four or five simulated generations, male members of the surviving lineages had a higher average age at death, a higher proportion married, a higher average number of children and a higher proportion with sons, than males of all simulated lineages. This provided

insights into the effects of selectivity in descent-type genealogies such as the Chinese, especially linked to their patrilineal nature and survival bias.

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 inherent biases before drawing conclusions about the general population. In this research, we examine the biases inherent in the process of genealogical reconstruction, particularly in ascendant genealogies, without focusing on any specific database. We conducted a series of experiments on synthetic populations, simulated using the SOCSIM demographic microsimulation programme (Hammel et al., 1976), and taking Sweden as a study case, which allowed us to cover several generations thanks to the availability of long-established vital statistics. Based on synthetic populations with 'fully-recorded' information, we replicated the construction of ascending family trees for a group of hypothetical genealogists to 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 ran demographic microsimulations over four centuries using the SOCSIM microsimulation programme and Swedish data (1751–2022) to obtain 'fully-recorded' synthetic populations, i.e., the register of every single individual who was ever alive during the simulation, including the information on their vital events and kinship relationships. SOCSIM is an open-source demographic microsimulation programme,

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, because they are only available after 1970. 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 summarised 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. The event competition thus follows a competing risks framework, where the probability of experiencing each event for which the individual of a given sex, age and marital status is at risk is independent of all others. All 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 ran 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 were retrieved via the 'HMDHFDplus' R-package (Riffe, 2015). To minimise the effects of microsimulation stochasticity without significantly compromising computational time, we ran ten simulations with the same initial population and input rates but different randomly generated seeds. This allowed 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 ran the simulator for 100 years using the age-specific rates for 1751 to produce a stable age structure. This resulted in populations of about 15,000 individuals in 1751, which were then subjected to the corresponding annual rates for 1751–2022. We let the simulator run for another 50 years with 2022 rates to avoid problems with the final populations. However, we conducted the experiments based on the synthetic populations alive at the end of 2022 (about 100,000 living individuals per simulation). Due to the lack of accurate age-specific marriage rates by sex for the entire period, we used 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 were chosen from all living single men to minimise 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 estimated period age-specific

fertility rates (ASFR) for women and age-specific mortality rates (ASMR) for both sexes, 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 5 in 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 than the final populations after the simulated populations have grown. Nevertheless, the gaps were reduced when the summary measures (TFR and e_0) and the average of the ten simulations for each measure were calculated. We chose to run ten 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 carried out a series of experiments on synthetic populations to assess the potential effect on demographic measures of some 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, on which the methodological design of this research was based. Therefore, throughout the experiments, we examined the biases that, in addition to lineage survival, may result from the retrospective reconstruction of family trees of given genealogists from the present. Based on the results of previous research, as well as some exploratory attempts to build our family trees, we defined three main sources of bias to investigate: 1) the *selection on direct lineages*, 2) the *incomplete recon-*

struction of family trees, and 3) the *missing information on some subpopulations*. Although the last two could also affect descendant genealogies, we focus here on the biases in ascendant genealogies.

To evaluate the three sources of bias, we replicated the process of reconstructing the family trees of a group of individuals alive at the end of each simulation, considered as our hypothetical genealogists. From each simulation output, we randomly selected a sample of 10% of individuals alive by the end of 2022. These individuals, hereafter called the 'genealogists', were the starting point to reconstruct individual family trees. We merged the family trees of all the genealogists selected from each simulation, to obtain the 'genealogical subsets' that replicated each source of bias.

We assessed 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 from each experiment. As measures of period fertility and mortality, we included ASFR, ASMR, TFR and e_0. We computed these demographic measures based on the 'bias-infused' genealogical subsets and compared the estimates with those derived from the whole 'fully-recorded' population. We followed the same approach for the three sources of bias, which are explained in more detail below, (see Table 1 for a summary of the experiments). To minimise the effects of microsimulation stochasticity, we ran all the experiments over each of the ten simulations, calculated the output demographic measures from both the 'bias-infused' and the whole 'fully-recorded' populations of each simulation, and then averaged the results. As a summary measure of the bias, for each simulation, we calculated the absolute difference between the genealogical subsets and the whole simulation, and then averaged 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 the childless are excluded by definition. In the first experiment, we traced all direct ancestors nine generations backwards (e.g., parents, grandparents, great-grandparents, ..., 7x great-grandparents) of each genealogist. Since more than one descendant of a lineage may have survived to the time of the genealogical reconstruction (in our case, the end of 2022), some genealogists may share a common ancestor, who may then be included in more than one genealogy. Therefore, duplicates may occur when combining the trees of multiple genealogists, which is a common problem when working with genealogical data, especially within online platforms that compile the trees of multiple genealogists. To assess the effect not only of demographic selection but also of duplicates, in this experiment, we computed demographic measures using the subset of only direct ancestors, both with and without duplicates, and compared them with estimates derived from the whole simulated population.

The second source of bias is related to the fact that *family trees reconstructed by genealogists are often incomplete* due to limited knowledge of all relatives or the choice of whom to include. Therefore, the extent and complexity of the kinship network considered in genealogies may vary between individuals or societies and over time, and some family trees may be incomplete. In the first experiment, we limited the genealogical reconstruction to the (up to 1022) 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 are the offspring 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. Therefore, in this experiment we analysed the effect of including in the genealogies not only the *direct ancestors* but also *their offspring*. Starting from the genealogists, we traced their

parents' offspring (i.e., their siblings), and the offspring of all their direct ancestors (i.e., aunts/uncles, great-aunts/uncles, (...), 7x-great-aunts/uncles) (see Table 1 for the included kin types). We gradually added one more kin type to the genealogical subset from each simulation, but for readability, we only present the results with all kin types in the main text. In this experiment, we removed the duplicates created after merging the family trees of multiple genealogists. We computed the defined demographic measures from the subsets including the direct ancestors' offspring and compared them to the estimates derived from the whole simulated population.

The third source of bias is due to *missing information on some subpopulations* who may be forgotten or omitted from a family tree, such as early deceased children and unmarried/childless women. This could lead to the underrepresentation of these subpopulations in genealogies compared to their actual share in a given population. We examined the effect on the 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 their offspring, hereafter referred to as the 'extended genealogy', (see Table 1 for the types of kin included). We followed a similar approach for both subpopulations. For children (Experiment 3A), we randomly removed, from the extended genealogy, a proportion of those who died before the age of five, 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 removed, 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 removed the duplicates from the trees of multiple genealogists, computed the demographic measures based on the subsets with omitted subpopulations and compared them to the estimates derived from both the whole simulation and the extended genealogy.

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	Exclusion of direct ancestors' offspring	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
Direct ancestors' offspring	No	All	All
Omission of children dead before age one or five	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
1x-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
7x-Great-grandparents	All	All	All
Siblings	No	Gradually/All	All
Aunts/uncles	No	Gradually/All	All
1x-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
7x-Great-aunts/uncles	No	Gradually/All	All

Results

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

Experiment 1. Selection on direct lineages

In the first experiment, we evaluated the bias of selection in direct lineages by comparing the genealogical subsets of direct ancestors reconstructed nine generations backwards with the whole simulated populations. For women in Sweden, Figure 1 compares the age-specific fertility and mortality rates in 1900–05, 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. A comparison of 1900–05 age-specific estimates with earlier (1800–05) and more recent years (2000–05) are provided in Figure 6 in Appendix

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, the direct ancestors' offspring and especially the childless are underrepresented in this type of genealogy. As the direct ancestors' children are not all included in the tree of a given genealogist, the ancestors' mothers may appear to have had fewer children than they gave birth to. For example, if a given woman had four children back in time but only

one is the direct ancestor of any genealogist, she would appear to have had only one child instead of four. However, some of the siblings of that direct ancestor may also belong to another tree as direct ancestors of another genealogist. Thus, this woman would be included in two (or more) family trees and two (or more) of her children would be counted for the fertility estimates. Yet it is unlikely that all of her four children are direct ancestors of any of the genealogists, as the latter were a selected sample of individuals who have survived to the present.

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 pronounced in the most distant periods (see Figure 6 in Appendix). The number of possible duplicates is likely to increase as we go back in time, since more distant ancestors are likely to be included in more family trees — and thus be common ancestors for more individuals — than more recent ancestors. For instance, a female direct ancestor may appear in two (or more) genealogies as the mother of two (or more) different children (i.e., siblings), who are direct ancestors of different genealogists. If duplicates are removed, the mother would appear only once in the denominator, but both children remain in the numerator as they are not the same individual. Therefore, the age-specific fertility rates are higher in the subset without duplicates because more duplicates are removed from the denominator than from the numerator. In terms of timing, the age distribution is close to that of the full simulation, but fertility peaks at lower ages (25-30) in the genealogical subsets.

The extent of this bias over time can be examined by looking at the evolution of the fertility summary measure (i.e., TFR). Panel c of Figure 1 shows the underestimation of fertility from genealogies of only direct ancestors (blue line with squares and olive line with dots) throughout the entire period. TFR from the genealogical subset with duplicates is always very close to one, as every female direct ancestor

has exactly one child who is a direct ancestor. Slight variations (above or below one) are probably due to approximations in the calculation of the TFR. Here, all births to women of a given age in a calendar year are divided by the female population of a given age at mid-year. Thus, the female population at risk is approximated by the female mid-year population.

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 twentieth century, when fertility was at its highest level (a fluctuating TFR always above 4 children per woman). Over the entire period, including only direct ancestors leads on average to -2.4 and -1.5 children per woman (or -66% and -42%) in the genealogical subsets with and without duplicates compared to the whole simulation. Before the fertility decline (approximated here by 1900), the effect was larger for the subset with duplicates resulting on average on -3.5 children per woman (or -76%). 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 all their offspring and without-descent ancestors, seems to be more pronounced as the real number of births per woman was also higher. Otherwise, due to the research design, births to mothers belonging to lineages that have not survived to the present were not included in this or subsequent experiments.

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 40. This pattern also holds for earlier and more recent years (see Figure 6). This suggests that, for adult and old ages,

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 40 are underestimated. 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. There are no deaths before the age of 35 in the early twenty-first century, probably because there are no early deaths among the parents of the genealogists.

Considering now the effect of omitting early deaths on the mortality summary measure (i.e., e_0), panel d of Figure 1 suggests an overestimation of life expectancy at birth over the whole period, ranging for females from almost no bias (0.6 and 0.7 years) to 34 and 33 years higher for the subsets with and without duplicates, respectively. This corresponds on average to +33% and +32% compared to the e_0 estimated from the whole simulation in the genealogical subsets with and without duplicates, respectively. The decrease in the size of the bias 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.

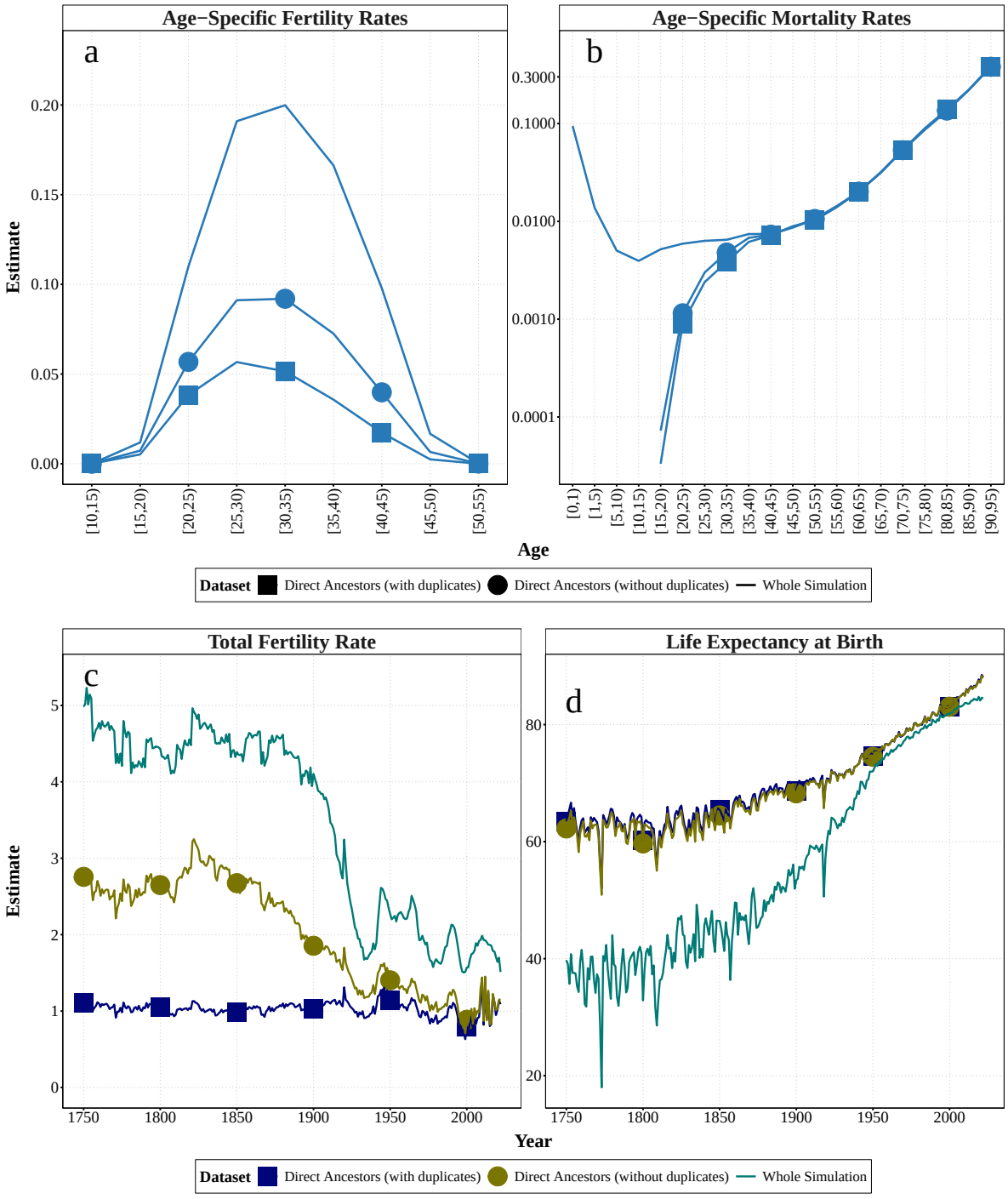

Fig 1: Experiment 1: Age-specific and summary demographic measures for women in Sweden derived from genealogical subsets of only direct ancestors (with and without duplicates) extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure represents the means 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). These values are not shown.*

Experiment 2. Incomplete reconstruction of family trees

In our second experiment, we examined the bias resulting from the incomplete reconstruction of family trees, by exploring the effect of adding all direct ancestors' offspring to the genealogies of direct ancestors. From this experiment onwards, we remove all the duplicates from the calculation. Although we included some relatives from the same generation as the genealogist (their siblings), the majority of the family tree belongs to previous generations. Therefore, demographic events corresponding to more recent years are only partially covered and the results for recent years might be taken with additional caution. For the sake of readability, we compare here the estimates from the genealogical subsets of only direct ancestors, direct ancestors with all their offspring, and the whole simulated population. For women in Sweden, Figure 2 compares age-specific fertility and mortality rates in 1900–05, taken as a mid-point example, and the evolution of the summary measures (TFR and e_0) over the whole period. A comparison of the estimates based on 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 Figure 7 and 8 in Appendix.

Concerning age-specific fertility, panel a of Figure 2 shows that when including the births of all direct ancestors' offspring (see Table 1 for details), the estimates from the genealogical subsets with direct ancestors and their offspring (lines with diamonds) became larger than the estimates from the whole simulation (lines without shapes) and the genealogical subset of only direct ancestors (lines with dots). As we go back in time, the overestimation of fertility is due to the inclusion of births of direct ancestors' offspring of more distant generations (see Figure 7 in Appendix.).

This trend could also be observed over time, see panel c of Figure 2. Estimates of TFR based on genealogies including direct ancestors' offspring (purple line with

diamonds) are larger than the estimates from the whole simulation (and the genealogies of only direct ancestors), implying an overestimation of +0.3 children per woman (or an increase in TFR of +6%) before the fertility decline and +0.5 children per woman (or +21%) over the whole period. Here, the bias is larger over the most recent periods. The differences in the accuracy of the estimates as an additional type of kin is progressively included are illustrated in Figure 8 in 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 their offspring (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 7). However, the estimates from the genealogical subset that includes direct ancestors' offspring 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 direct ancestors' offspring in the genealogical reconstruction improves the estimation of early deaths.

The addition of direct ancestors' offspring 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 direct ancestors' offspring, 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 0.9 years (or +1.8%). Therefore, the accuracy of demographic estimates based on genealogies improves significantly after the inclusion of all direct ancestors' offspring, as each generation backwards provides progressively more information about the demographic events of each period when the direct ancestors' offspring are included. Figure 8 in Appendix compares the estimates that progressively include an additional relative.

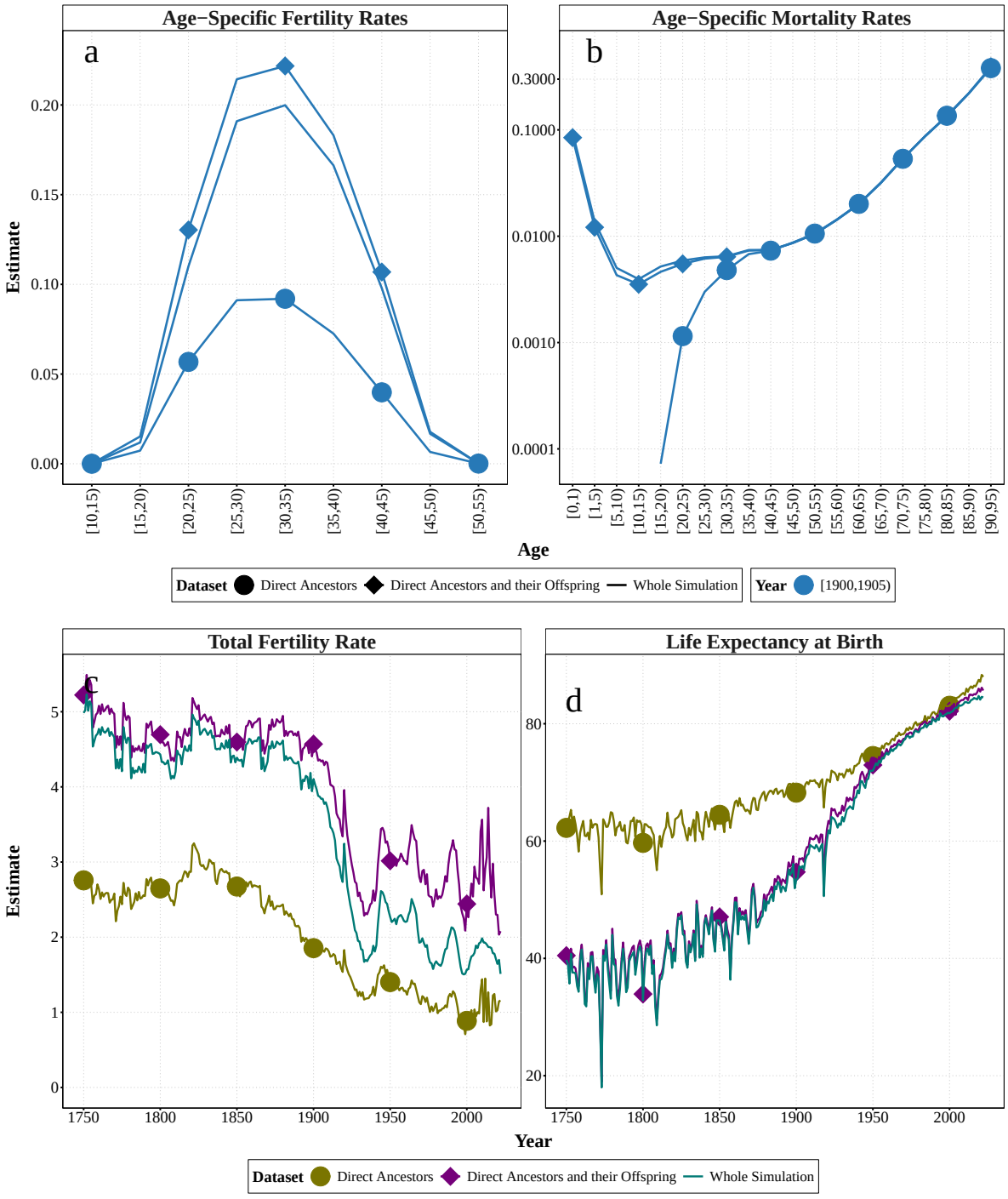

Fig 2: Experiment 2: Age-specific and summary demographic measures for women in Sweden derived from genealogical subsets of only direct ancestors and together with all direct ancestors' offspring extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure 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). These values are not shown.*

Experiment 3. Missing information on some subpopulations

In our third experiment, we examined the bias associated with missing information on two types of relatives who are often omitted from genealogies: children who died at an early age and childless women. We analysed the effects of omitting both subpopulations from the extended genealogies separately using a similar approach, but without considering their combination. Figures 3 and 4 compare, for women in Sweden, age-specific fertility and mortality rates in 1900–05, 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 five or childless women, respectively. For readability, we only show the estimates with 25% and 100% of omission. We compared 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 their offspring, which is 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 former, we also explored using a threshold of age one, but the results (not included here) show very similar patterns to those obtained using the threshold of age five, except for the age-specific mortality rates below the age of one or five. Figures comparing the age-specific estimates from 1900–05 with earlier (1800–05) and more recent years (2000–05) for each subpopulation are provided in Figures 9 and 10 in Appendix.

On the one hand, Figure 3 compares the estimates derived from omitting children who died before age five. 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 (lines with no shapes), with the bias increasing as the percentage of omission increases. The estimates from the genealog-

ical subset became lower than the whole simulation when 100% of early-deceased children. The age distribution is almost unaffected. This is true during periods of high fertility and mortality, but the gap with the extended genealogy is almost imperceptible in recent years when infant and child mortality is very low (see Figure 9).

Looking at the changes in the summary measure over time, as shown in panel c of Figure 3, the effect of omitting early deceased children on underestimating fertility increases the further back in time one goes. This bias increases in proportion to the percentage of omission, especially before the twentieth century, when both fertility and under-five mortality were high in Sweden. Before 1900, a 25% omission of early-deceased children leads on average to an underestimation of -0.04 children per woman (or a reduction of -0.73% in the TFR) compared to the whole simulation, while a 100% of omission leads on average to an underestimation of -0.94 children per woman (or a reduction of -21% in the TFR). Under-five mortality was particularly high in earlier centuries and, in the case of Sweden, began to decline from the eighteenth century onwards. Therefore, as we go back in time, a larger number of children ever born would be missing if those who died at an early age were omitted from the genealogies. From the twentieth century onwards, the gap to the extended genealogy begins to close and is minimal in recent decades.

For mortality, the age-specific estimates are lower only for ages zero to one and one to five, being non-existent with 100% omission (see panel b of Figure 3). This also holds for earlier and more recent years (see Figure 9 in Appendix). However, the omission of early deceased children has a large effect on the overestimation of life expectancy at birth (e_0) (see panel d of Figure 3). It increases significantly going back in time and with the proportion of omission. A 25% omission of early-deceased children can lead to an overestimation of e_0 by up to 2.6 years (i.e., $+5.8\%$), while a 100% omission to an overestimation of e_0 by up to 9.3 years (i.e., $+22\%$).

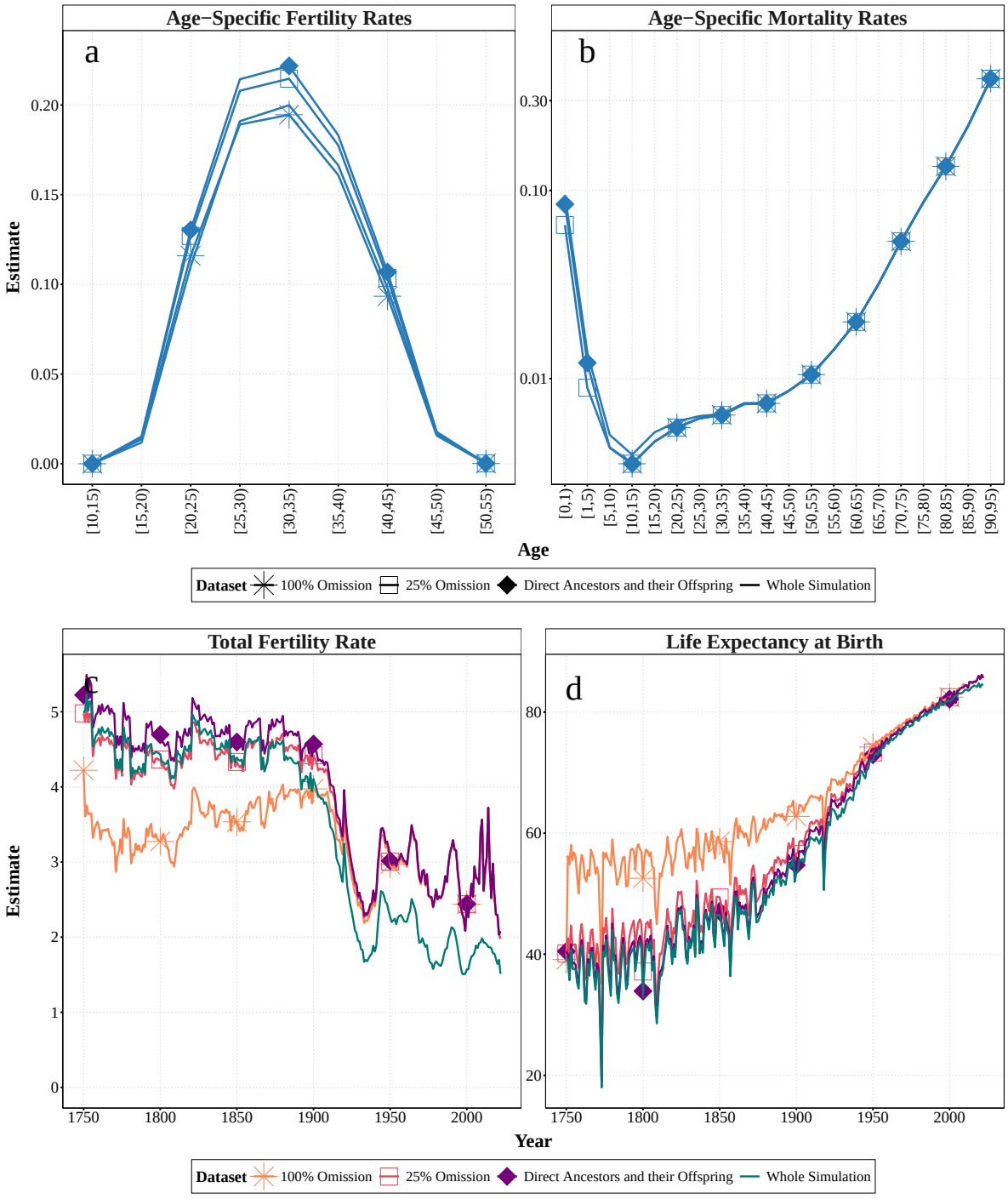

Fig 3: Experiment 3A: Age-specific and summary demographic measures for women in Sweden derived from genealogical subsets omitting different proportions of children who died before the age of five extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure 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). For the dataset with* 100% *omission all mortality rates below age 10 are 0. These values are not shown.*

On the other hand, Figure 4 compares the estimates obtained by omitting childless women. For fertility, this omission led to a slight underestimation of the age-specific rates compared to those derived from the extended genealogy, though they remain higher than the estimates from whole simulation (panel a of Figure 4. This also holds for earlier and more recent years (see Figure 10). Again, the age distribution is unaffected. This trend can be observed when looking at the evolution of the total fertility rate. Omitting childless women from the genealogies led to slightly lower estimates of TFR compared to the extended genealogies, but the estimates remain higher than the whole simulation over the entire period (see panel c of Figure 4). Otherwise, there is no significant change in the magnitude of the bias overtime.

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 1900–05, but also for previous and more recent centuries (see Figure 10 in Appendix). 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 1.05 years (i.e., +2.1%), while a 100% omission to an overestimation of e_0 by up to 1.5 years (i.e., +9.57%).

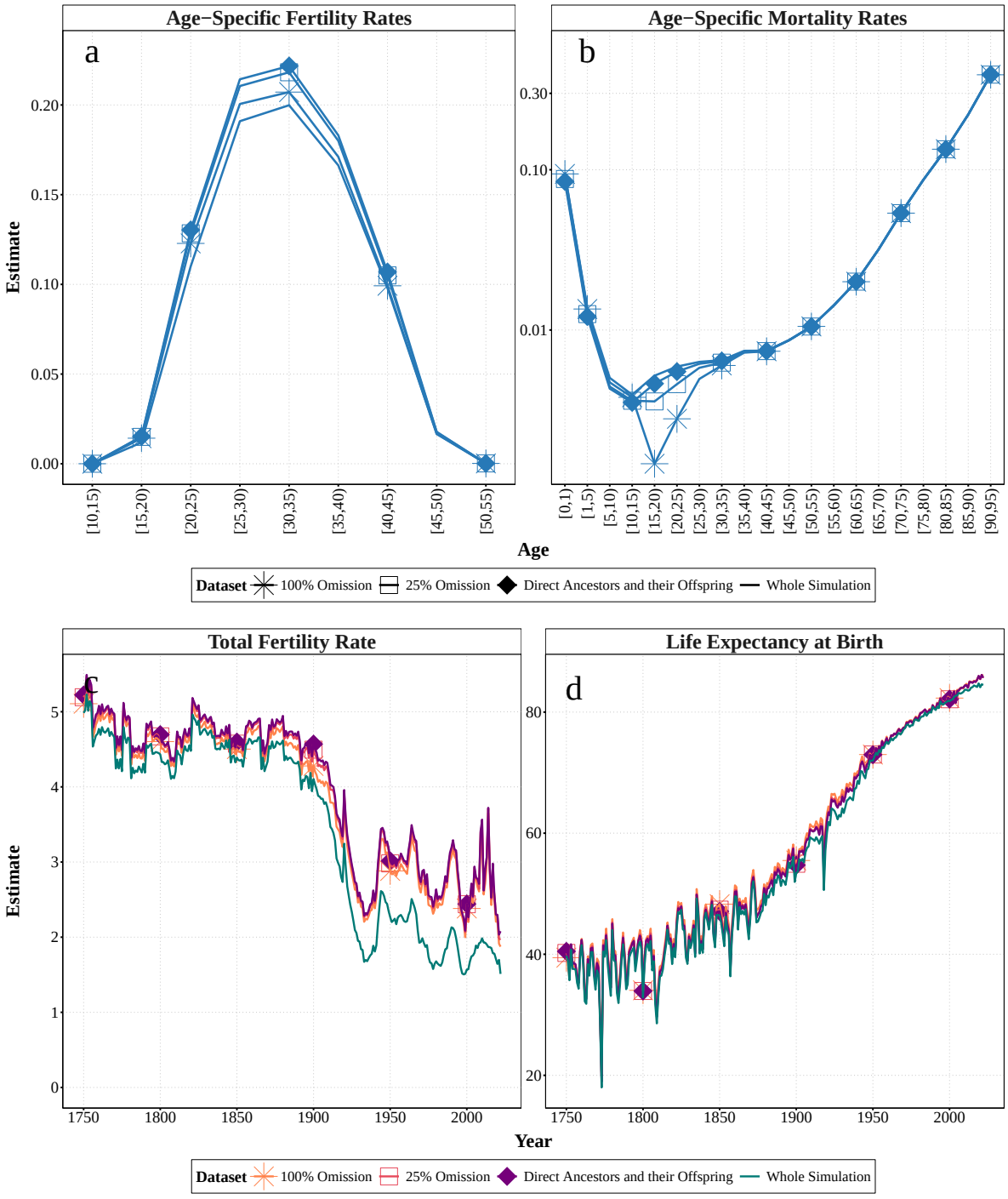

Fig 4: Experiment 3B: Age-specific and summary demographic measures for women in Sweden derived from genealogical subsets omitting different proportions of childless women extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure represents the means 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). These values are not shown.*

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 programme and taking Sweden (1751–2022) as a study case, we generated fully recorded synthetic populations that were then used as benchmarks for our experiments on ascendant genealogies. From the simulated populations, we traced the family trees of a group of hypothetical genealogists alive in 2022, and extracted different genealogical subsets that reproduced the three sources of bias. Hence, based on lineages that survived to the present, we explored: the selection in direct lineages leading to the exclusion of the childless, the incomplete reconstruction of family trees involving the inclusion/exclusion of all the direct ancestors' offspring, 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 trees, approximated here by the types of ancestors included, seem to affect the accuracy of demographic estimates based on ascendant genealogies. With the inclusion of the direct ancestors' offspring, the fertility underestimation turned into an overestimation, while mortality becomes closer to the real values. On the other hand, such effects do not appear to be linear, as their size varies over time, particularly between periods characterised by high and low fertility and mortality levels. Finally, the omission of subpopulations that are normally underrepresented

in genealogies seems to follow a similar pattern of variation, although it is more pronounced for children.

According to previous research, genealogical data are biased toward higher fertility and lower mortality. We find 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 (+33%) or omitted subpopulations. The expected higher fertility depends on the completeness of family trees. The estimates derived from our genealogical subsets are lower than those derived from the whole simulation in Experiment 1 which considers only direct ancestors and underestimates the number of births a woman might have given (−42%). After including all direct ancestors' offspring in Experiment 2, fertility from genealogies overestimates that from the whole simulation by +21%, while the overestimation of life expectancy at birth is reduced to +1.8% over the whole period. Both changes in the estimates suggest that the extension 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 thus cannot reproduce the whole complexity of their dynamics and structures. For instance, we do not consider the familial transmission of fertility and mortality behaviour, 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 the absence of reliable age-specific marriage rates by sex for the en-

[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.

tire study period, we modelled marriage and fertility behaviour using the 'marriage after childbirth' option in 'rsocsim'. It allowed us to address the lack of accurate historical data on marriage, although it may oversimplify real-world dynamics and the growing complexity of family structures. This approach performs reasonably well in periods when marital fertility was the norm, producing relatively accurate estimates of female fertility. It may introduce biases in contexts where childbearing outside marriage is prevalent or where family structures involve multiple partners, step-families, or complex dynamics, but these are not the focus of this analysis.

Third, in our research design, genealogists are randomly selected from individuals alive 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.

Fourth, 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 their offspring nine generations backwards without any restrictions. However, it may be more difficult for real 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. 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.

Fifth, 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, nieces, nephews or grandchildren of the genealogists, could provide some information. We also exclude affinal and in-law relatives 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 important to analyse 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 direct ancestors' offspring in the ancestors-only genealogies improved the accuracy of our estimates, particularly for mortality. This shows that the completeness of family trees within ascendant genealogies is crucial and should be carefully assessed before leveraging these data. Researchers working with these 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. As shown in this study, the underestimation of fertility may indicate the exclusion of some types of ancestors' offspring in genealogies, 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, although estimates can become more accurate if they are conditional upon survival to age five, ten, 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 and the kinship networks are complete enough so that the vital events of their members can be representative.

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 by 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.

Appendix

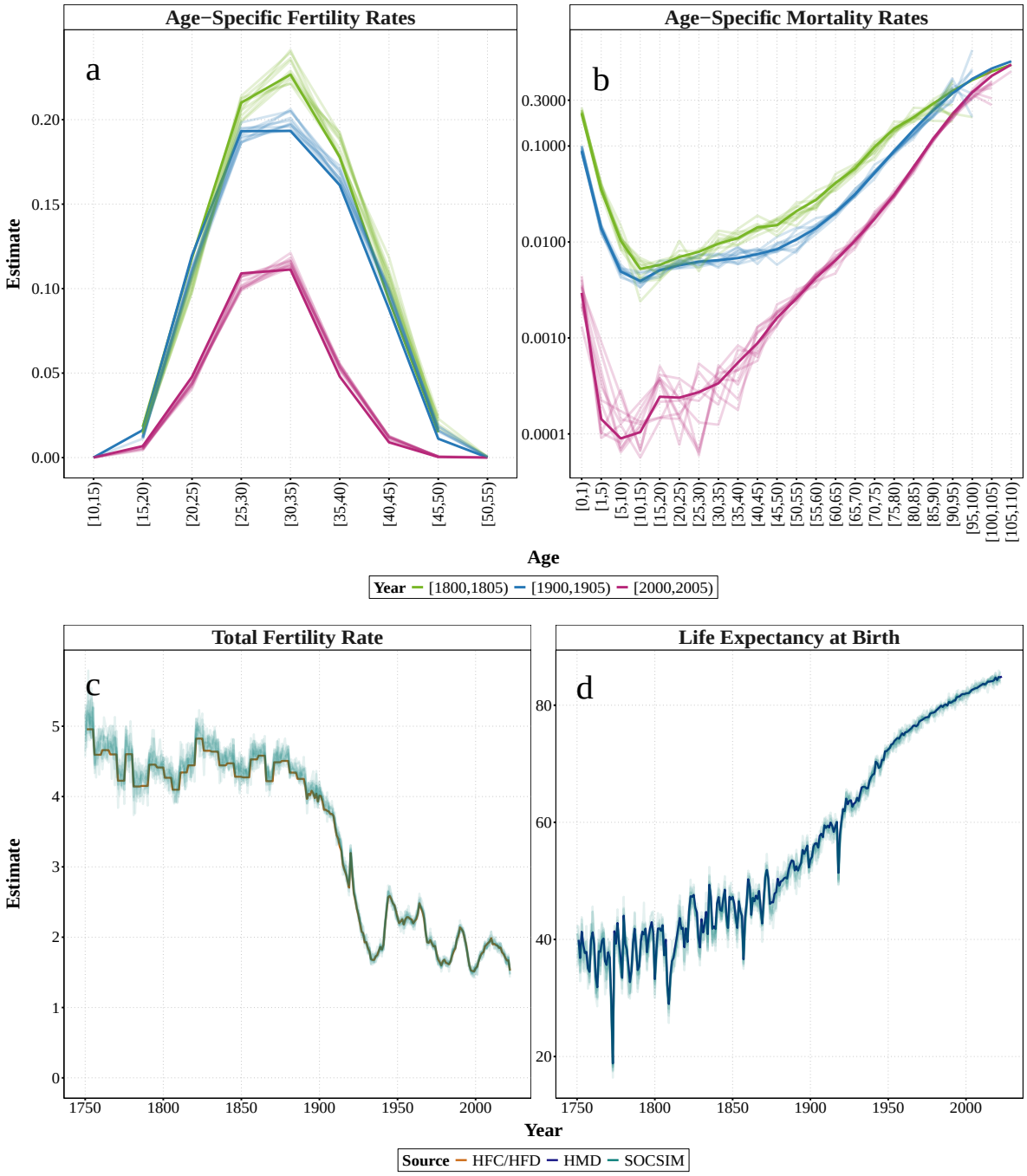

Fig 5: Age-specific and summary demographic measures for women in Sweden, retrieved from the Human Fertility Collection (HFC), the Human Fertility Database (HFD), the Human Mortality Database (HMD), and SOCSIM outputs. *Notes: The bold lines correspond to HFC/HFD and HMD estimates and the transparent 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, infinite (x/0) or NaN (0/0). These values are not shown.*

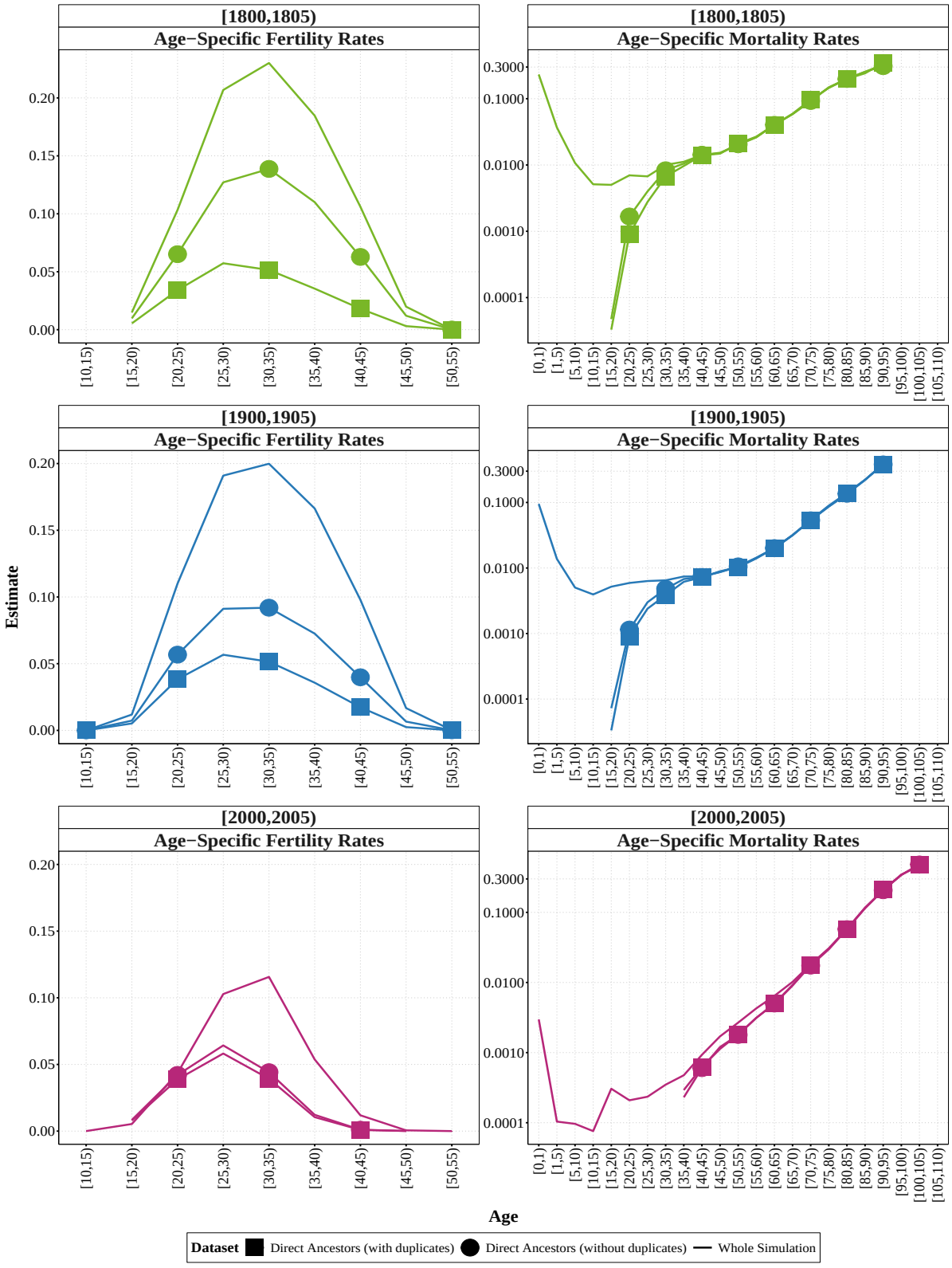

Fig 6: Experiment 1: Age-specific demographic measures for women in Sweden derived from genealogical subsets with only direct ancestors (with and without duplicates) extracted from SOCSIM simulated populations versus the whole simulations in three selected periods. *Notes: In each panel, the figure represents the means for each dataset. 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, infinite (x/0) or NaN (0/0). These values are* 31 *not shown.*

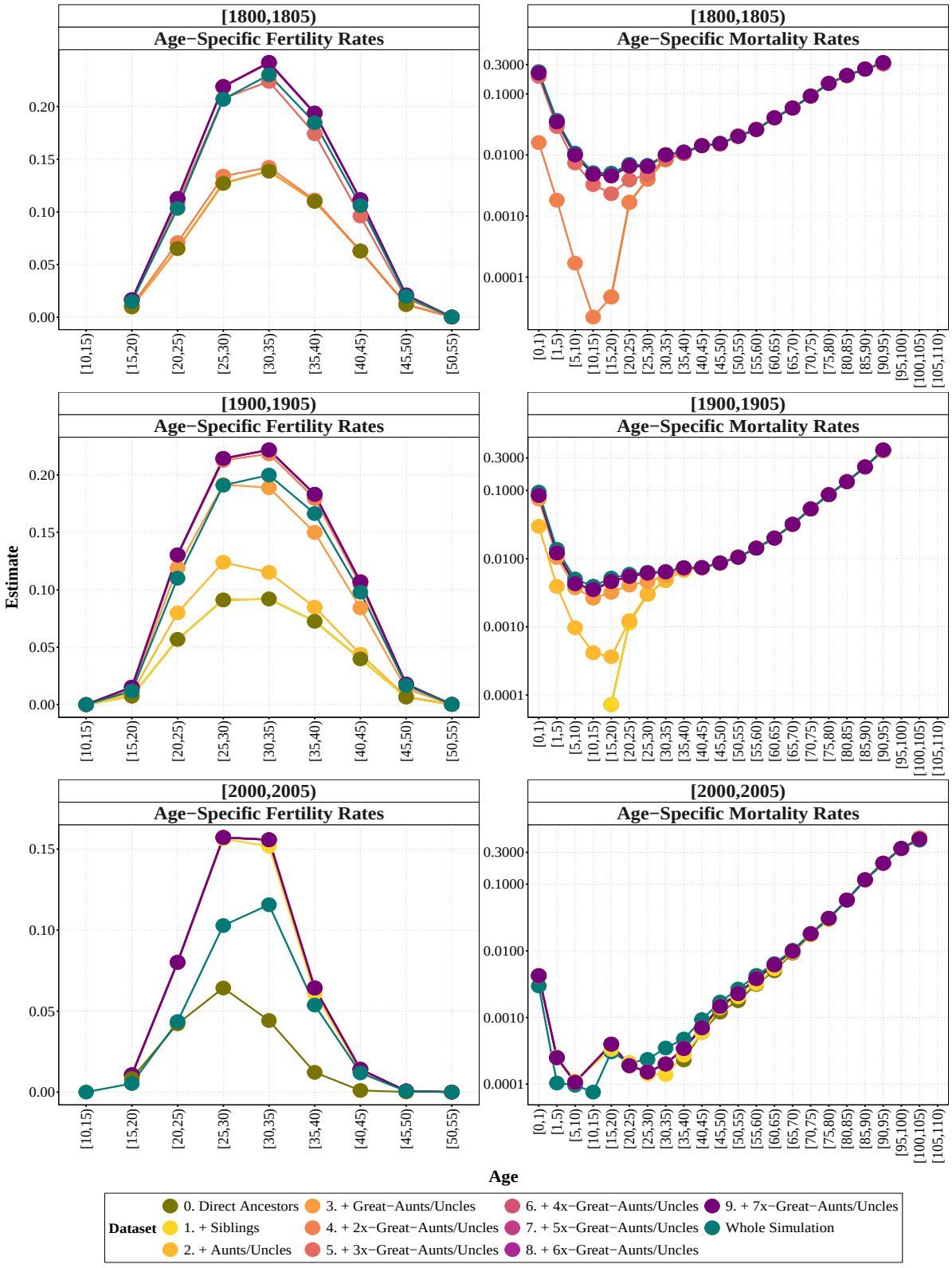

Fig 7: Experiment 2: Age-specific demographic measures for women in Sweden derived from genealogical subsets with only direct ancestors and with progressive inclusion of direct ancestors' offspring extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure represents the means for each dataset. 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, infinite (x/0) or NaN (0/0).* 32 *For the dataset with* 100% *omission of early-deceased children all rates below age 10 are 0. These values are not shown.*

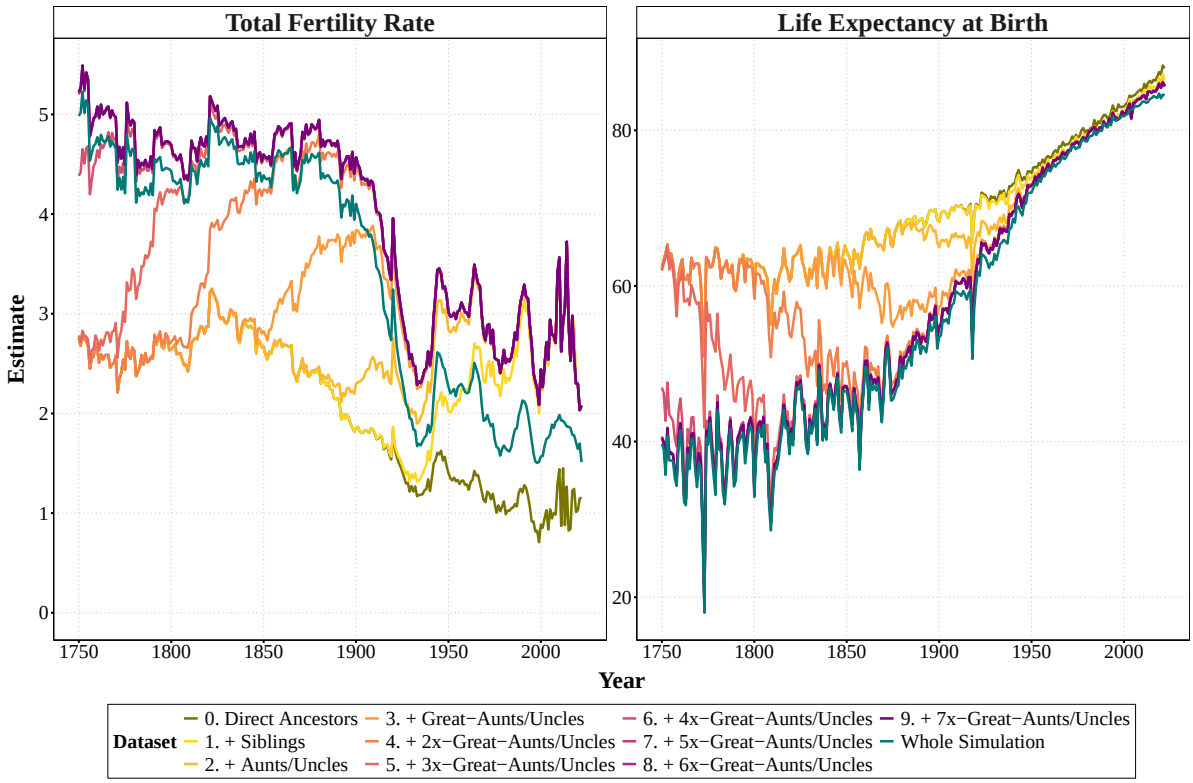

Fig 8: Experiment 2: Summary demographic measures for women in Sweden derived from genealogical subsets with only direct ancestors and with progressive inclusion of direct ancestors' offspring extracted from SOCSIM simulated populations versus the whole simulations. *Notes: In each panel, the figure represents the means for each dataset.*

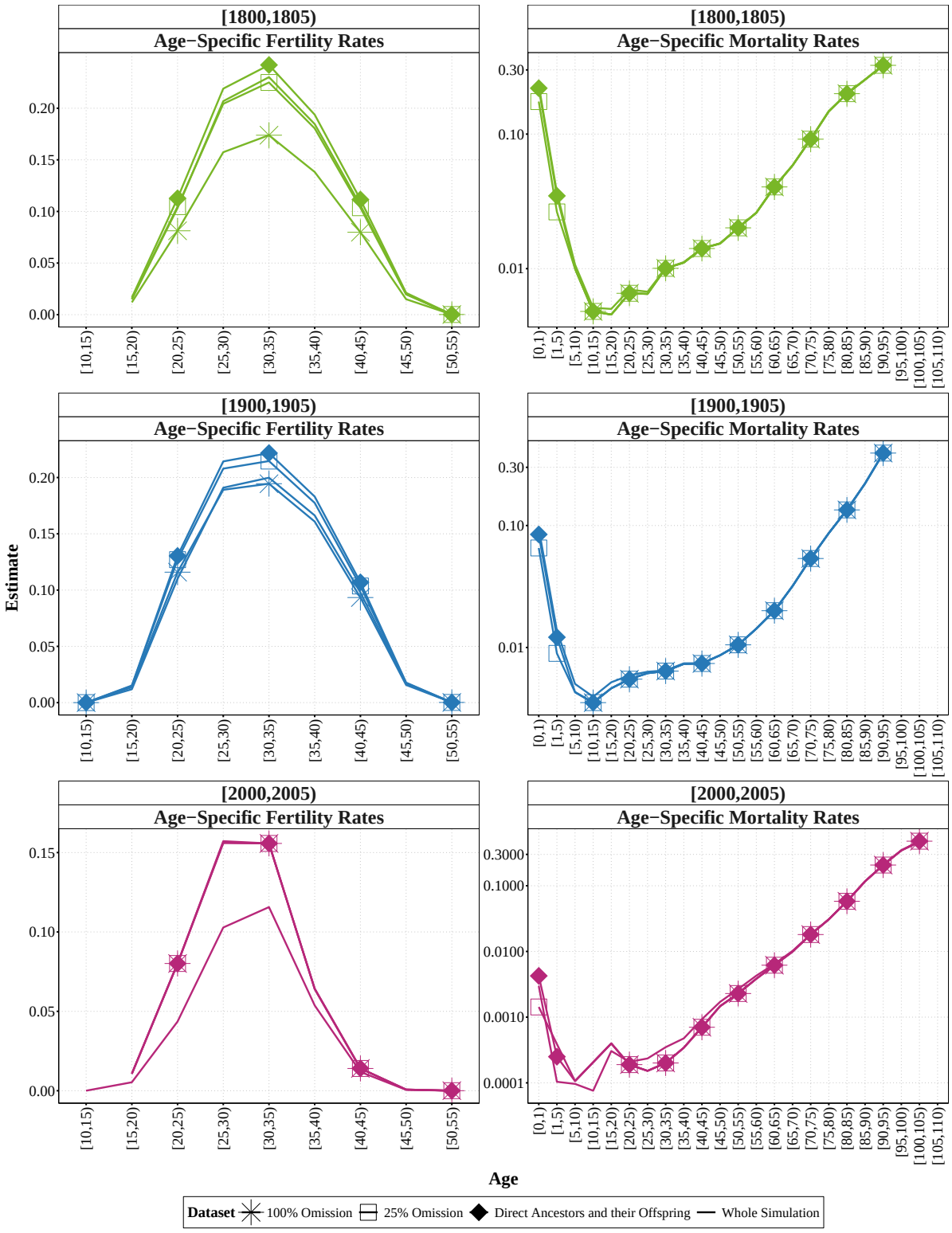

Fig 9: Experiment 3A: Age-specific demographic measures for women in Sweden derived from genealogical subsets omitting different proportions of children who died before the age of five extracted from SOCSIM simulated populations versus the whole simulations in three selected periods. *Notes: In each panel, the figure represents the means for each dataset. At young and very old 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). For the dataset with* 100% *omission all mortality rates below age* 10 *are 0. These values are not shown.*

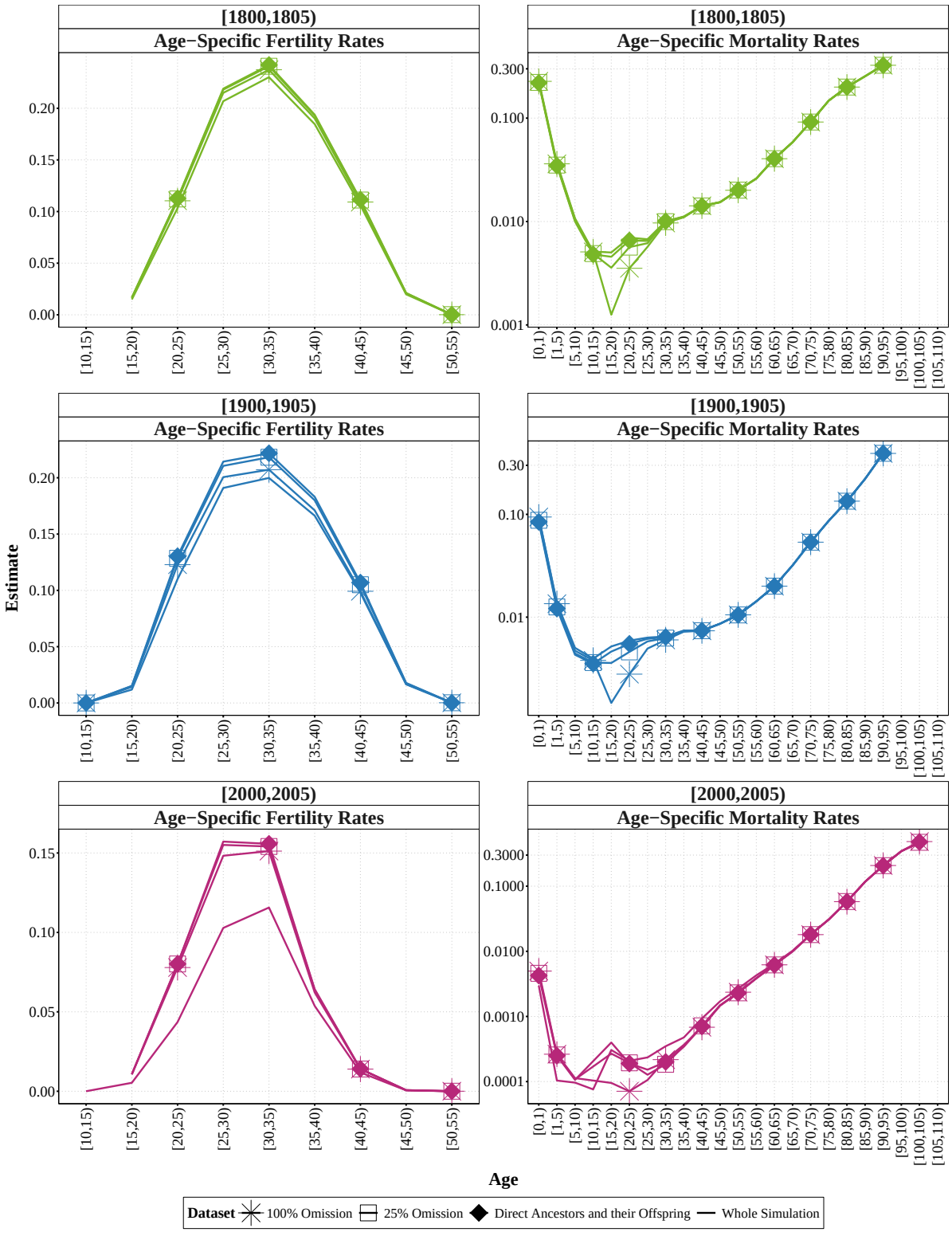

Fig 10: Experiment 3B: Age-specific demographic measures for women in Sweden derived from genealogical subsets omitting different proportions of childless women extracted from SOCSIM simulated populations versus the whole simulations in three selected periods. *Notes: In each panel, the figure represents 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). These values are not shown.*

Acknowledgments We thank Tom Theile for assistance with programming, particularly with rsocsim and code review. We thank Jim Oeppen for relevant exchanges during the early stages of the research, David Hacker for useful comments on a preliminary version of the manuscript, presented at the 2023 Population Association of America Annual Meeting. We thank Martin Kolk for useful feedback throughout the different stages of the research. 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. An updated version was presented at the PhD Half-Time Seminar at Stockholm University, where Linus Andersson provided valuable feedback. Liliana Calderón-Bernal gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS).

Data and code availability 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.

Colasurdo, A. and Omenti, R. (2024). Using online genealogical data for demographic research: An empirical examination of the FamiLinx database. *Demographic Research*, 51:1299–1350.

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.

Henry, L. (1956). Anciennes familles genevoises. Etude démographique : XVIe siècle - XXe siècle. - Présentation d'un cahier de l'I.N.E.D. *Population*, 11(2):334–338.

Hollingsworth, T. H. (1964). The Demography of the British Peerage. *Population Studies*, 18(Supplement 2).

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

Hollingsworth, T. H. (1977). Mortality in the British Peerage Families Since 1600. *Population*, 32(1):323–352.

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 (2024). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.fertilitydata.org.

Human Fertility Database (2024). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.humanfertility.org.

Human Mortality Database. HMD (2024). 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. (2024). Historical Patterns in the Intergenerational Transmission of Lifespan and Longevity: A Research Note on U.S. Cohorts Born Between 1700 and 1900. *Demography*, 61(4):979–994.

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.

Oeppen, J. (2021). Genealogies as a Source for Demographic and Genetic Studies: some estimates of bias. Unpublished Working Paper.

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.