Validation of the simulation

Frans Willekens

2024-02-20

Contents

List of Figures

List of Tables

1 Introduction

A simulation is considered valid if the simulated (virtual) population is sufficiently close to the real population being simulated, which we call reference population. The validity of the simulation depends on the model, the data and the assumptions. To assess the validity of the simulation, indicators of the virtual population are compared with indicators estimated for the reference population.

The vignette reports on the validity of the virtual population of the United States generated from mortality rates by age and sex and fertility rates by age and parity. The data used in the simulation pertain to 2021. In demography, they are known as *period data*. Validity tests require that at least two conditions are satisfied:

a. The virtual population is sufficiently large to limit the effect of chance. In this vignette, a virtual population of 10,000 individuals is used.

b. Periods of observation of the reference and virtual populations are comparable. Since the summary indicators of the real population are based on observations of the population during a given period, the same period should be used in the simulation. In other words, the virtual population and the real (sample) population must have the same observation window. Otherwise, the summary indicators are not comparable.
c. The methods to esstimate the indicators should be differ between the reference and virtual populations. If that is not feasible, at least they should be comparable.

Four tests are used to assess the validity of the virtual (simulated) population. The first is a comparison of the simulated distribution of ages at death with the age at death distribution implicit in the period life table for 2021, published as part of the Human Mortality Database (https://www.mortality.org). In the second test, the distribution of women by number of children and the distribution by age and number of children are compared with the distributions observed in the Current Population Survey (CPS) of June 2018. For comparison, the virtual population is generated from rates of 2018 (VP2018). In the third test the distribution of number of children ever born by age age mother (VP2018) is compared with the distribution observed in the CPS of June 2018. The fourth test considers children who lost their mother. It compares the age distribution of the children at mother's death in the virtual population with the distribution observed in the Survey of Income and Program Participation (SIPP) of 2021. For comparison, the virtual population based on rates of 2021 is used. In tests two, three and four, simulated data are compared with retrospective survey data. Consequently, the observation window of the virtual population is setnequal to that of the respondents in the CPS (test 2 and 3) and SIPP (test 4). The observation window is from birth to the age at survey (real population). In the virtual population, the same observation window is simulated by starting the simulation at birth and ending it at an age sampled from the age distribution of respondents at survey date.

Note that the standard fertility table disregards mortality, meaning that all individuals complete the reproductive period. Note also that a census or survey do not cover the entire population, but only persons who are alive at Census Day or survey date. Any assessment should account for these peculiarities.

2 Ages at death

Consider the 2021 period death rates of the population of the United States, by single years of age and sex. The data are part of the period life tables, found here for females and here for males^1[It is necessary to first log in to access the data in the Human Mortality Database.]. The data for 2021 are included in VirtualPop as data object *rates*. Consider a virtual population of 10,000 individuals. To generate lifespans that are consistent with the empirical age-specific death rates, the highest age possible must be defined. The maximum age is set to be 120. The death rate for persons aged 110-120 applies to all survivors at ages above 110. The function Lifespan() of VirtualPop is used to simulate lifespans.

```
data(rates,package="VirtualPop")
data(dLH_USA2021_6_2000,package="VirtualPop")
refyear <- attr(dLH,"refyear")
print (refyear)
#> [1] 2021
countrycode <- attr(dLH,"country")
nsample <- nrow(dLH[dLH$gen==1,])
dd <- data.frame(ID=1:nsample)
dd$bdated <- 2000
dd$sex <- sample(x=c(1,2),size=nsample,replace=TRUE,prob=c(0.5,0.5))
dd$sex <- factor(dd$sex,levels=c(1,2),labels=c("Male","Female"))
d <- VirtualPop::Lifespan(data=dd, ASDR=rates$ASDR, mort = NULL)
```

Figure 1 shows the age distribution at death, by sex. The histogram is overlaid with the distribution of ages at death in the period HMD life table of USA in 2021 (dashed line in black). The distribution of simulated ages at death is close to the distribution of life-table ages.

The mean and the variability (standard deviation) of age at death are 73.25 years and 17.73 for males and 80.32 and 14.64 for females. The HMD shows a life expectancy of 73.62 for males and 79.37 for females. The U.S. National Center for Health Statistics published life expectancies of 73.2 and 79.1, respectively (Health Statistics 2022). Differences are due to method and chance. In the conventional life table, used in the HMD, the survival function is a piecewise linear function (Wilmoth et al. 2021, 36). In the simulation it is a piecewise exponential function.

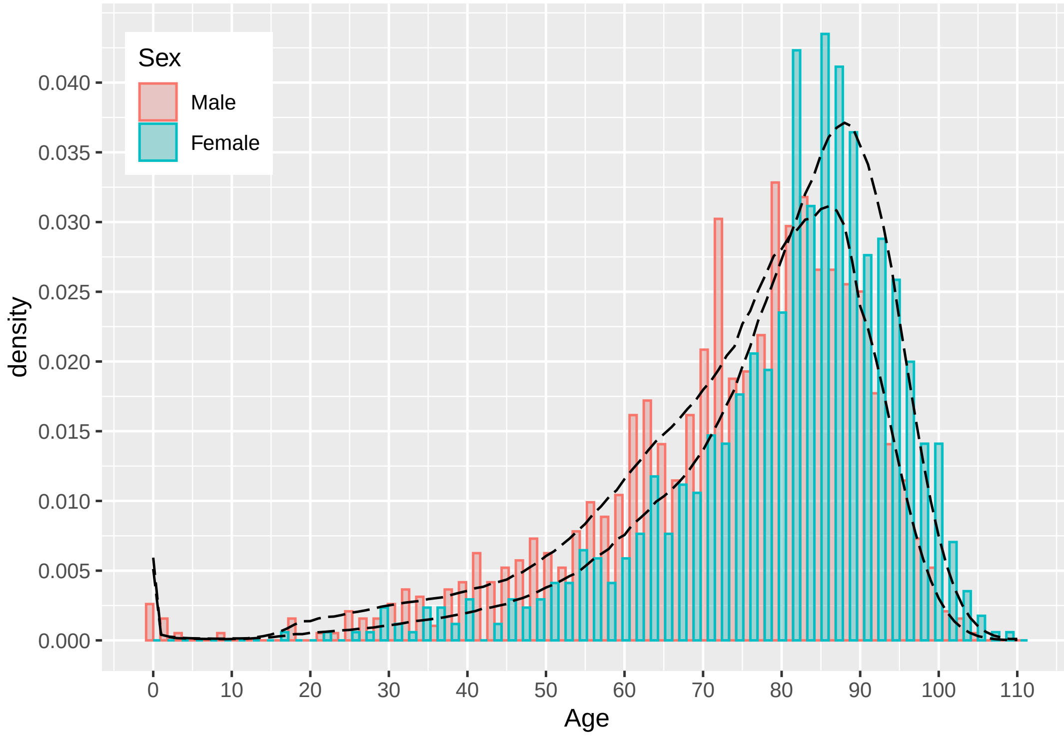

Figure 1: Ages at death in the virtual population, USA, 2021

3 Women with children by number of children ever born

```
dLH <- VirtualPop::CreateVirtual(user,pw_HMD,pw_HFD,
                             country="USA",
                             refyear=2018,
                             ncohort=2000,
                             ngen=2)
dLHnm <- VirtualPop::CreateVirtual(user,pw_HMD,pw_HFD,
                             country="USA",
                             refyear=2018,
                             ncohort=2000,
                             ngen=2,mort=FALSE)
pathSave <- "/Users/frans/VirtualPop_data/"
# fileSave="dLH_USA2018_3_1000.RData"
fileSave <-paste0("dLH_",attr(dLH,"country"),attr(dLH,"refyear"),"_",
                  max(dLH$gen),"_",length(dLH$ID[dLH$gen==1]),".rda")
# save(dLH,file=paste0(pathSave,fileSave))
fileSave <-paste0("dLHnm_",attr(dLHnm,"country"),attr(dLHnm,"refyear"),"_",
                max(dLHnm$gen),"_",length(dLHnm$ID[dLHnm$gen==1]),".rda")
# save(dLHnm,file=paste0(pathSave,fileSave))
```

```
data(dLH_USA2018_3_2000,package="VirtualPop")
data(dLHnm_USA2018_3_2000,package="VirtualPop")
```

To compare number of children in the virtual population with the figures reported in the period fertility table, the effects of mortality should be removed. To remove mortality, the argument mort of the GetGenerations() function is set equal to FALSE (see *Tutorial*). In the absence of mortality, a woman in the virtual population (generation 1) has 1.721 children, on average, lower than the total fertility rate (TFR) of 1.727 reported in the period fertility table (2018; TFR in 2021: 1.662). The TFR of 2918 reported by the National Center for Health Statistics was 1.730 (2021: 1.664) (Osterman et al. 2023, 13). The proportion of women remaining childless is 19.38 percent, a little higher than the 19.76 percent in the period fertility table (2018; 2021: 22.90 percent). In the presence of mortality, women in the virtual population have 1.645 children, on average and 22.02 percent remain childless.

Table 1 shows the distribution of women with children by the number of children they have in the virtual population in the presence of mortality (vp2018m) and in the absence of mortality (vpnm), the period fertility table 2018 (ft2018) and the June 2018 CPS survey (CPS2018). The distributions in the virtual population and the fertility table are close. The difference can be attributed to the effect of mortality and the method used to compute probabilities from rates. Consider the fertility rate of childless women aged 32. The rate is 0.10119. The probability of having a first child within a year is m/(1+0.5m)=0.10119/(1+0.5*0.10119)=0.09632. In the exponential model, the probability is 1-exp[-m]=1-exp[-0.10119]=0.09624. An exponential survival function with constant rate implies a lower transition probability than a linear survival function with uniform distribution of events. The cumulative effect over all ages is a higher childlessness in the piecewise exponential model than in the piecewise linear model.

The distribution of women with children by number of children ever born is similar to that recorded in the Current Population Survey (CPS) 2018. The result is unexpected because the CPS records the number of children ever born by age of mother at survey date (June 2018).

```
p2018m<- round (table (dLH$nch[dLH$gen==1 &dLH$nch!=0])/
              sum(table (dLH$nch[dLH$gen==1 &dLH$nch!=0])),2)
p2018nm<- round (table (dLHnm$nch[dLHnm$gen==1 &dLHnm$nch!=0])/
              sum(table (dLHnm$nch[dLHnm$gen==1 &dLHnm$nch!=0])),2)
p2018m[5] <- sum(p2018m[5:6])
p2018nm[5] <- sum(p2018nm[5:6])
d_ft18 <- c(0.31,0.39,0.19,0.04,0.07)
d_CPS18 <- c(0.30,0.39,0.19,0.08,0.03)
dvp18 <- data.frame(vp2018m=as.numeric(p2018m)[1:5],
                vp2018nm=as.numeric(p2018nm)[1:5],
                ft2018=d_ft18,CPS2018=d_CPS18)
d2 <- knitr::kable(dvp18,
     caption = paste0("Distribution of women with children by ",
```

```
      "number of children ever born"),
     format="latex",align="c",booktabs=TRUE,linesep = "")
kableExtra::add_footnote(d2,
        c("Source: Period fertility table 2018 (HFD) and CPS2018",
        "https://www.census.gov/data/tables/2018/",
        "demo/fertility/women-fertility.html"),
         notation = "none")
```

4 Women by age at censoring and number of children ever born

As an additional validity check, the distribution of number of women in the virtual population by age and number of children ever born is compared with the distribution observed in the CPS of June 2018.

To obtain comparable figures, respondents in the CPS and individuals in the virtual population should be followed during the same segments of life. To meet that requirement, the female members of the first generation are selected, the competing risk of death is omitted, and the CPS censoring scheme is imposed onto the virtual population. In the CPS 2018, 13.5 percent of respondents are interviewed at an age between 15 and 20. 13.9 percent at an age between 20 and 25, etc. The age of interview is the age at censoring. The same age distribution of censoring is imposed onto the virtual population. Individuals are assigned an age group at censoring randomly by sampling a multinomial distribution with parameters the probability distribution of respondents in the CPS of June 2018. The exact ages at censoring (interview) are obtained by assuming a uniform age distribution within a 5-year age interval. It is implemented by sampling a uniform distribution with minimum value 0 and maximum value 5 and adding the result to the minimum age of the selected age group. Once the exact age at censoring is known, the calendar date of censoring is obtained by adding the age at censoring to the age at birth. The following code implements the procedure:

```
# Age at censoring
dLHnm$stop <- NA
# Number of females in CPS by age group
nfemCPS <- c(10294,10607,11476,10889,10727, 9896,12524 )
# Percentage
percCPS <- nfemCPS/sum(nfemCPS)
# Allocate age at censoring to members of virtual population
nbreaks <- c(15,20,25,30,35,40,45,50)
kk <- sample (nbreaks[1:(length(nbreaks)-1)],
          nrow(dLHnm),prob=percCPS,replace=TRUE)
dLHnm$stop  <- kk + runif(nrow(dLHnm),min=0,max=5)
```

To assesses whether the age distribution at censoring in the virtual population is the same as the age distribution at CPS survey, use the following code chunk:

```
nbreaks <- c(15,20,25,30,35,40,45,50)
namagegroup <- vector(mode="character",length=7)
for (i in 1:6)
{ namagegroup[i] <- paste (nbreaks[i],"-",nbreaks[i+1]-1,sep="")
}
namagegroup[7] <- paste (nbreaks[i+1],"-",nbreaks[i+2],sep="")
namagegroup[1] <- "<20"
namagegroup[7] <- ">=45"

# Age distribution at censoring (interview), virtual population, females
age_VP <- cut (dLHnm$stop[dLHnm$sex=="Female"],
                    breaks=nbreaks,include.lowest=TRUE,labels=namagegroup)
n <- table (age_VP)
age_VP <- round (100 * n/sum(n),2)
# Age distribution of respondents at survey date, CPS 2018
names(percCPS) <- names(n)
age_CPS <- round (100*percCPS,2)
c <- data.frame(VP=data.frame(age_VP)$Freq,CPS=age_CPS)
```

Table 1: Age distribution at censoring in CPS and virtual population

	VP	CPS
<20	11.96	13.47
20-24	13.73	13.88
25-29	15.42	15.02
30-34	15.42	14.25
35-39	13.38	14.04
40-44	13.93	12.95
>=45	16.17	16.39

The number of females in the virtual population of 10,000 individuals is 2542. The age $dLHnm\$stop$ is the age at censoring.

The following table shows the number of children ever born, by age of mother, observed by CPS at survey date. The numbers are given for 5-year age groups from 15 to 50. A total of 76,413 women are included in the CPS in June 2018, 13.5 percent was 15-19 years of age at time of survey, 13.9 percent was 20-24, etc. Of those aged 15-19, 96.9 percent had no children at survey date, 2.1 percent has 1 child and 0.8 percent has 2 children. Of those 45-50 at survey, 15.4 percent are childless. More than one third (35.5 percent) has two children.

```r
# ====  Number of children ever born  CPS June 2018    Women, by age   ====
# https://www.census.gov/data/tables/2018/demo/fertility/
#               women-fertility.html#par_list_57
#   table t1.xlsx
nfemCPS <- c(10294,10607,11476,10889,10727, 9896,12524 )
# round (100 *nfemCPS/sum(nfemCPS),2)
#     44.2 percent have 0 children
# 15 to 50 years
everBorn <- matrix (c(96.9,2.1,0.8,0.1,0,0.1,0,
  78.6,14,6,1,0.3,0.2,0,
  54.2,20.4,16.2,6.5,2.1,0.5,0.1,
  33.6,22.3,24.6,12.8,4.4,1.9,0.3,
  20.0,19.2,32.6,17.4,7.3,3.2,0.4,
  15,18.7,34.6,18.6,8.7,3.8,0.7,
  15.4,19.8,35.4,17.3,7.4,3.6,1.2),nrow=7,byrow=TRUE)
everBornTot <- c(44.2,16.8,21.7,10.7,4.3,1.9,0.4)

dCPS <- rbind (everBorn,Allages=everBornTot)
dCPS <- cbind (Females=c(nfemCPS,sum(nfemCPS)),dCPS)
dimnames(dCPS) <- list (AgeGroup=c(namagegroup,"Total"),
       Number_of_children_ever_born_CPS=c("nfemales",0,1,2,3,4,"5-6","7-8"))
attr(dCPS,"refyear") <- 2018
```

Generation 2 consists of children of women of generation 1. In the following code, numbers of children ever born at censoring date are computed.

```r
# ========  Dataframe of mothers of members of generation 2  ========
dfw <- data.frame(ID=dLHnm$ID[dLHnm$gen==1 & dLHnm$sex=="Female"])
dfw$bdated <- dLHnm$bdated[dfw$ID]
dfw$stop <- dLHnm$stop[dfw$ID]
dfw$stopY <- dfw$bdated + dfw$stop
dfw$stopG <- cut (dfw$stop,breaks=nbreaks,include.lowest=TRUE,labels=namagegroup)
# Remove women who die before survey
dfw <- subset(dfw,dLHnm$x_D[dfw$ID]>=dfw$stop)
# Number of women by age at censoring and TOTAL number of children in lifetime
dfw$nch <- dLHnm$nch[dfw$ID]
tab_nch <- addmargins(table (dfw$stopG,dfw$nch))
# For each woman, number of children born before censoring
dfw$nchC <- apply(dfw,1,function(x)
  { idch <- Families::IDch(as.numeric(x["ID"]))
    j <- length(which(dLHnm$bdated[idch] < x["stopY"]))
  })
# Number of women by age group at censoring and number of children at censoring
z <- table (dfw$stopG,dfw$nchC)
```

```r
tab_nchC <- addmargins(table (dfw$stopG,dfw$nchC))
y <- data.frame(table (dfw$stopG))$Freq
tab_nchC <- cbind (nfemales=c(y,sum(y)),tab_nchC)
rownames(tab_nchC)[length(y)+1] <- "Total"
# Percentage
zT <- colSums(z)
tab_p <- rbind (z,Total=zT)
ncheverPerc <- round(100*proportions(tab_p,1),2)
ncheverPerc[1,][all(is.na(ncheverPerc[1,]))] <-
                c(100,rep(0,(ncol(ncheverPerc)-1)))
dvp18b <- cbind(nfemales=c(y,sum(y)),ncheverPerc)
names(dimnames(dvp18b))[2] <-
          "Number_of_children_ever_born_VirtualPopulation"
attr(dvp18b,"refyear") <- 2018
```

The distribution of numbers of children ever born, by age group at censoring, is shown in the following table.

```r
# CPS 2018
y <- knitr::kable(dCPS,
     caption = paste0("Distribution of women by number of children ",
     "ever born, by age (CPS)"),
     format="latex",align="c",booktabs=TRUE,linesep = "")
yy <- kableExtra::add_header_above(y,c("AgeGroup"=1,
       "Number_of_children_ever_born_CPS"=7))
kableExtra::add_footnote(yy,
        c("Source: https://www.census.gov/data/tables/2018/demo/",
         "fertility/women-fertility.html#par_list_57"),
         notation = "none")
```

Table 3: Distribution of women by number of children ever born, by age (virtual population)

AgeGroup	Number_of_children_ever_born_CPS							
	nfemales	0	1	2	3	4	5	6
<20	119	99.16	0.84	0.00	0.00	0.00	0.00	0.00
20-24	141	82.27	14.89	2.13	0.71	0.00	0.00	0.00
25-29	165	63.03	19.39	10.91	4.85	1.82	0.00	0.00
30-34	146	41.78	27.40	21.92	6.85	1.37	0.00	0.68
35-39	129	27.13	29.46	21.71	11.63	4.65	5.43	0.00
40-44	144	29.86	27.78	28.47	11.11	1.39	0.69	0.69
>=45	157	23.57	31.21	28.66	8.92	3.18	2.55	1.91
Total	1001	51.35	22.08	16.68	6.39	1.80	1.20	0.50

Table 2: Distribution of women by number of children ever born, by age (CPS)

AgeGroup	Number_of_children_ever_born_CPS							
	nfemales	0	1	2	3	4	5-6	7-8
<20	10294	96.9	2.1	0.8	0.1	0.0	0.1	0.0
20-24	10607	78.6	14.0	6.0	1.0	0.3	0.2	0.0
25-29	11476	54.2	20.4	16.2	6.5	2.1	0.5	0.1
30-34	10889	33.6	22.3	24.6	12.8	4.4	1.9	0.3
35-39	10727	20.0	19.2	32.6	17.4	7.3	3.2	0.4
40-44	9896	15.0	18.7	34.6	18.6	8.7	3.8	0.7
>=45	12524	15.4	19.8	35.4	17.3	7.4	3.6	1.2
Total	76413	44.2	16.8	21.7	10.7	4.3	1.9	0.4

Source: https://www.census.gov/data/tables/2018/demo/fertility/women-fertility.html#par_list_57

The results of the simulation are relatively close to the observed figures. The simulated and observed distributions of number of children ever born differ for two reasons. First and foremost, the number recorded in the CPS is collected retrospectively and is the outcome of a history of varying demographic rates. In the CPS, young respondents have different age- and parity-specific rates than old respondents when they were young. The two generations experience the first stage of the reproductive career in historical contexts with different social and economic conditions. In the virtual population, the effect of historical context is missing. Age- and parity-specific fertility rates are constant rates collected during a single calendar year (reference year 2019). The second reason is the effect of sampling. A comparison of Tables 3 and 4 reveal three important characteristics of fertility change in recent decades. First, the proportion childless increased. Second, childless women increasingly have their first child between ages 25 and 34. Third, the two-child norm became less manifest.

Differences between numbers of children ever born in the virtual population and the CPS data would be much larger if the simulation did not account for the censoring of observations in the CPS. The relative closeness of the figures in the virtual population and the CPS survey shows the power of simulation and the computational approach. It also justifies the use of virtual populations to gain insight into demographic processes.

5 Age of a child at mother's death

Download data: to compute the age distribution of respondents (Stata file pu2022.dta) etc

```
# install.packages("haven")
library(haven)
pathSTATA <- "/Users/frans/Documents/R/00MAC/SimulABM/HMD_HFD+paper/Paper/Markdown/Famil
dstata <- haven::read_dta(file=paste0 (pathSTATA,"pu2021.dta"))
```

In the Survey of Income and Program Participation (SIPP), respondents were asked a series of questions regarding parental mortality, including whether their biological parents were still alive at the time of the survey, and, if not, the respondent's age at which they died (Scherer, Berchick, and Kreider 2021). United States Census Bureau (2023) reports the age distribution of individuals at mother's death (in 5-year age groups) based on the SIPP 2021 data. For a comparison of SIPP findings and the virtual population, I retrieved the microdata from the SIPP website (https://www.census.gov/programs-surveys/sipp.html) and computed the ages of children at mother's death. The age distribution is shown in Figure 3. To enable a comparison with the age distribution of children at mother's death in the virtual population, data must be constructed that mimic the SIPP data. The data are made comparable in three steps. In a first step, the observation window used in the SIPP (observation starts at birth and ends at survey date) is imposed onto the virtual population by allocating to each individual an age at censoring drawn randomly from the age distribution of respondents in the 2021 SIPP survey, by sex. The distributions of ages of respondents in the SIPP and ages at censoring in the virtual population are shown in Figure 2. In the SIPP, a male respondent is 42.61 years and a female respondent 45.01 years, on average. The standard deviations are 23.98 years and 24.11 years, respectively. In the virtual population, the mean ages at censoring and their standard deviations are the same. The dates of censoring are computed by adding the ages at censoring to the dates of birth of the individuals whose observation is being censored.

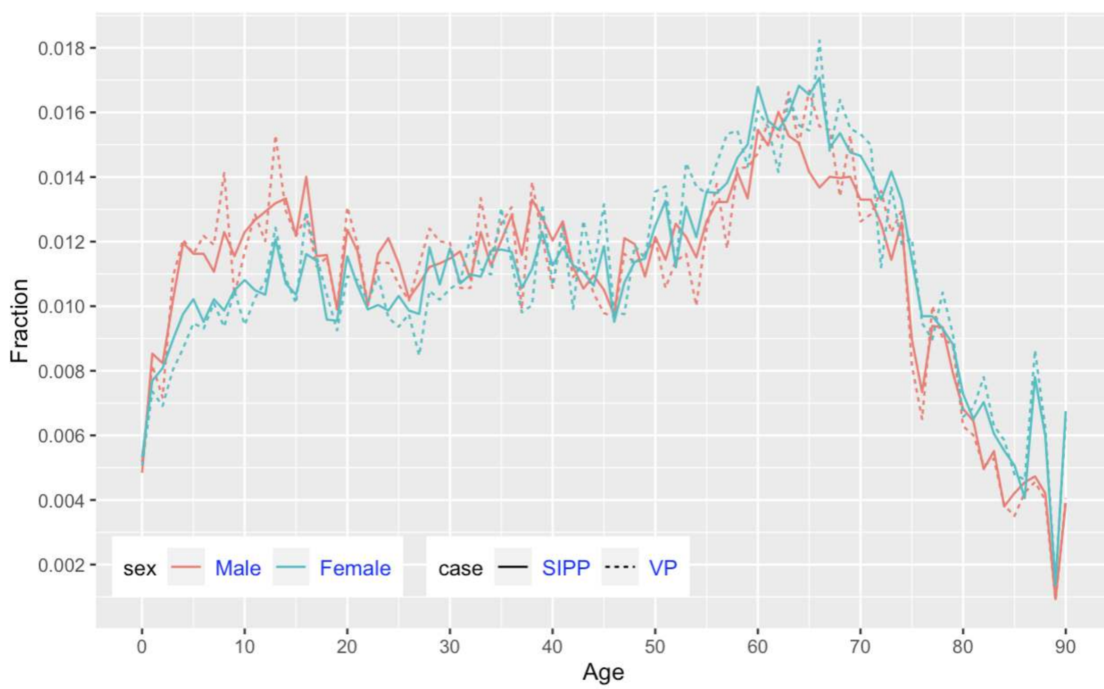

Figure 2: Age distribution of population, at censoring, SIPP 2021 and virtual population USA 2021

In a second step, members of the second generation are selected. They are the children of the members of the first generation of which their mother is part. To be selected, a focal individual must be alive at censoring. In a third step, the mothers of focal individuals are identified and their dates of death retrieved. If death occurs before the focal individual's censoring date, the age of the focal individual at death of the mother is computed.

The age distribution of members of the second generation at death of their mother, provided the death occurs before censoring, is shown in Figure 3 (curve VP_x_C). In the virtual population, children lose their mother at younger ages than in SIPP. Part of the reason is that in the virtual population women have their children at a higher age (2021 period rates) than in the SIPP. If censoring is omitted, deaths of mothers at higher ages are included, causing the age distribution of children at mother's death to shift to the right (curve VP_x_D). The distribution of ages observed in SIPP lie between the two. That is expected because children of SIPP respondents are born earlier than may be concluded based on the mortality and fertility rates of 2021. Since SIPP omits respondents who lose their mother after survey date, while they are included in the virtual population in the absence of censoring, the distribution of ages of children at mother's death includes higher ages.

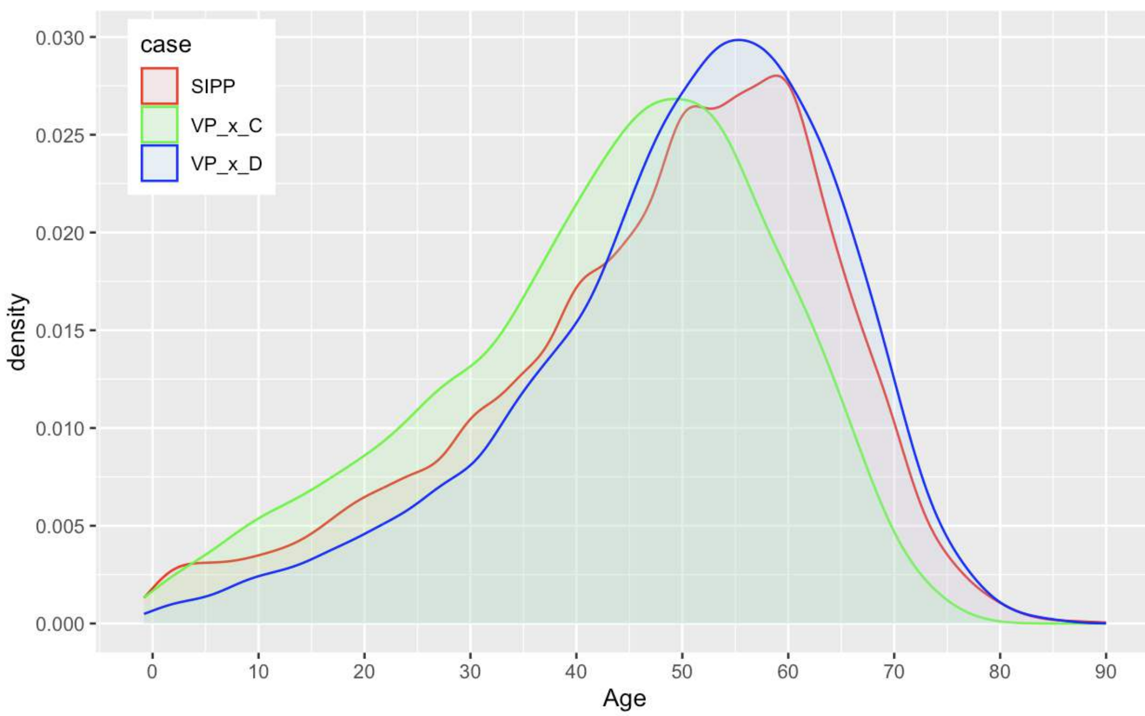

Figure 3: Age distribution of children at death of mother. SIPP2021 and virtual population USA 2021

6 Conclusion

The virtual population generated from mortality rates by age and sex, and fertility rates by age and parity are is an accurate picture of the real population observed during the same period. A comparison of simulated and real populations requires comparable observation windows.

References

Health Statistics, National Center for. 2022. "Life Expectancy in the u.s. Dropped for the Second Year in a Row in 2021." U.S. Department of Health & Human Services, Centers for Disease Control; Prevention, National Center for Health Statistics. https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/20220831.htm.

Osterman, Michelle JK, Hamilton Brady E, Martin Joyce A, Driscoll Anne K, and Valenzuela CP. 2023. "Births: Final Data for 2021." *National Vital Statistics Reports* 72 (1). https://doi.org/10.15620/cdc:122047.

Scherer, Zachary, Edward R. Berchick, and Rose M. Kreider. 2021. "Risk of Parental Death Across the Life Course." SEHSD Working Paper 2020-21 294. Washington, D.C.: U.S. Census Bureau.

United States Census Bureau. 2023. "Losing Our Parents: Recent Statistics on Parent Mortality and Change During the Coronavirus (COVID-19) Pandemic." Census Interactive Gallery. Washington, D.C.: United States Census Bureau. https://www.census.gov/library/visualizations/interactive/losing-our-parents.html.

Wilmoth, John R, Kirill Andreev, Dmitri Jdanov, Dana A Glei, Tim Riffe, C Boe, M Bubenheim, et al. 2021. "Methods Protocol for the Human Mortality Database. Version 6 [26 January 2021]." https://www.mortality.org/File/GetDocument/Public/Docs/MethodsProtocolV6.pdf; University of California, Berkeley,; Max Planck Institute for Demographic Research, Rostock.