Assessing the validity of microsimulated kinship networks using Swedish population registers *

Liliana P. Calderón-Bernal †

Max Planck Institute for Demographic Research and Stockholm University

Diego Alburez-Gutierrez ‡

Max Planck Institute for Demographic Research

Martin Kolk §

Stockholm University

Emilio Zagheni ¶

Max Planck Institute for Demographic Research

June 30, 2025

*Short title: Assessing microsimulated kinship networks

†Corresponding author, e-mail: calderonbernal@demogr.mpg.de, ORCiD: 0009-0002-7629-9718

‡ORCiD: 0000-0002-9823-5179

§ORCiD: 0000-0002-7175-4040

¶ORCiD: 0000-0002-7660-8368

Abstract

Estimating kinship networks is a data-intensive undertaking, typically carried out using empirical sources or demographic models. Empirical data, like population registers, provide a realistic picture but are scarce and limited by truncation and survivorship bias. Demographic models, like microsimulation, are less data-demanding, but only minimally account for population heterogeneity, family similarity and multipartner fertility. This study assesses the validity of kinship networks derived from microsimulation using Swedish population registers. We estimated the number of kin from grandparents to grandchildren for the Swedish population (cohorts 1915–2017) based on a SOCSIM microsimulation output and compared the results with the Swedish register counts. Mean numbers and distributions of most kin are similar between the sources. The microsimulation produces slightly lower numbers of kin for cohorts unaffected by register truncation, but better accounts for kin from earlier cohorts that are under-registered due to missing parent-child links in the registers.

Keywords: Kinship, Microsimulation, Population Registers, Sweden

Introduction

Kinship networks are a fundamental aspect of human life, surrounding individuals from birth to death. These networks extend beyond the nuclear family to include relatives linked by biological, legal or normative ties. Despite their importance across the life course, as a source of support and intergenerational exchange, our knowledge of kinship networks remains limited, particularly from an international and comparative perspective. Few countries worldwide have high-quality demographic data with the information necessary to study kinship, making it essential to consider alternative methods for examining other contexts where such detailed data are not available.

In this research, we use the Swedish population registers as a benchmark to assess the accuracy of kinship estimates derived from a relatively simple microsimulation model using population-level demographic rates as inputs. By comparing these two approaches, we examine the potential and limitations of using demographic microsimulation to count kin. While our analysis is based on the unique Swedish data context, our goal is to enhance our understanding of the accuracy of synthetic kinship networks, which could be employed to study kinship in other settings where only demographic rates are available.

Given a population register of sufficient quality, it should be possible to reconstruct the entire kinship network of a given individual (hereafter referred to as 'ego') based on data on parents and children. Starting from ego's parents and children, one can move up and down generations to identify ego's parents' parents (i.e. ego's grandparents) and ego's parents' children (i.e. ego's siblings); ego's children's children (i.e. ego's grandchildren); ego's grandparents' children (i.e. ego's aunts and uncles); ego's siblings' children (i.e. ego's nieces and nephews); ego's aunts' and uncles' children (i.e. ego's cousins), and so on. Kinship networks can be extended by including relatives through marriage. However, these reconstructions can be more challenging in practice due to limitations in the available data necessary to link parents and children across multiple generations.

There are two main approaches to estimating kinship networks within a population. The first approach is to derive them empirically using sources such as administrative registers, surveys, censuses or ethnographic methods. The second approach relies on modelling techniques such as formal demographic methods (Caswell, 2019, 2020; Caswell and Song, 2021; Caswell, 2022; Caswell et al., 2023) and demographic microsimulation (Ruggles, 1993; Zagheni, 2015; Zhao, 2006). Each approach has its advantages and limitations, which can result in differences in kinship estimates. On the one hand, empirical kinship counts can be derived from individual-level data and provide a realistic picture, but are limited by availability, coverage, truncation and survivorship bias. They can be calculated from available parent-child relationships in empirical data, following the procedure described earlier. On the other hand, analytical or microsimulation models can be run using population-level data, but may not capture the full complexity of the dynamics of real populations. In such cases, the average number of kin can be estimated analytically using mathematical models or by tracing kinship relationships from synthetic individuals produced by microsimulation, following the same parent-child link procedure.

As high-quality empirical data at the individual level are scarce, demographic models of kinship are a good alternative for estimating kinship networks using population-level data. Specifically, demographic microsimulation is an individual-based and computationally intensive tool used to model individual demographic events that are consistent with macro-demographic rates to gain insights into life course transitions (Zagheni, 2015). It allows the reconstruction of the demographic history and the kinship networks that would have been observed at any given point in time during the simulation (Murphy, 2010b). Among the most widely used microsimulation programmes, SOCSIM requires age-specific fertility rates and probabilities of death, while CAMSIM uses parity progression ratios, probabilities of death, mean ages at first marriage and proportions of never-married. Therefore, the demographic events experienced by the synthetic individuals (and their timing) are determined by the input parameters and the assumptions underlying the microsimulation. However, the two simulators have different approaches. SOCSIM takes a

population perspective and considers changes in demographic rates over time, whereas CAMSIM takes an 'ego-centric' perspective and is based on the assumption of demographic stability (Zhao, 2006). SOCSIM's population- and time-driven approach makes it suitable for estimating kinship networks over time and comparing them with empirical data over multiple generations.

Some data features or assumptions of the microsimulations may lead to differences in the synthetic kinship networks compared to those based on empirical data. First, since microsimulations are the outcomes of the input data (and parameters), the only uncertainty in the simulated kinship networks should arise from their stochasticity (Zagheni, 2015). The demographic events experienced by the simulated individuals are randomly assigned and executed, often using a competing risk model, according to predetermined rates or probabilities (Ruggles, 1993; Zagheni, 2015; Zhao, 2006). However, demographic data, usually available as aggregate measures, may not fully reflect the heterogeneity of real populations. For instance, data by marital status or parity, which would allow for a more accurate kin distribution, are usually unavailable. Such a lack of data disaggregation may result in more homogeneous synthetic populations.

Second, the resemblance in demographic behaviour between members of the same kinship group is often disregarded in microsimulation models. According to Ruggles (1993), ignoring these correlations can bias the mean number of kin downwards and result in a more homogeneous distribution than would be the case in real populations. In some simulators, this could be partially accounted for by adjusting some parameters (e.g. the heterogeneity and heritability of fertility) or by defining groups based on demographic resemblance within families. However, this would require more detailed data including family links that are rarely available and/or significant calibration of the input rates.

Third, partnership and fertility models must rely on simplifications of real-world dynamics. Microsimulation models can be closed, where partners are selected from living

simulated individuals, or open, where partners with suitable characteristics are created when needed for marriage (Zagheni, 2015; Zhao, 2006). Simulating childbearing within marriages or partnerships according to women's fertility schedules may not introduce bias into estimates for periods when marital fertility is the norm. However, when childbearing does not necessarily occur within marriage, accurately simulating fertility behaviour requires data disaggregated by age, parity, and marital status, which are rarely available. Furthermore, realistic kinship networks would require capturing both maternal and paternal lines, necessitating a two-sex fertility and partnership model with substantial data requirements. When such detailed data are unavailable, microsimulation results may underestimate multipartner and non-marital fertility, and may produce less accurate male fertility estimates. These limitations may be most apparent when studying family relationships that originate from a separation followed by a new union.

Otherwise, some features of the empirical data may also lead to differences between empirical and microsimulated kinship networks. On the one hand, empirical kinship counts are often affected by left-truncation, whereby some information is missing for individuals born close to the start of the registration system or data collection period. In particular, truncation of parent–child links in most administrative registers can limit the depth of complete kinship reconstruction, unlike simulations which record the entire population. Additionally, population registers often suffer from survivorship bias as they tend to count only individuals who have survived to a given date. Administrative sources count the current population of a country, but they may miscount migrants or their ancestors. Furthermore, international migration can truncate administrative sources if vital events occur outside the country's borders and are therefore not recorded.

Given the discrepancies between the kinship networks derived from empirical and microsimulated data, systematically comparing their estimates may provide a more accurate assessment of their validity. Despite its necessary simplifications and assumptions, microsimulation is a powerful tool for estimating kinship relationships when only

population-level information is available. Using available demographic rates, it enables the entire kinship network of synthetic individuals, along with their vital events, to be created and traced over a selected period without any coverage, truncation or survivorship bias. Therefore, comparing kinship estimates derived from microsimulation with those derived from empirical sources could provide researchers with insights into the model's strengths and limitations. To our knowledge, only Wachter et al. (1997) has previously tested the validity of microsimulated kinship networks against empirical data using US national surveys. However, such an assessment has not yet been conducted using a complete enumeration of a 'gold-standard' source. Our research contributes to filling this gap by providing a comprehensive assessment of the validity of microsimulation for demographic and kinship studies.

Among the empirical sources for counting kin, administrative population registers are invaluable, and Sweden is renowned for having the longest historical series of high-quality register data. This makes Swedish population registers an ideal basis for comparison. Here, we examine the consistency in size and parity distribution of kinship networks derived from microsimulated and empirical data, to understand the main differences between the two approaches. We compare kin counts derived from the SOCSIM demographic microsimulation programme (Hammel, 1976) using fertility and mortality data for Sweden with empirical kin counts estimated from the Swedish population registers as provided in the 'Swedish Kinship Universe' by Kolk et al. (2023) (see Online Appendix 2).

With this systematic comparison, we aim to shed light on the accuracy of synthetic kinship networks derived only from fertility and mortality data. This offers valuable insights for researchers seeking to use microsimulation to study kinship in contexts where more detailed demographic data are unavailable.

Data and Methods

Using the SOCSIM microsimulation programme, we ran a demographic microsimulation of the Swedish population from 1751 to 2017 to obtain a complete register of all individuals ever alive during the simulated period, including their vital events and kinship relationships. Originally developed at the University of California, Berkeley (Hammel et al., 1976), and written in C programming language, SOCSIM is an open-source microsimulation programme that has been used for decades in demographic research to study, among others, kinship networks' availability (Alburez-Gutierrez et al., 2021; Margolis and Verdery, 2019; Murphy, 2010a,b, 2011; Verdery and Margolis, 2017), and loss (Zagheni, 2011). The microsimulator requires as input an initial population file with individual information on the date of birth and sex, and monthly age-specific fertility rates and age-specific probabilities of death that apply over the simulated period to individuals of a given sex, group, and marital status. If available, parity-specific fertility can be included using conditional age-specific rates by birth order. During the simulation, SOCSIM schedules and executes demographic events (births, marriages, and deaths) for each synthetic individual in the initial population and their descendants.

The operation of the microsimulator is described 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 every synthetic individual to be executed in the future. Only one event can be scheduled for each individual at any one time. After the execution of an individual's event (except in the case of death) or a change in marital status or parity, a new event is scheduled for that individual. The next event to be scheduled is determined by creating random waiting times depending on the individual's 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 scheduled events are sequentially executed in a random order. Then, SOCSIM increments the month and repeats the event execution. At the end of the simulation, SOCSIM writes an output population file with information about every synthetic individual who ever lived and a marriage file containing information about each simulated marriage.

We ran one large simulation using the 'rsocsim' R-package (Theile et al., 2023), with an initial population of 50,000 synthetic individuals and age-specific rates obtained from the Human Fertility Collection (HFC) (1751–1890), the Human Fertility Database (HFD) (1891–2017) and the Human Mortality Database (HMD) (1751–2017). The last two were retrieved via the 'HMDHFDplus' R-package (Riffe, 2015). Although microsimulations involve stochastic variation, fluctuations diminish as the population size increases, resulting in stable, near-deterministic outcomes. Conversely, smaller simulations show greater variability in kin counts and their distribution, resulting in noisy estimates. Therefore, we relied on a single large-scale simulation to reduce such noise and yield more reliable estimates of kinship networks.

Due to data availability, we used age-specific fertility rates for all birth orders combined up to 1969 and conditional age-specific fertility rates by birth order from 1970 to 2017. Using parity-specific rates improved the accuracy of kinship distributions for the most recent cohorts compared to a simulation using combined rates over the entire period (results not shown). However, as parity-specific data is only available from 1970, this enhancement only affects the fertility of earlier cohorts partially.

Otherwise, our microsimulation setup did not incorporate additional parameters for demographic heterogeneity or heritability, since such detailed information is lacking in the input data and alternative long-term data sources. Although SOCSIM offers a 'heterogeneous fertility' option that models heterogeneity and maternal-line heritability, the default parameters of this option underestimate fertility, particularly during periods of

high fertility, and would require significant rate calibration. Furthermore, while this option can mitigate the underestimation of certain collateral kin types to some extent (e.g. siblings, aunts and nieces), it also increases the occurrence of both null and very large kin sets. This results in over-dispersion, which also requires calibration. Therefore, to ensure broader applicability, especially in contexts lacking comparable calibration data or gold-standard validation, we opted for a simpler simulation setup without the 'heterogeneous fertility' option.

In line with earlier research using SOCSIM (see Hammel (2005)), we initially ran the simulator for 100 years using the first available set of age-specific rates (1751) to produce a stable age structure. After this century of 'demographic stability', the population tripled to 150,000 individuals alive at the beginning of 1751, which were then exposed to the corresponding annual rates for the period 1751–2017. At the end of the simulation (2017), the synthetic population reached about one million living individuals.

In the absence of accurate, age-specific marriage rates by sex for the entire period, we employed the 'marriage after childbirth' directive in 'rsocsim' to establish lifelong partnerships and select living, single partners whenever a previously single woman gave birth. Despite its name, this option corresponds to lifelong partnerships, rather than real marriages in the sociological sense. Following Alburez-Gutierrez et al. (2021), partners for each woman were selected from all living single men to minimise the squared difference between the observed distribution of 'man's age - woman's age' and a normal distribution with a mean of two and a standard deviation of three. Further information can be found in the supplemental material of Alburez-Gutierrez et al. (2021). It is worth mentioning that this partnership option is not intended to predict multipartnering, which limits comparisons of multipartner fertility, half-relatives, and stepfamilies. However, it provides a solution to the lack of empirical data on marriage, separation, divorce and non-marital fertility over the whole period and overall produces an accurate estimate of female fertility. Moreover, using this option, which is less demanding in terms of data,

for comparison with a 'gold-standard' source allows us to assess the accuracy of kinship networks simulated using only fertility and mortality data. We deliberately used a 'naïve' microsimulation setup (with no heterogeneous fertility or marriage rates, and no further disaggregation of the data) so that the insights about its accuracy could be helpful in other settings where only aggregated demographic data are available.

After checking the accuracy of the microsimulation output against the input age-specific demographic rates and summary measures (see Figure A1 in the Appendix), we estimated the synthetic kinship networks following the approach outlined in the 'Swedish Kinship Universe'. Using the SOCSIM output, we traced the kinship networks (i.e. children, parents, siblings, grandchildren, grandparents, aunts and uncles, nieces and nephews and cousins) of the synthetic individuals who were alive at the end of 2017, based on parent–child links, and we replicated the figures in Kolk et al. (2023). To facilitate comparison, we also reproduced the register-based figures using the aggregated data provided in the Online Appendix 2 of Kolk et al. (2023).

The empirical counts are based on the Swedish Total Population Register and the Swedish Multigenerational Register. These registers contain information on all individuals ever registered in Sweden since 1960, including their country of birth, sex, dates of birth and death, and the names of their biological mother and father. The analytical sample comprises all individuals born in Sweden between 1915 and 2017 who were recorded in the registers and were living in Sweden at the end of 2017 (N = 8,243,185). For these reference individuals (used as the denominator), the kinship counts (numerator) include all relatives, whether alive or deceased in 2017. We applied the same restrictions to our microsimulation models to provide an account of kinship links in 2017 and estimate the number of kin of each type in each age group, which corresponds to the cohort as the period is fixed. To provide an assessment of the two approaches, we reproduced the same figures, used the same definitions of kin and disaggregated in the same way as in Kolk et al. (2023)

The advantages and limitations of using Swedish register data to enumerate kinship networks were described in detail in the Online Appendix 1 of Kolk et al. (2023). However, some of the data features should be acknowledged before comparing microsimulation estimates with those based on empirical data. Modern Swedish registers began with the introduction of a unique personal identity number, or 'personnummer', in 1947–48, alongside the collection of data on parenthood for all surviving children under the age of 16 living in Sweden. After 1947, data on children's parents were obtained directly from birth records. The earliest cohort that can be linked to their parents and siblings are those born in 1932. Consequently, grandparents must have given birth to their children (i.e. parents, aunts, and uncles) after 1932 for links between cousins to be traceable. The annual population registers were only digitised after 1968, marking the starting point for the availability of demographic event records. Another condition for inclusion in the Swedish register data, as discussed in Kolk et al. (2023) study, is survival until 1960, since the first digitised censuses were those of 1960 and 1965. Thus, despite their quality, the Swedish registers are biased by left-truncation before 1932 and survivorship until 1960.

Results

This study compares two sets of results concerning the number of kin in Sweden in 2017: one based on a newly created microsimulation model, and the other based on previously published empirical results. The SOCSIM-based figures are generally consistent with the register-based figures from Kolk et al. (2023) in terms of mean numbers and distribution by type of kin. However, there are some differences due to the design and limitations of the microsimulation, or due to incomplete information on parent-child relationships before 1932 in the registers. To systematically compare the average number and distribution of different types of kin derived from microsimulation and Swedish register data, the following figures are structured into three panels. The top left panel (a) shows the microsimulation-based estimates and the top right panel (b) shows the corresponding

register-based estimates reproduced using the data from the Online Appendix 2 of Kolk et al. (2023). The figures are organised by birth cohort, or age measured in 2017, and illustrate the average number or the distribution of a given type of kin at different ages of individuals alive in 2017. The bottom panel (c) shows the difference between the two approaches, calculated as the microsimulation estimates minus the register-based estimates, indicating whether the microsimulation overestimates or underestimates the number or proportion of kin. The replication and comparison of the figures from Kolk et al. (2023) that are not included in this section can be found in the paper's Appendix.

Average numbers of kin

The microsimulation closely aligns with the registers in estimating direct kinship ties, such as the number of living children, parents, and grandparents, with slight differences between cohorts. Figure 1 shows that the average total number of children per woman is relatively consistent across sources, although it is lower in the microsimulation, especially for early cohorts. This underestimation may be partly due to differences in the number of childbearing partners. While the microsimulation overestimates the number of children with only one childbearing partner for women born after 1930, it underestimates the number of children with two or three childbearing partners across nearly all cohorts. This results from the partnership model used in the simulation, which is based on lifelong partnerships and only considers re-partnering after the death of the previous partner. Therefore, as the simulation set-up is not intended to model multipartner fertility, the distribution of children by childbearing partners may be biased, but the total female fertility remains overall comparable.

Otherwise, estimates of male fertility in the microsimulation are generally lower and show greater variation between cohorts (see Figure A2 in Appendix). In addition to the differences due to multi-partner fertility, the discrepancy between sexes may arise from the SOCSIM fertility model, which is based on women's fertility schedules. Men's fertility then depends on their partnerships with women who are scheduled to have children.

For living grandchildren, the estimates are relatively consistent for women born after 1940, except for the early cohorts, who have fewer children and then fewer grandchildren in the simulation (see Figure A3 in Appendix). However, the accuracy in the number of grandchildren per grandfather shows more variation across cohorts. As with fathers, the difference in accuracy between sexes is probably due to the dependence of the SOCSIM fertility model on women's fertility schedules.

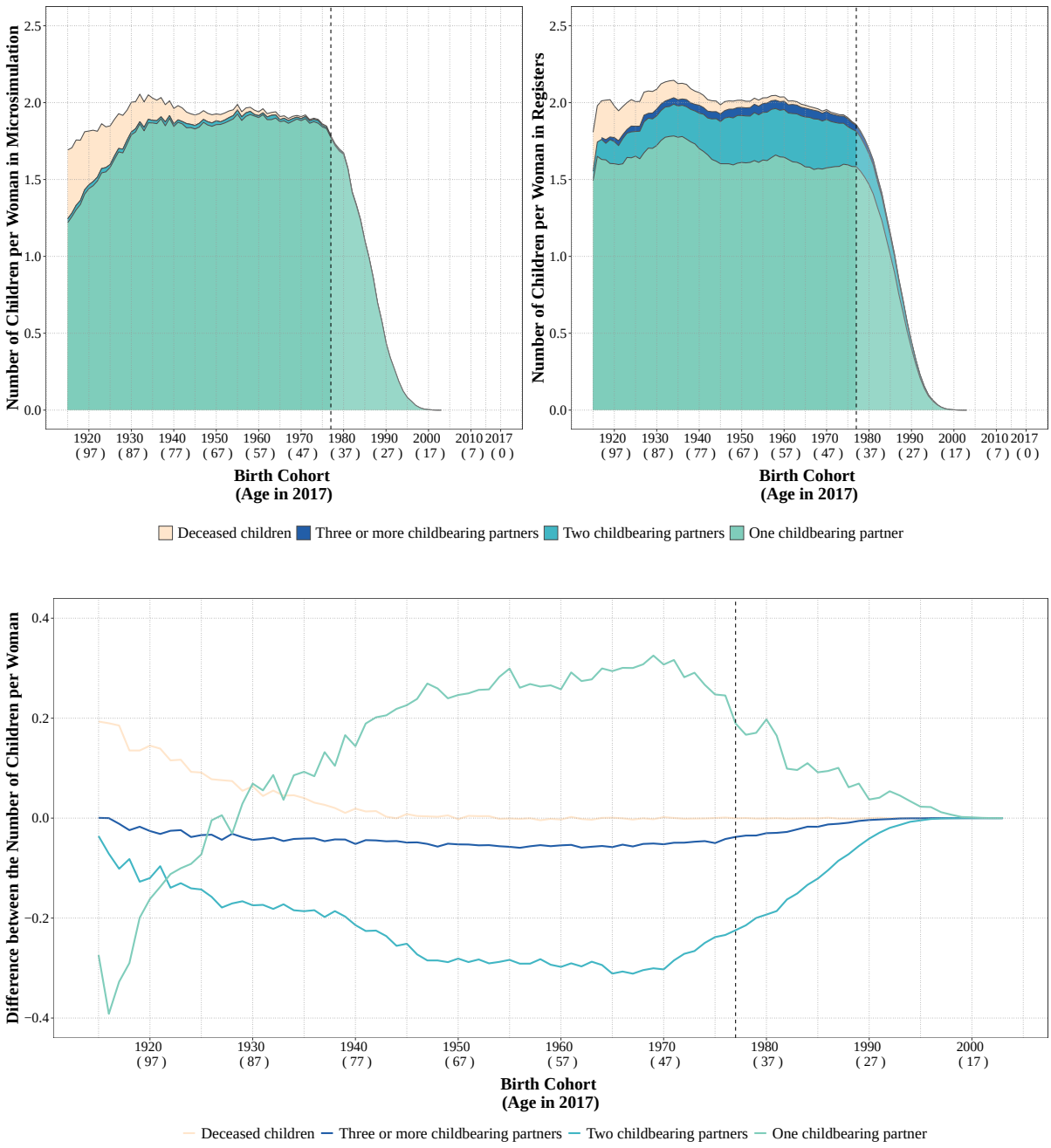

Figure 1: Average number of living and dead children per woman in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.*

For living parents, Figure 2 suggests that the numbers are overall consistent, though the microsimulation estimates slightly more living fathers for some cohorts (1945–65) and fewer for others (1967–90), likely compensating for the 'unregistered parents'. For living grandparents, Figure 3 shows that the mean number is similar up to the 1990 cohort, but is higher in the microsimulation for the most recent cohorts. The slightly higher counts of parents and grandparents in the simulation may take into account some of those labelled 'unregistered', who are not in the registers due to migration but are likely to be alive.

For other types of living kin, such as siblings, aunts, uncles, nieces, nephews and cousins, the microsimulation generally produces lower numbers for the cohorts with almost complete information in the registers (i.e. the benchmark), but higher numbers for earlier cohorts. Looking first at siblings, Figure 4 shows that the microsimulation produces lower numbers for the benchmark cohorts (post-1940). This undercount results from the differences in estimating full- and half-siblings between sources. The microsimulation produces more full siblings for most cohorts, except for those born in 1933–53, and fewer (almost zero) half-siblings in all cohorts. This underrepresentation of half-siblings is expected, given the partnership and fertility models used in the simulation, which do not include marriage rates. As explained before, the microsimulation set-up is intentionally simple to assess the accuracy of synthetic kinship networks derived only from fertility and mortality data.

The numbers of living nieces, nephews, aunts, uncles, and cousins follow a similar pattern, being lower in the microsimulation for the benchmark cohorts but higher for early cohorts. Specifically, the number of living nieces and nephews is lower for people born after 1940 (see Figure A4 in Appendix). This difference stems from the underestimation of half-siblings, leading to fewer nieces and nephews from half-siblings in the microsimulation. Although the microsimulation estimates more children of full siblings in most cohorts, this does not offset the lower number of nieces and nephews of half-siblings.

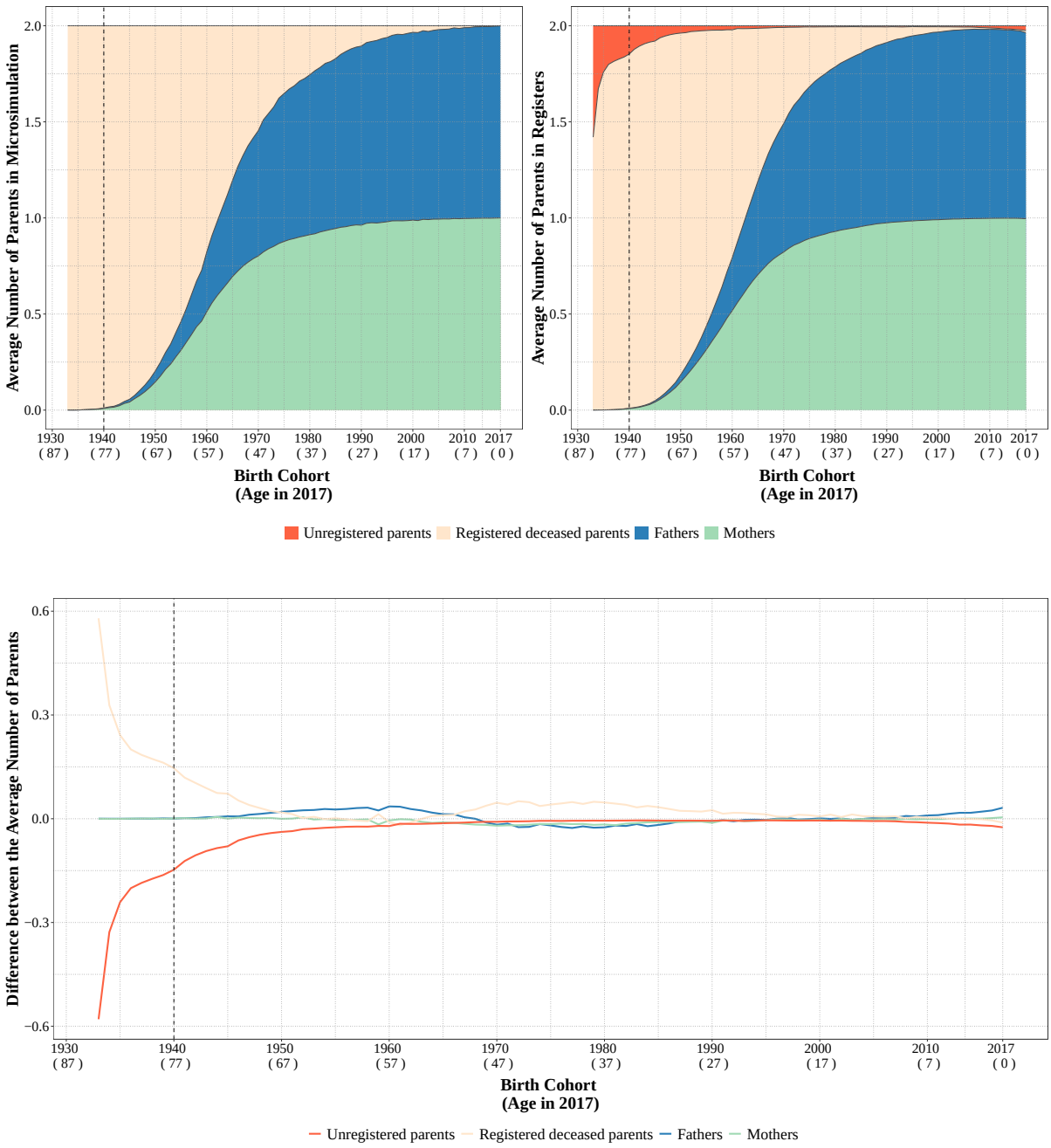

Figure 2: Average number of living, dead, and unregistered parents in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the left side represent cohorts with incomplete coverage of parents due to missing parent-child links in the registers.*

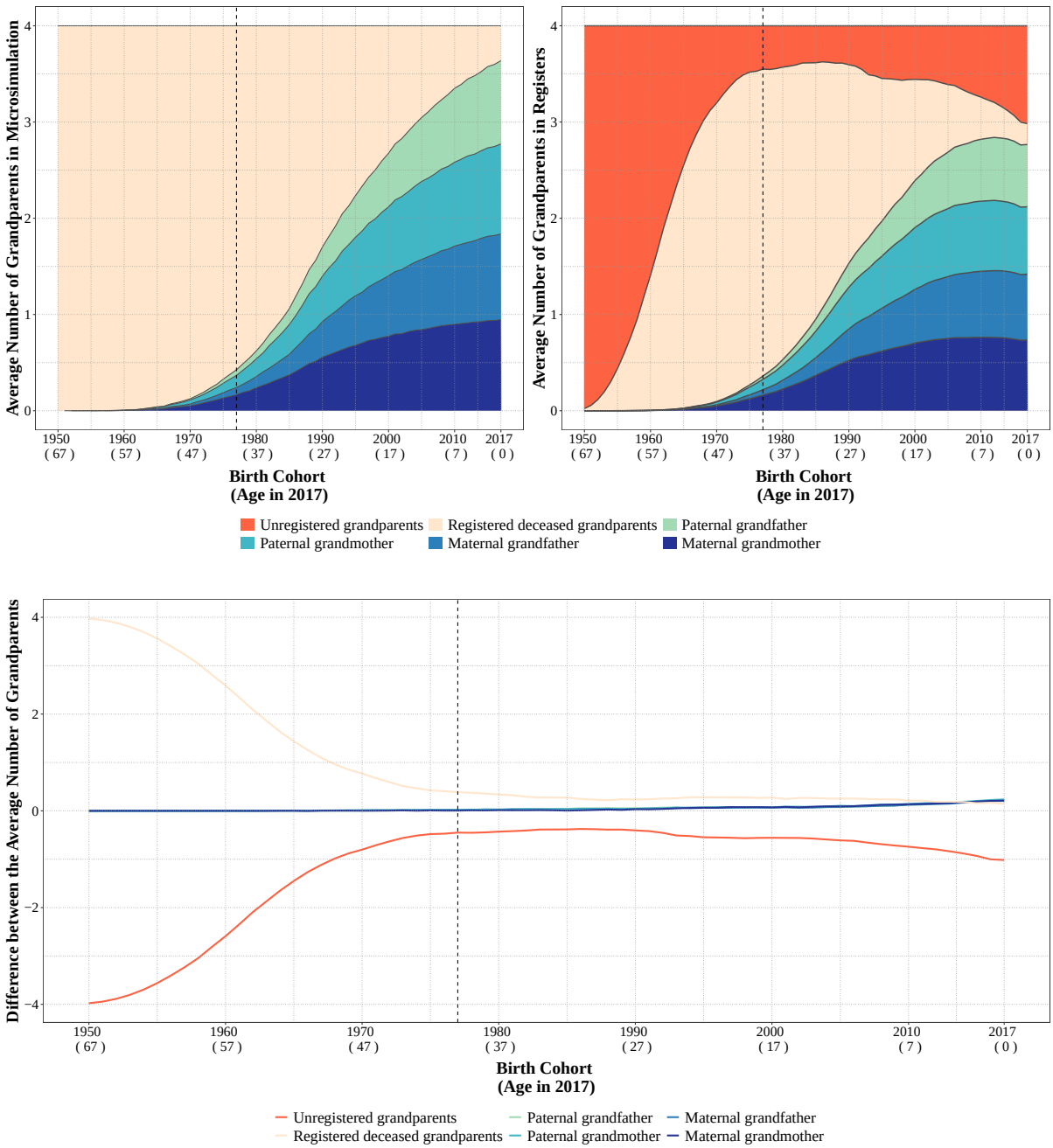

Figure 3: Average number of living, dead, and unregistered grandparents in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the left side represent the cohorts with incomplete coverage of grandparents due to missing parent-child links.*

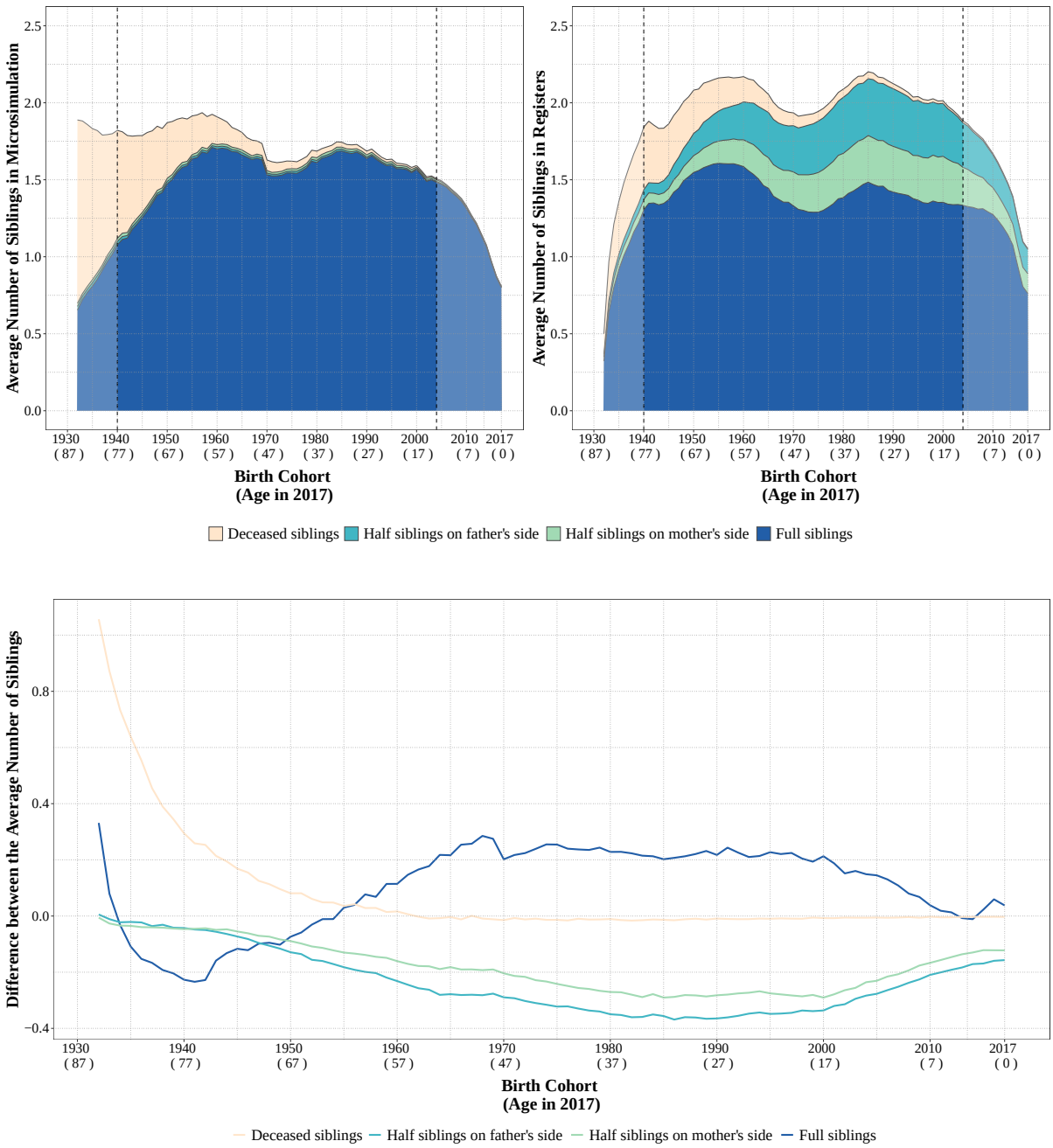

Figure 4: Average number of siblings, whether full or half-siblings and by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their siblings, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links.*

Similarly, the microsimulation estimates fewer living aunts and uncles for people born after 1965 (see Figure A5 in Appendix) and fewer living cousins for people born after 1970 (see Figure 5). However, for early cohorts, missing information on grandparents in the registers (see 'unregistered grandparents' in Figure 3), leads to missing aunts and uncles and then missing cousins in the empirical data. Therefore, while the microsimulation produces fewer horizontal kin for the benchmark cohorts, it provides higher estimates for the early cohorts affected by parent-child truncation in the registers.

Differences in the number of living kin may also be due to different numbers of deaths, which also play a role in shaping kinship networks. The microsimulation produces similar numbers of deceased children and grandchildren to the registers for most cohorts, but higher numbers for other types of kin for people born before 1970. On the one hand, the number of deceased children is similar for women (see Figure 1) and men (see Figure A2 in Appendix) born after 1940, while the number of deceased grandchildren is almost identical for both sexes in all cohorts (see Figure A3 in Appendix). On the other hand, the numbers of deceased parents for most cohorts (see Figure 2) and of deceased grandparents for all cohorts (see Figure 3) are higher in the microsimulation. The larger numbers likely account for some individuals classified as 'unregistered' in the empirical data. Given that the microsimulation records the entire population, any simulated parents or grandparents who are not alive are necessarily deceased, unlike the Swedish registers, where the status of the 'unregistered' individuals remains uncertain. This may be due to migration since the Swedish data only account for people registered in Sweden.

Otherwise, the numbers of deceased aunts and uncles (see Figure A5 in Appendix) and of deceased cousins (see Figure 5) are higher in the microsimulation for the early cohorts (up to +4.3 aunts and uncles and +2 cousins for the 1950 cohort). Missing information on deceased grandparents in the registers may result in undercounting aunts, uncles, and cousins, which might be better estimated in the simulation.

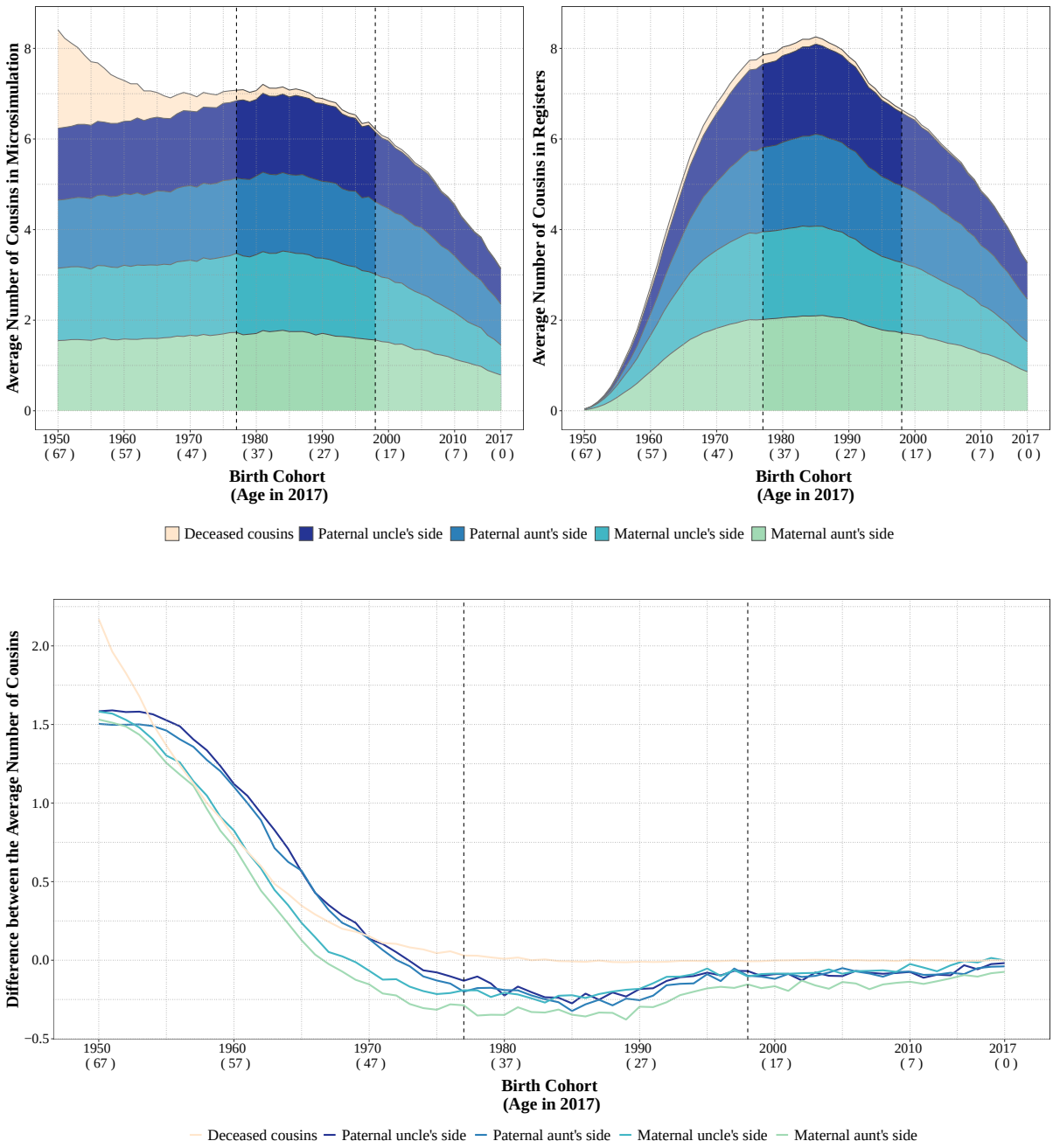

Figure 5: Average number of cousins, by birth cohort and by type of aunt or uncle, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their cousins, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links in the registers.*

To summarise the differences in the mean numbers of kin, Figure 6 shows that the microsimulation produces similar numbers of living kin to the registers for the benchmark cohorts (i.e. 1920, 1940 or 1970 depending on the kin), albeit slightly lower for some kin. These counts include children from all childbearing partners, as well as full and half-relatives. However, the discrepancies are more pronounced for early cohorts, where truncation and survivorship bias in the registers can lower kin counts.

Distributions of living kin

While the microsimulated and registered-based counts are relatively consistent in terms of the mean numbers of living kin, the differences increase when examining the parity distribution. For most cohorts and kin types, the microsimulation tends to over-represent individuals with the smallest family sizes while under-representing individuals with the largest family sizes. As shown in Figure 7, the microsimulation estimates a larger proportion of women with only one child, but a smaller proportion of women with two children, especially for the pre-1950 cohorts, which are less covered by parity-specific data. A similar pattern is also observed for other types of kin. For most cohorts, the microsimulation over-represent individuals with one grandchild (see Figure A7 in Appendix) or one sibling (see Figure A8 in Appendix), but under-represents individuals with four grandchildren, four or more siblings or eleven or more cousins (see Figure A9 in Appendix). For early cohorts, the microsimulation also overestimates women without children or grandchildren, but underestimates individuals without siblings or cousins, which are over-represented in the registers due to missing information.

Likewise, the microsimulation shows a more even distribution of the total number of living kin across all simulated cohorts (see Figure A10 in Appendix). Therefore, fewer individuals from each simulated cohort have null or large sets of kin compared to the empirical counts. The registers may lack information on some kin of individuals in the early cohorts, resulting in more null sets of kin, but are more likely to include above-average family sizes than the microsimulation.

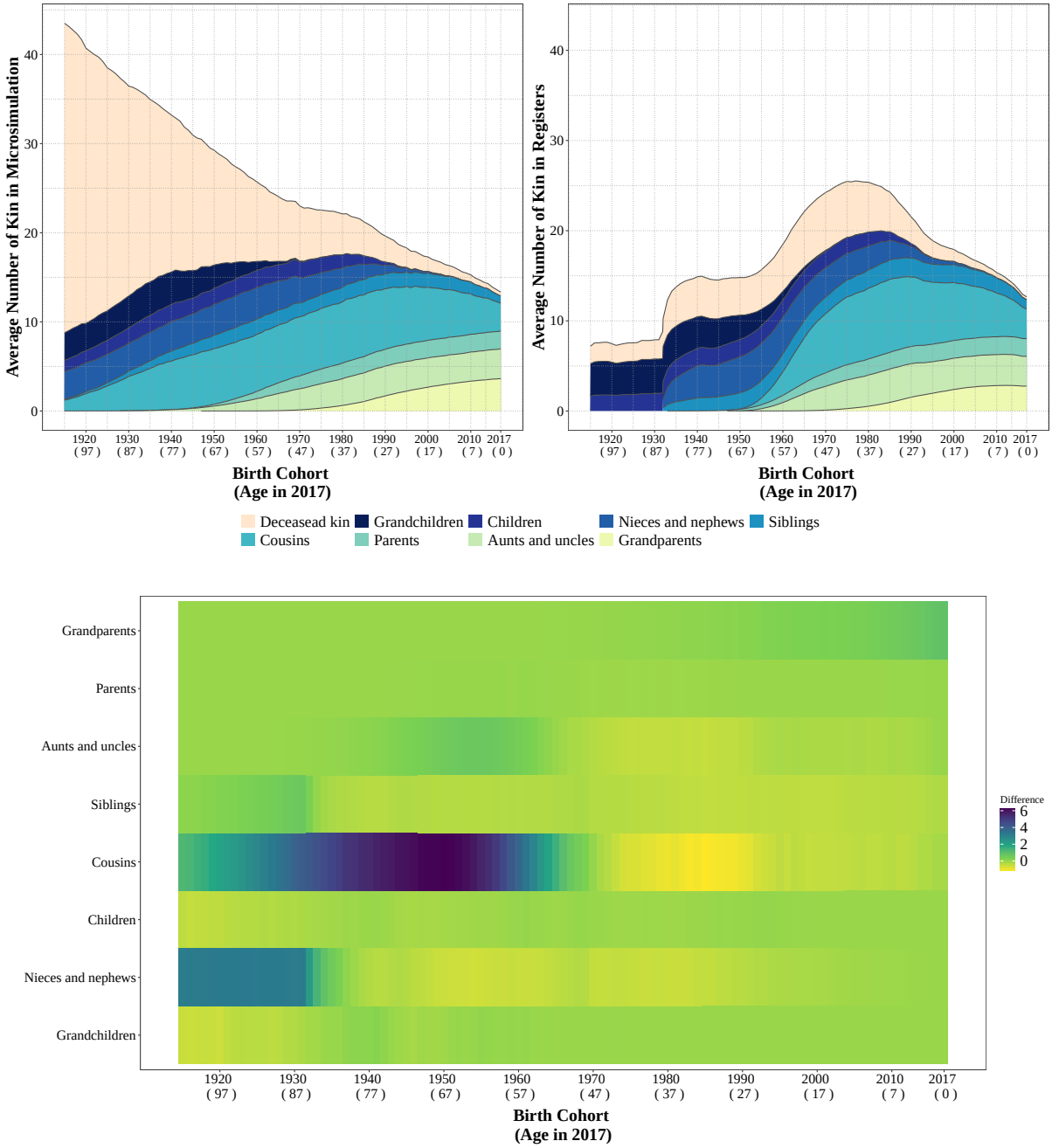

Figure 6: Average number of all types of kin in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

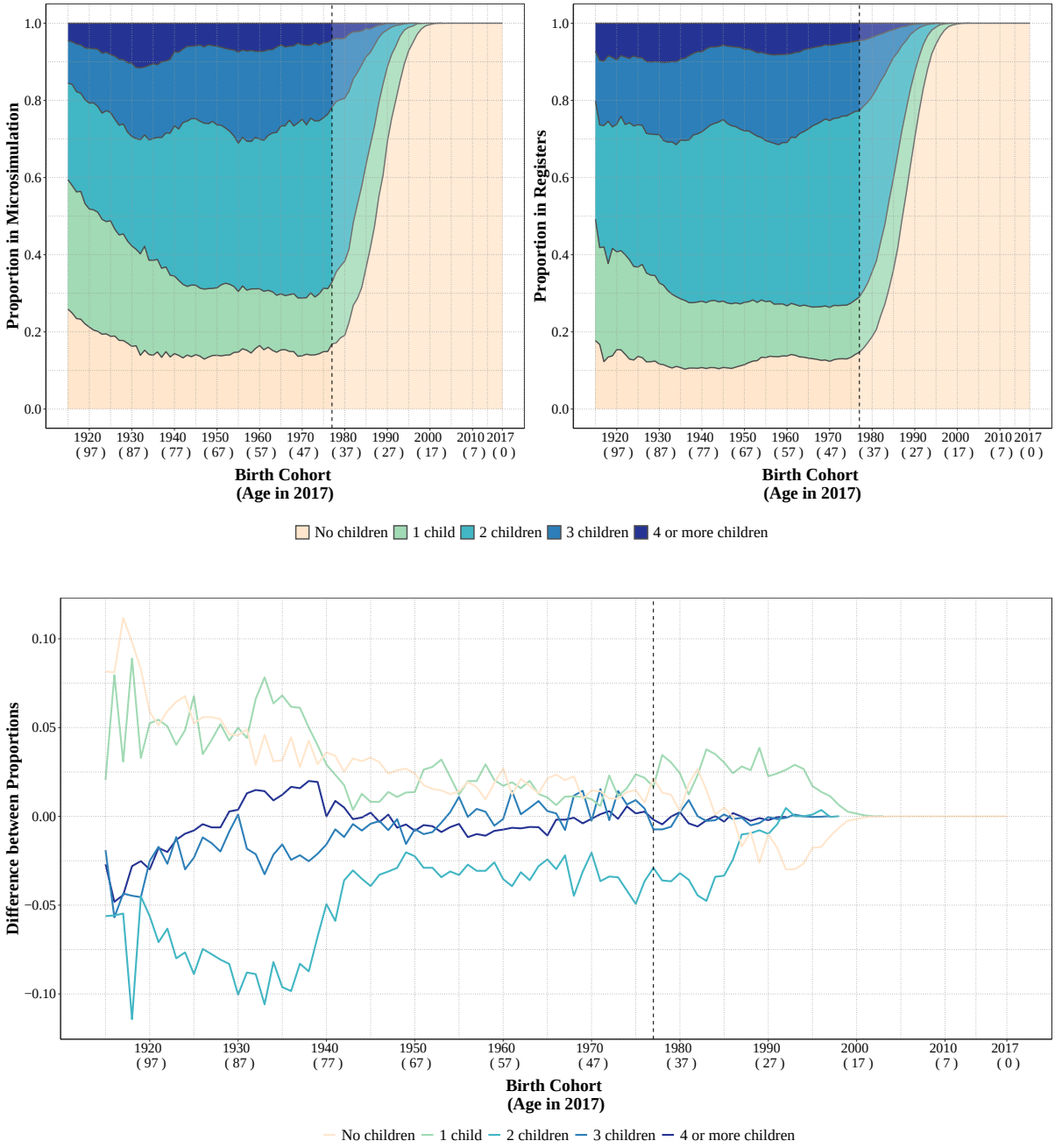

Figure 7: Proportional distribution of the number of living children per woman in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom. *The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.*

Discussion

Estimating kinship networks can be challenging due to the limited availability of data on vital events and family ties, particularly when considering relatives beyond the nuclear family or household. The two most common approaches to reconstructing these networks are using population registers and demographic microsimulation. In this study, we compared kinship networks for the contemporary Swedish population, estimated using both microsimulated and empirical data, to understand the main differences between the two sources and their effect on kin counts. We aim to shed light on the accuracy of synthetic kinship networks derived from a relatively simple microsimulation model by comparing them to a 'gold-standard' empirical source. The resulting insights could be valuable for future research using microsimulation to study kinship in contexts where only population-level data are accessible and a comparison with empirical data is not possible.

As shown on previous pages, the numbers of kin derived from the microsimulation and the registers are broadly similar, particularly regarding the mean number of living kin. However, the microsimulation produces slightly lower numbers of most types of kin for cohorts not affected by intergenerational parent-child truncation in the registers. This includes children for the 1920–70 cohorts; grandchildren for the 1920–35 cohorts; siblings for the 1940–2000 cohorts; nieces and nephews for the 1940–70 cohorts; aunts and uncles for the post-1975 cohorts, and cousins for the 1977–2000 cohorts. This underestimation is partly due to the limited estimation of multipartner fertility in the microsimulation, leading to lower numbers of children of multiple childbearing partners, half-siblings and nieces and nephews of half-siblings.

According to Ruggles (1993), the underestimation of some kin may also result from ignoring correlations in the demographic behaviour between members of the same kin group. The SOCSIM microsimulator enables different fertility and mortality rates to be assigned to specific groups. This could be used to replicate family clustering in de-

mographic behaviour by creating groups of families with distinct fertility and mortality rates. However, simulating these correlations would require a more complex microsimulation setup and substantial calibration of the input rates, because the original data are only disaggregated by sex and age. This would not align with the paper's goal of using a simple microsimulation to assess the validity of synthetic kinship networks.

As expected, the microsimulation produces less dispersion in kin counts across all kin types, reflecting limitations of the fertility model. Empirical registers show a higher proportion of individuals with no kin, especially in older cohorts, probably due to missing data, whereas the simulation by design includes all individuals. Conversely, large family sizes, which occur naturally in the real-world population but are less likely under the model's simplified fertility assumptions, are underrepresented in the simulation results.

Despite the undercount and underdispersion of most types of kin in the microsimulation for the benchmark cohorts, a 'fully registered' synthetic population allows better accounting for some relatives of the early cohorts who are likely to be omitted from the registers due to missing information on parent-child links, conditioning on survival or migration. This is the case of parents, siblings, nieces and nephews of people born before 1940, and grandparents, aunts, uncles, cousins and all types of deceased kin of people born before 1970. Moreover, microsimulation estimates can also go further back in time where reliable register data are unavailable.

The results presented here have some limitations that we would like to acknowledge. Our simulated results are based on synthetic populations, which are not 'real world' populations and therefore cannot reproduce the complexity of their dynamics and structures. On the one hand, we do not consider the familial transmission of fertility and mortality behaviour, because the input data are only disaggregated by sex and age. Thus, there is no predefined family clustering, beyond the individual stochasticity resulting from the microsimulation. On the other hand, compared with a simulation using fertility rates for

all birth orders combined over the whole period (results not presented here), the accuracy of the results improved slightly upon the inclusion of age-specific rates by birth order for 1970–2017. This suggests that using parity-specific data enables the microsimulation to perform even better than when all birth orders are combined, although such detailed data are rarely available. Finally, in the absence of data at higher levels of disaggregation (by age, parity and marital status) for the whole period, our modelling of partnership and fertility may oversimplify real-world dynamics and the complexity of modern families. While these simplifications may limit the accurate estimation of non-marital fertility, male fertility and multipartner fertility, as well as complex family structures involving half-siblings, estimates of female fertility remain reasonably accurate.

Despite these limitations, the SOCSIM microsimulation model produces kinship networks that closely match those in high-quality population registers in terms of average kinship. It also provides reasonable estimates of the parity distribution of certain kin types. This validates the use of microsimulations as a valuable tool for reconstructing kinship networks when only aggregate data is available and provides indirect validation of prior studies using SOCSIM. Furthermore, microsimulations can be employed to estimate kinship structures in historical periods lacking micro-level data, as well as to project future kinship dynamics based on demographic trends. By systematically comparing microsimulated networks with empirical data, this study strengthens the methodological basis of microsimulation in demographic research and encourages further investigation into kinship dynamics across diverse populations and contexts.

Acknowledgments

We thank Emma Pettersson for providing the R scripts developed for the Swedish Kinship Universe (Kolk et al., 2023), which enabled the reproduction of the figures in their paper and served as the basis for our replication using microsimulation. We thank Tom Theile for assistance with programming, especially with rsocsim. The manuscript was discussed at the Work-in-Progress Workshop of the Department of Digital and Computational Demography at the Max Planck Institute for Demographic Research, 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).

Author's contribution

LPCB led the study. LPCB, DAG, and MK conceptualised the study. LPCB, DAG, MK and EZ designed the methodology. LPCB implemented the analysis and visualisations and drafted the manuscript. DAG, MK and EZ reviewed and edited the manuscript. All authors approved the final version of the manuscript.

Disclosure statement

The authors report there are no competing interests to declare.

Data and code availability statement

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

Appendix

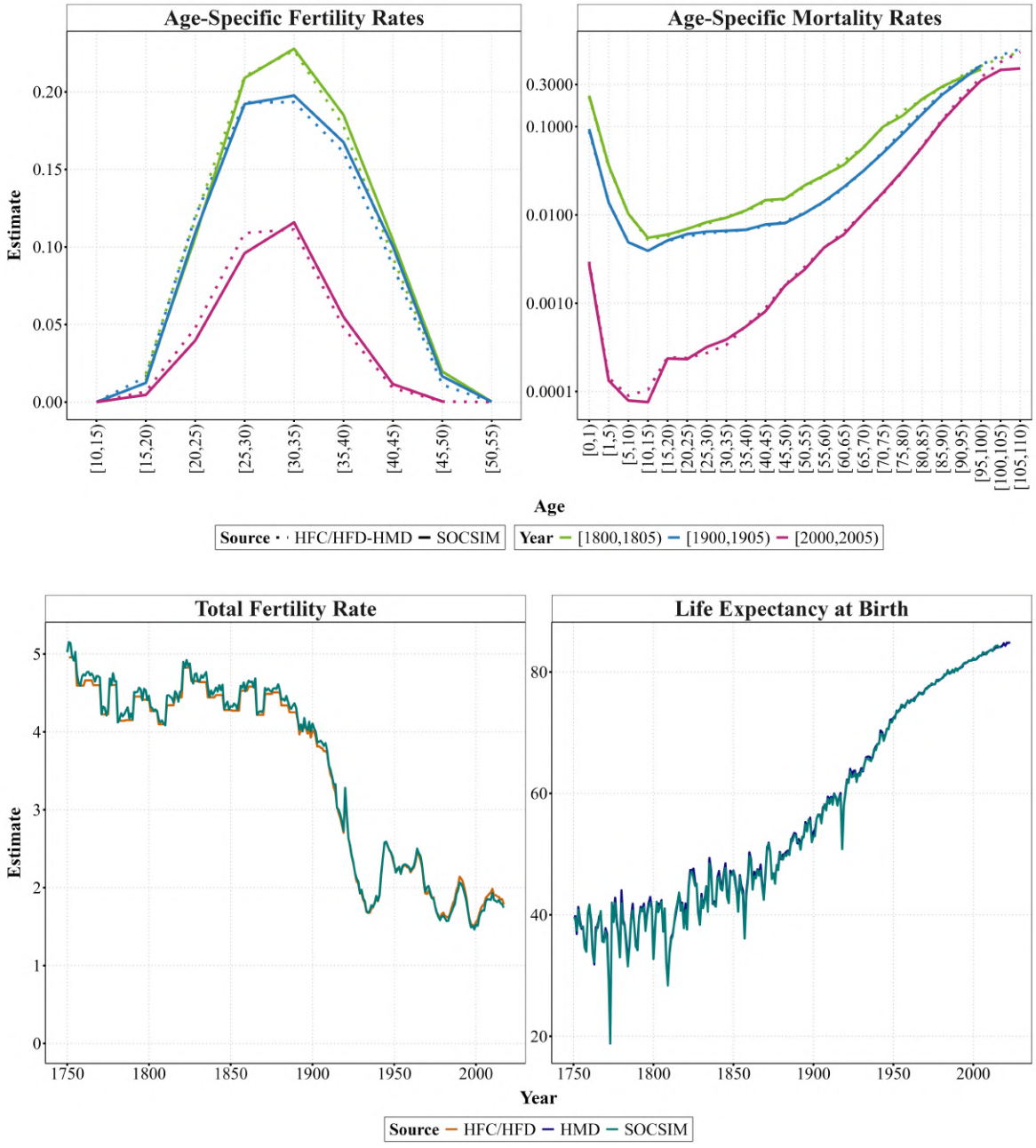

Figure A1: Age-specific and summary demographic measures for women in Sweden, retrieved from the Human Fertility Collection (HFC), the Human Fertility Database (HFD), the Human Mortality Database (HMD) and SOCSIM output.

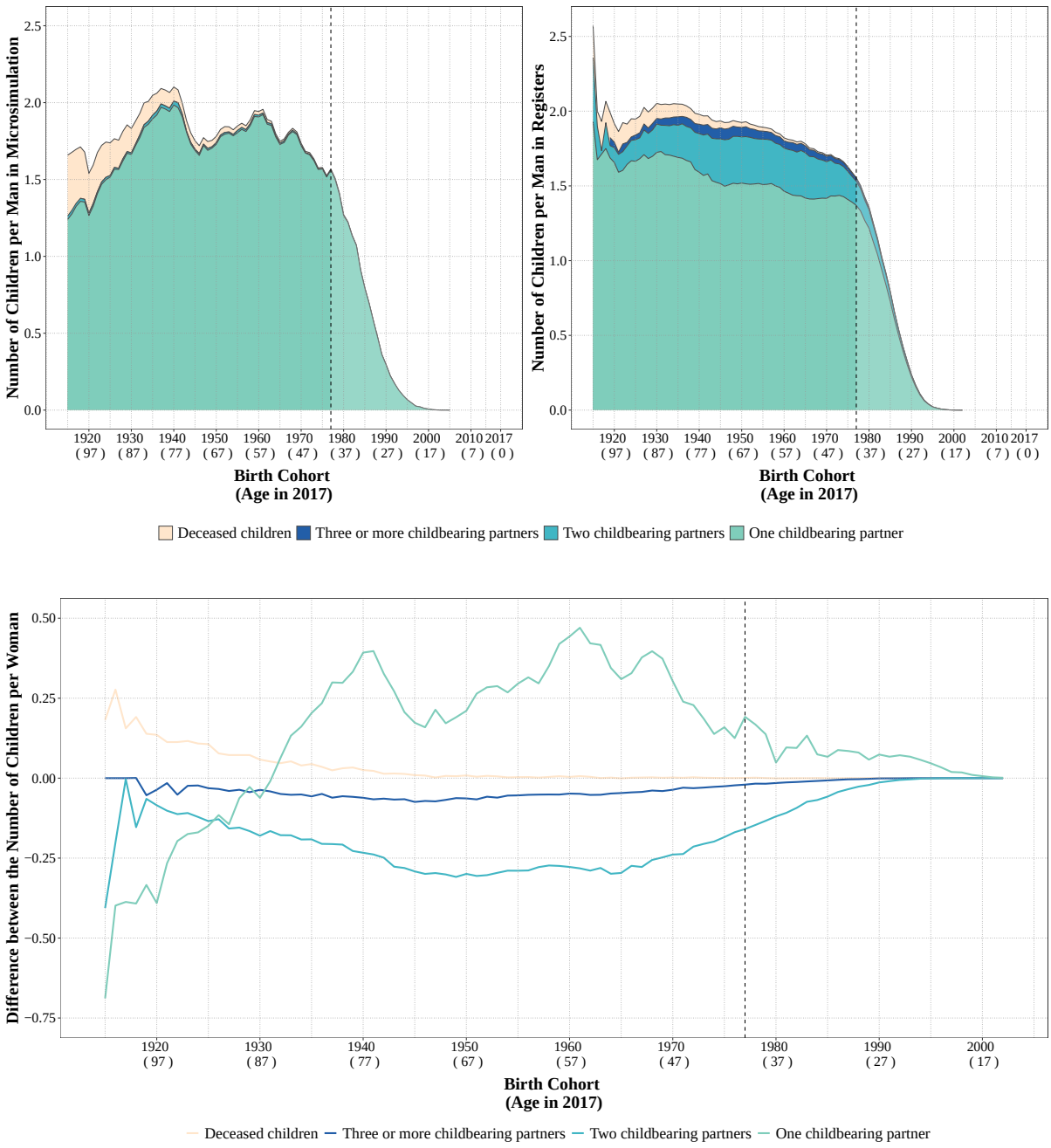

Figure A2: Average number of living and dead children per man in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.

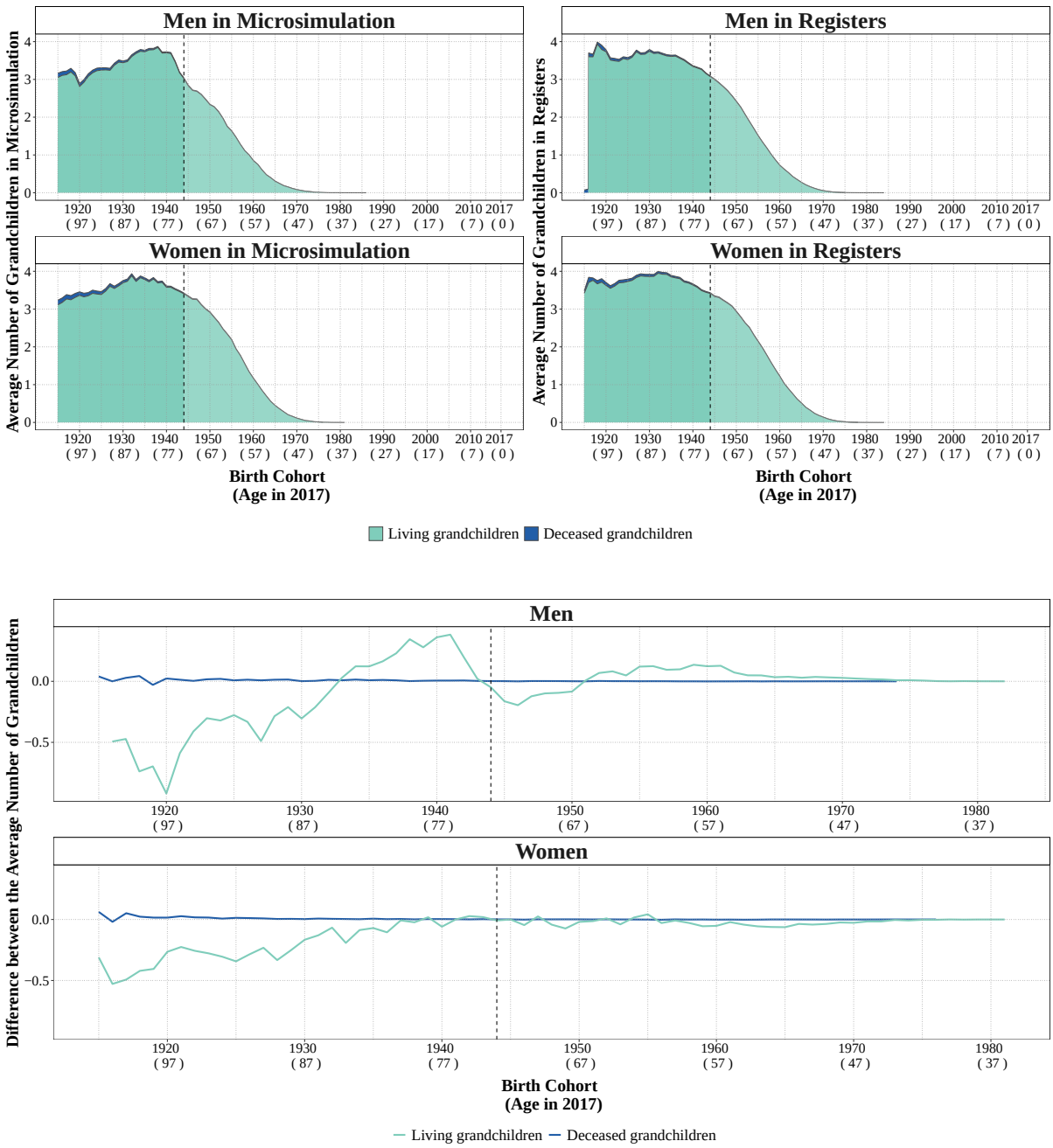

Figure A3: Average number of living and dead grandchildren in 2017, by sex and birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their grandchildren, as they were not yet born in 2017.

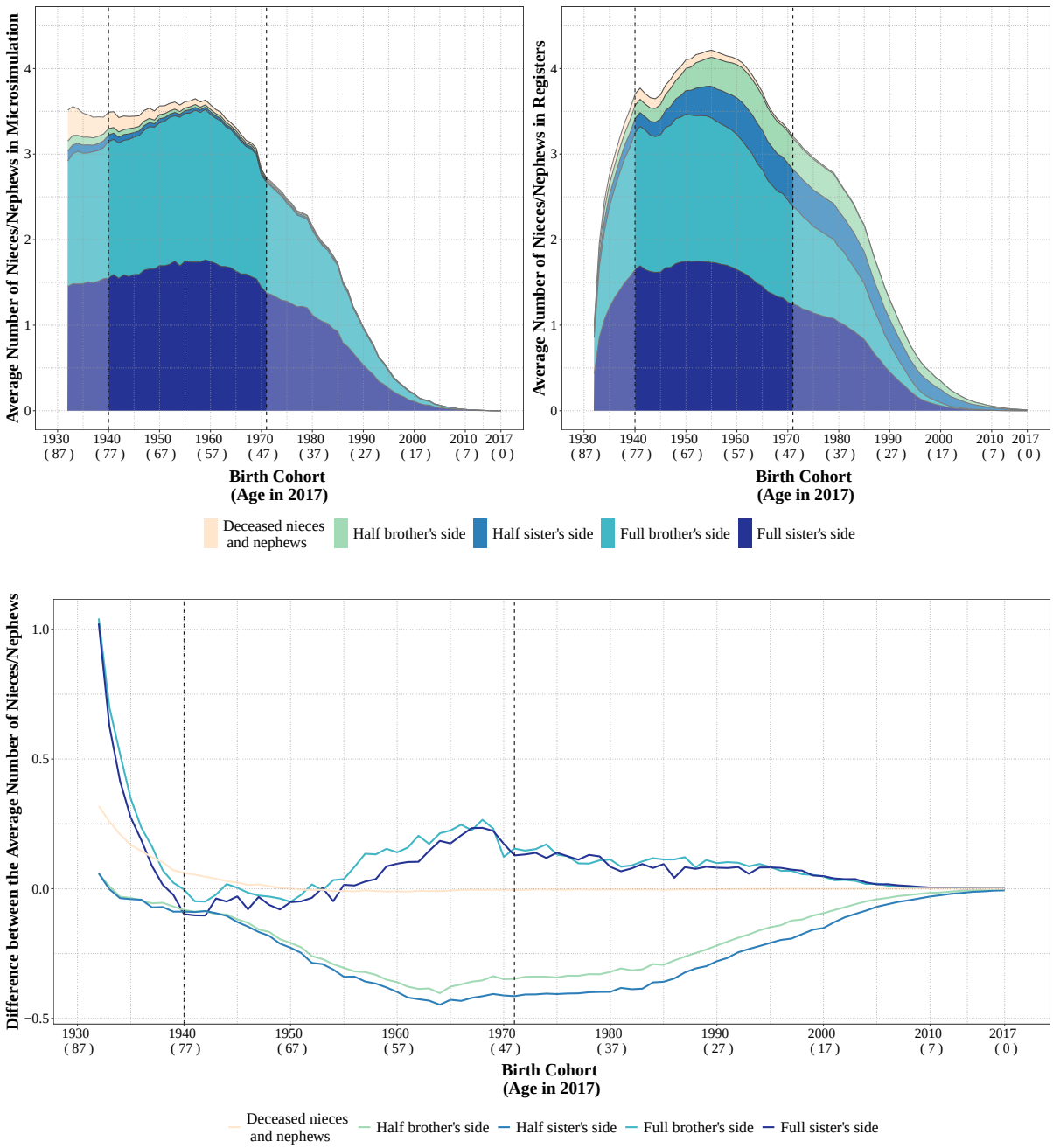

Figure A4: Average number of nieces and nephews, by birth cohort and through full- or half-siblings, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their nieces and nephews, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links in the registers.

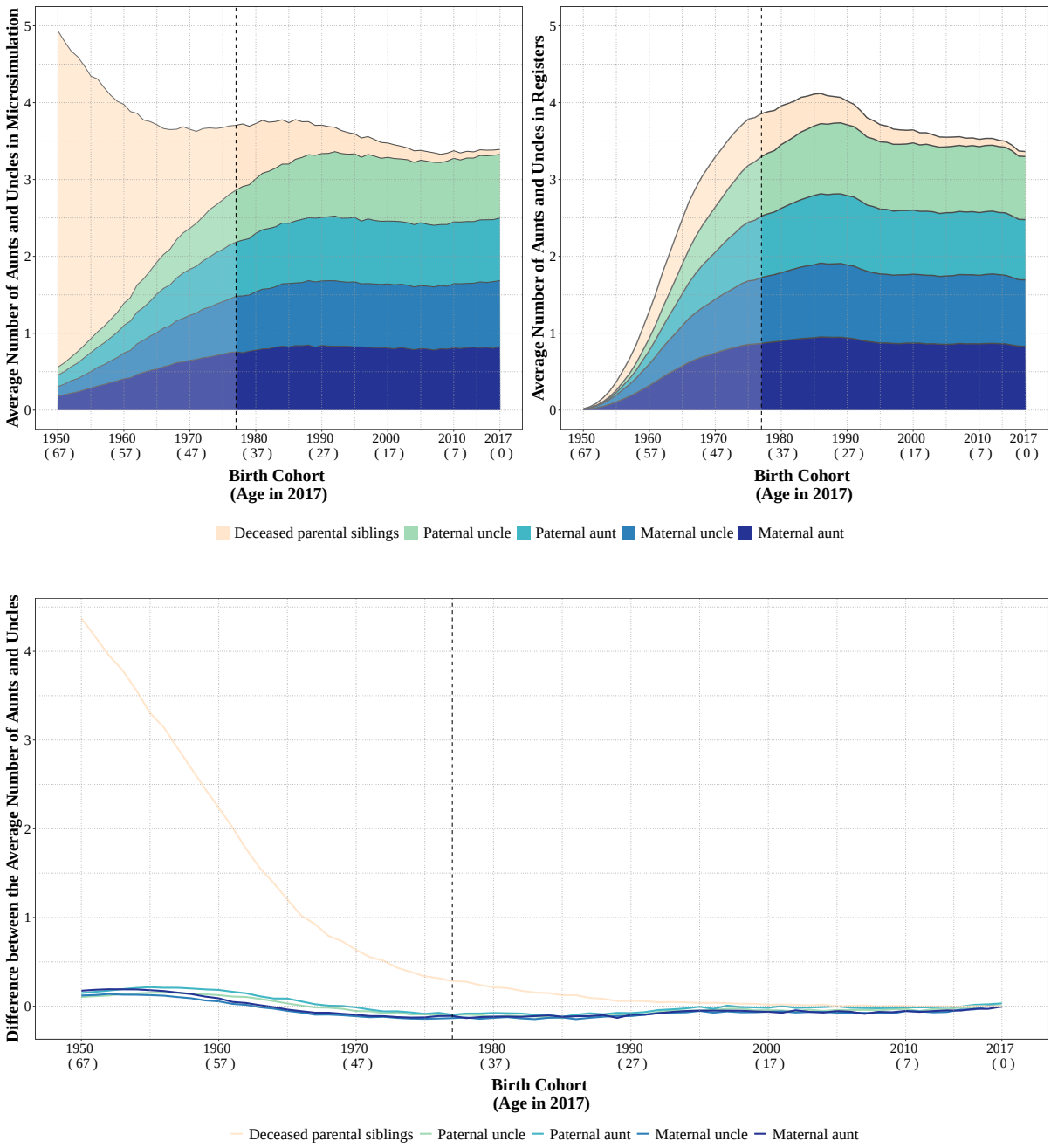

Figure A5: Average number of parent siblings, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.
Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the left side represent the cohorts with incomplete coverage of aunts and uncles due to missing parent-child links.

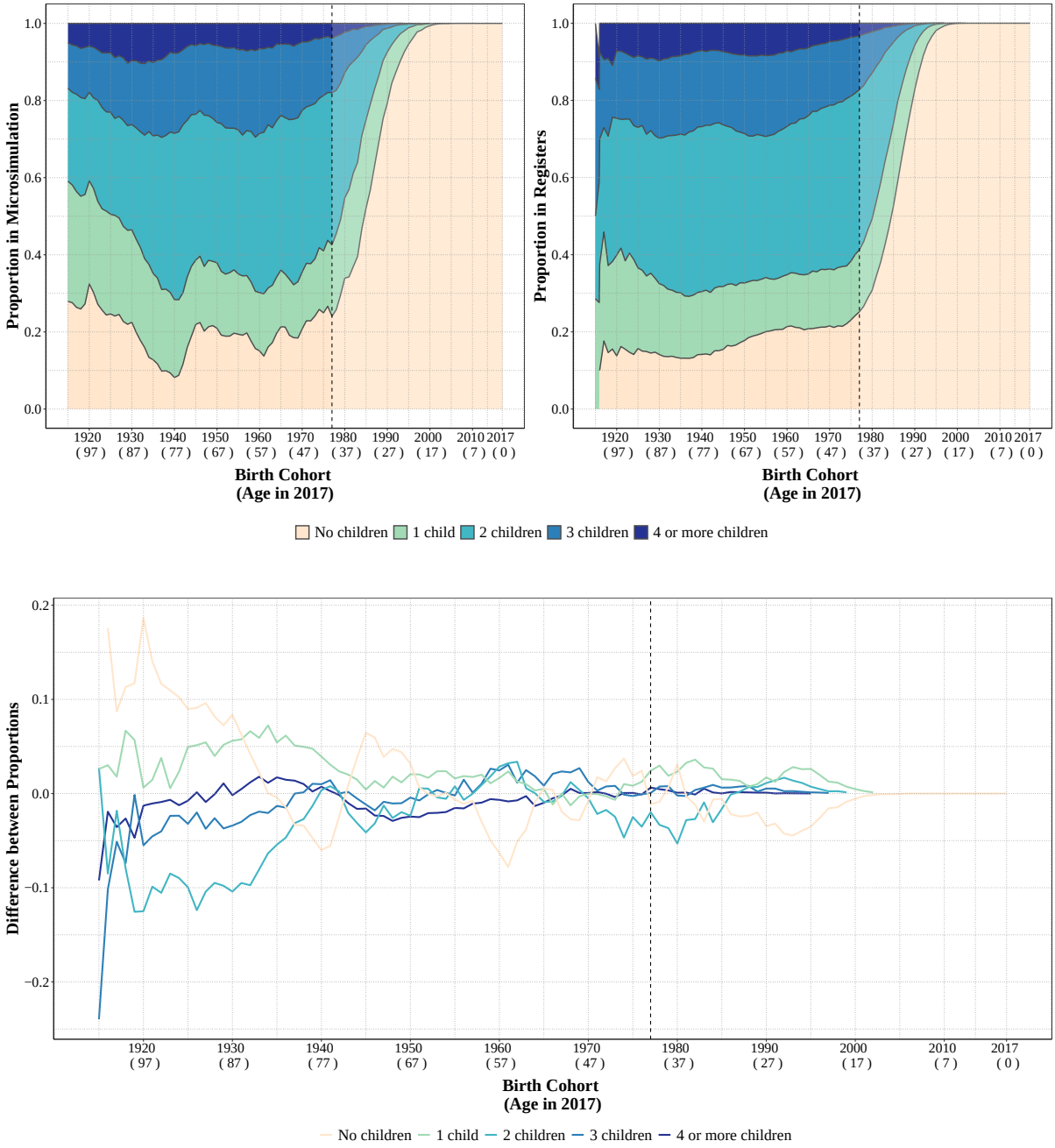

Figure A6: Proportional distribution of the number of living children per man in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.

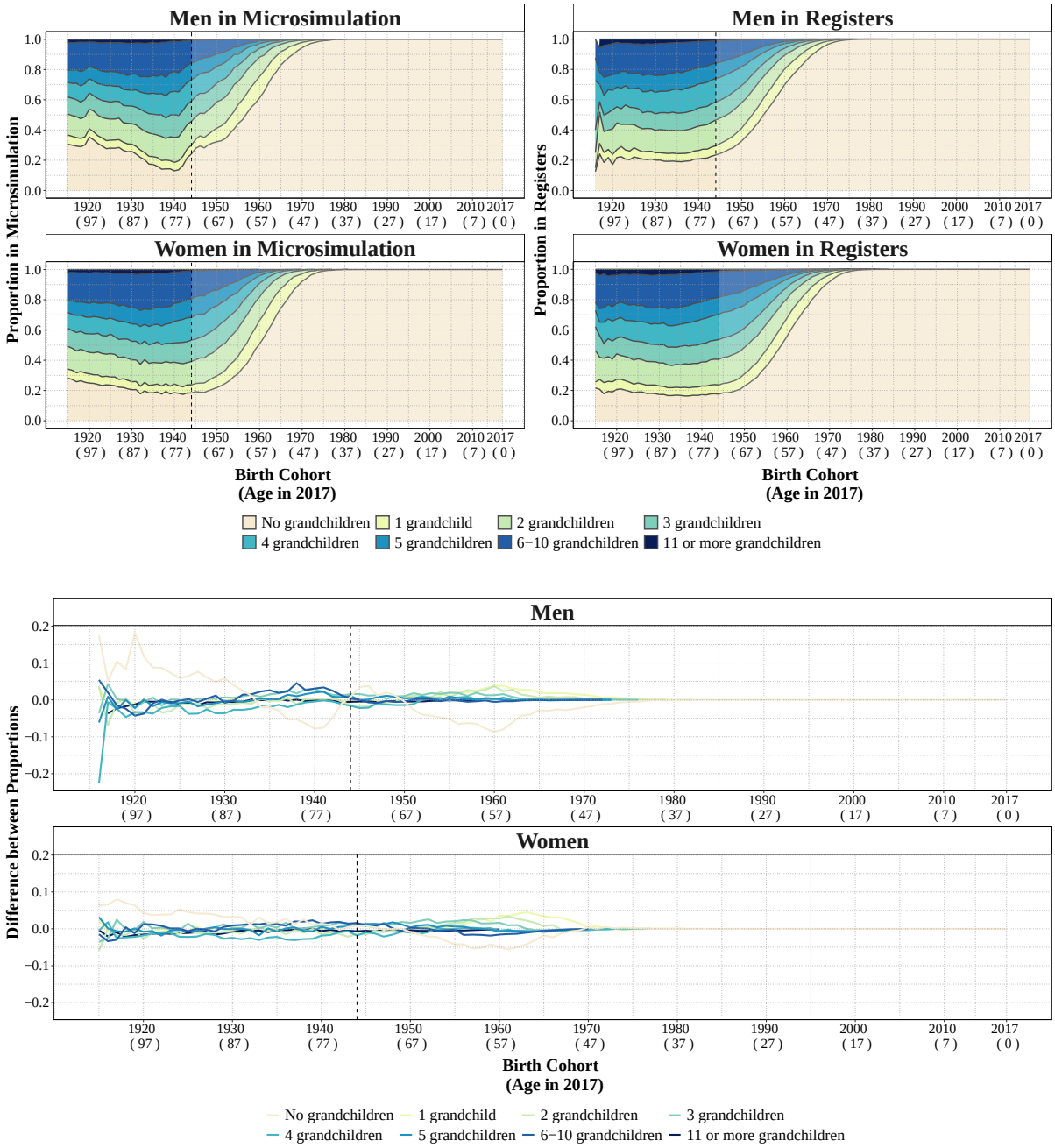

Figure A7: Proportional distribution of the number of living grandchildren in 2017, by sex and birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their grandchildren, as they were not yet born in 2017.

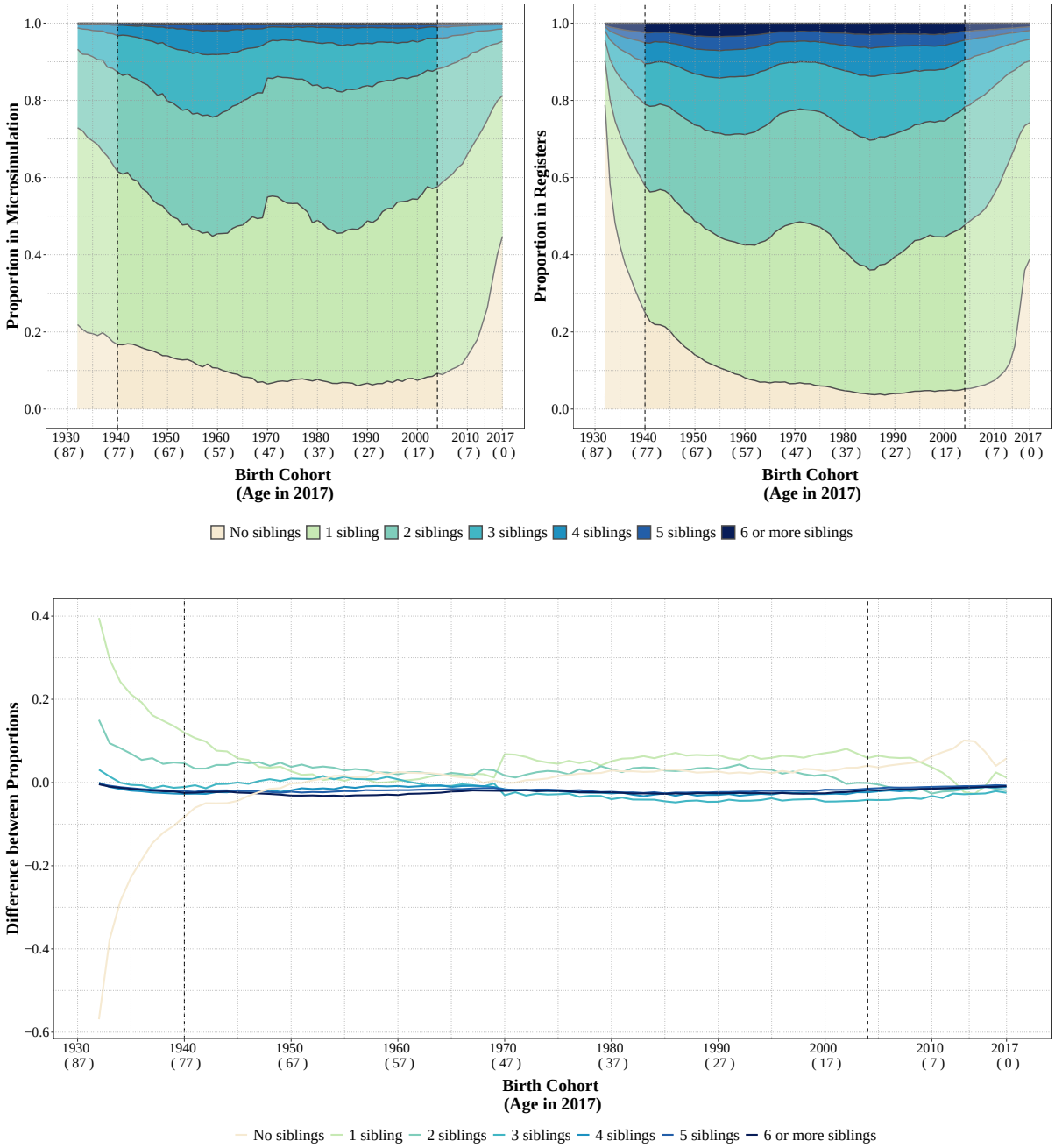

Figure A8: Proportional distribution of the number of siblings (half or full), by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their siblings, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links.

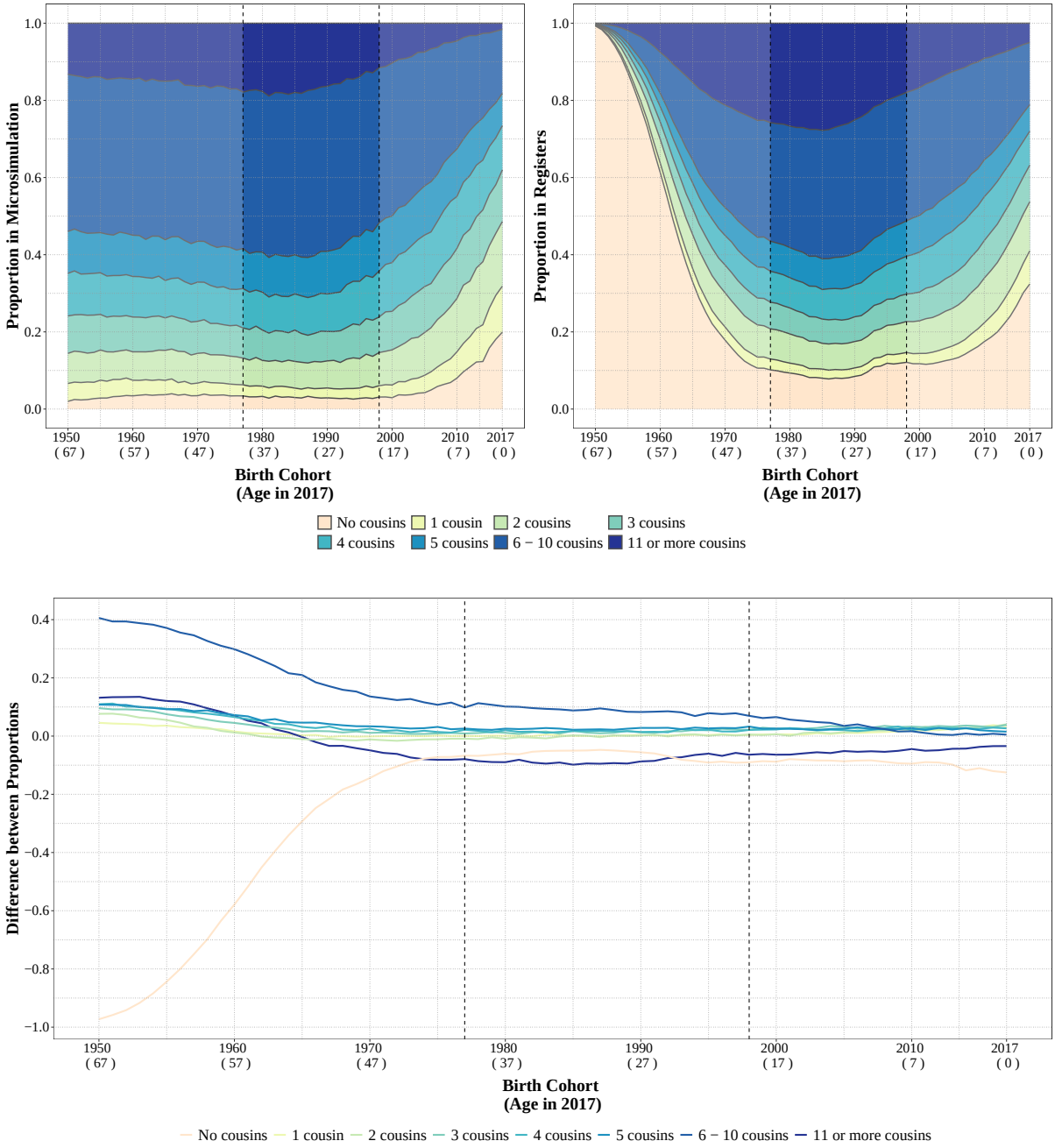

Figure A9: Proportional distribution of the number of cousins, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their nieces and nephews, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links in the registers.

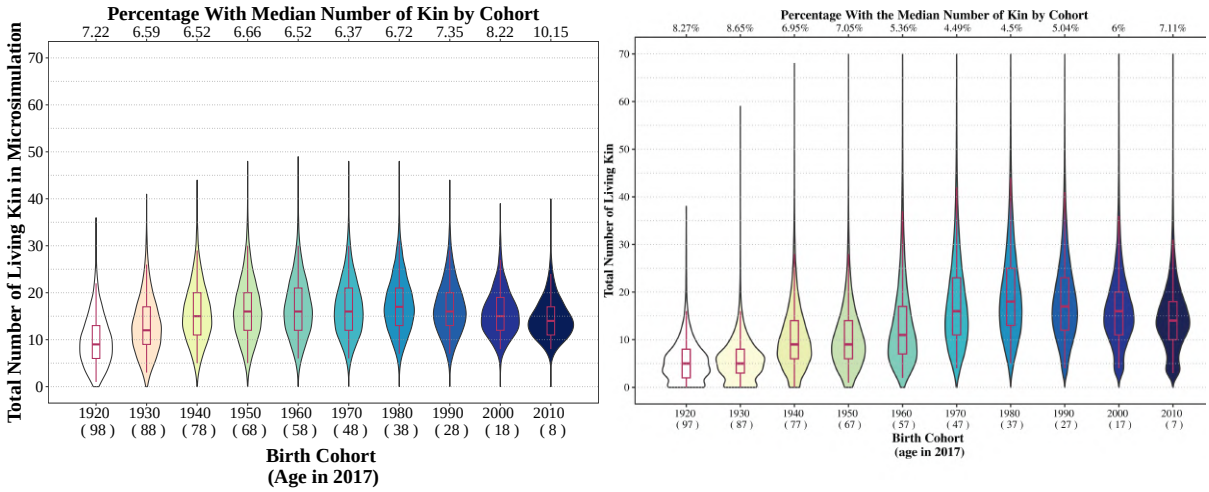

Figure A10: Distribution of the total number of kin and percentage with the median number of kin by birth cohort 1915–2017, estimated from a SOCSIM Microsimulation (left) and the Swedish Registers (right). The figure for the Swedish Registers corresponds to Figure 8 in Kolk et al. (2023), licensed under: CC BY-NC-ND 4.0

References

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.

Caswell, H. (2019). The formal demography of kinship: A matrix formulation. *Demographic Research*, 41:679–712.

Caswell, H. (2020). The formal demography of kinship II: Multistate models, parity, and sibship. *Demographic Research*, 42:1097–1146.

Caswell, H. (2022). The formal demography of kinship IV: Two-sex models and their approximations. *Demographic Research*, 47(13):359–396.

Caswell, H., Margolis, R., and Verdery, A. (2023). The formal demography of kinship V: Kin loss, bereavement, and causes of death. *Demographic Research*, 49:1163–1200.

Caswell, H. and Song, X. (2021). The formal demography of kinship III: Kinship dynamics with time-varying demographic rates. *Demographic Research*, 45:517–546.

Hammel, E. A. (2005). Demographic dynamics and kinship in anthropological populations. *Proceedings of the National Academy of Sciences*, 102(6):2248–2253. Publisher: Proceedings of the National Academy of Sciences.

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.

Human Fertility Collection (2025). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.fertilitydata.org.

Human Fertility Database (2025). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) Available at www.humanfertility.org.

Human Mortality Database. HMD (2025). 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.

Kolk, M., Andersson, L., Pettersson, E., and Drefahl, S. (2023). The Swedish Kinship Universe: A Demographic Account of the Number of Children, Parents, Siblings, Grandchildren, Grandparents, Aunts/Uncles, Nieces/Nephews, and Cousins Using National Population Registers. *Demography*, 60(5):1359–1385.

Margolis, R. and Verdery, A. M. (2019). A Cohort Perspective on the Demography of Grandparenthood: Past, Present, and Future Changes in Race and Sex Disparities in the United States. *Demography*, 56(4).

Mason, C. (2016). *Socsim Oversimplified*. Demography Lab, University of California, Berkeley.

Murphy, M. (2010a). Changes in family and kinship networks consequent on the demographic transitions in England and Wales. *Continuity and Change*, 25(1):109–136.

Murphy, M. (2010b). Family and Kinship Networks in the Context of Ageing Societies. In Tuljapurkar, S., Ogawa, N., and Gauthier, A. H., editors, *Aging in Advanced Industrial States*, pages 263–285. Springer Netherlands, Dordrecht.

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

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. (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.

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.

Wachter, K. W., Blackwell, D., and Hammel, E. A. (1997). Testing the validity of kinship microsimulation. *Mathematical and Computer Modelling*, 26(6):89–104.

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

Figure Titles

- **Figure 1**. Average number of living and dead children per woman in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

  Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.

- **Figure 2**. Average number of living, dead, and unregistered parents in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

  Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the left side represent cohorts with incomplete coverage of parents due to missing parent-child links in the registers.

- **Figure 3**. Average number of living, dead, and unregistered grandparents in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

  Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the left side represent the cohorts with incomplete coverage of grandparents due to missing parent-child links.

- **Figure 4**. Average number of siblings, whether full or half-siblings and by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their siblings, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links.

- **Figure 5**. Average number of cousins, by birth cohort and by type of aunt or uncle, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

  Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their cousins, as they were not yet born in 2017. The shaded areas on the left side represent the cohorts with incomplete coverage due to missing parent-child links in the registers.

- **Figure 6**. Average number of all types of kin in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

- **Figure 7**. Proportional distribution of the number of living children per woman in 2017, by birth cohort, estimated from a SOCSIM microsimulation (top left) and the Swedish Registers (top right), with the difference between estimates (microsimulation minus registers) shown at the bottom.

  Notes: The vertical dashed lines distinguish the cohorts with complete and incomplete kin counts in the registers. The shaded areas on the right side represent the cohorts for which we may not be able to observe all their children, as they were not yet born in 2017.