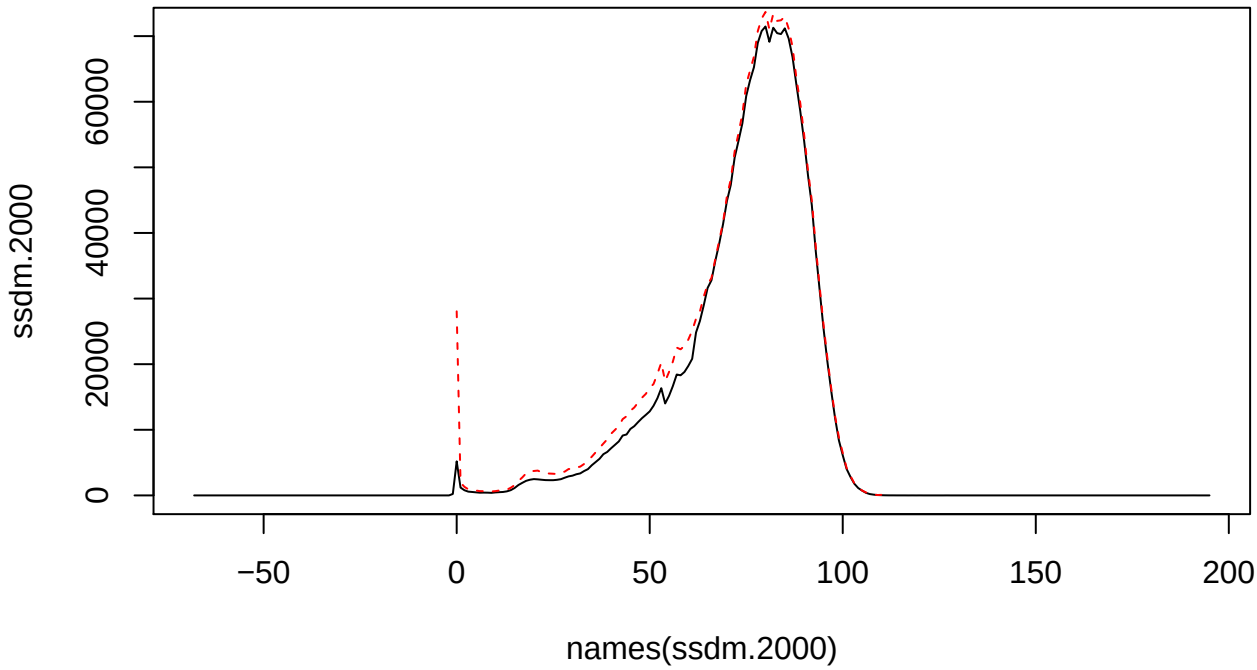
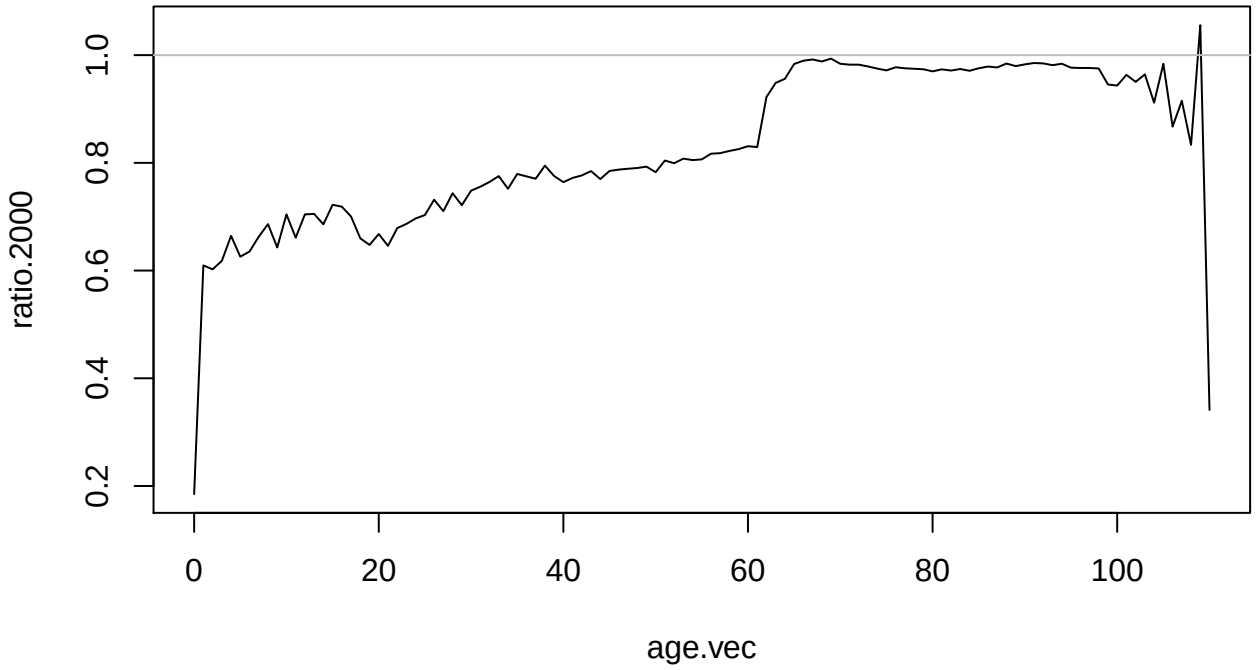
HMD comparison: Social Security Death Masterfile Coverage

Joshua R. Goldstein

Sept. 25, 2016

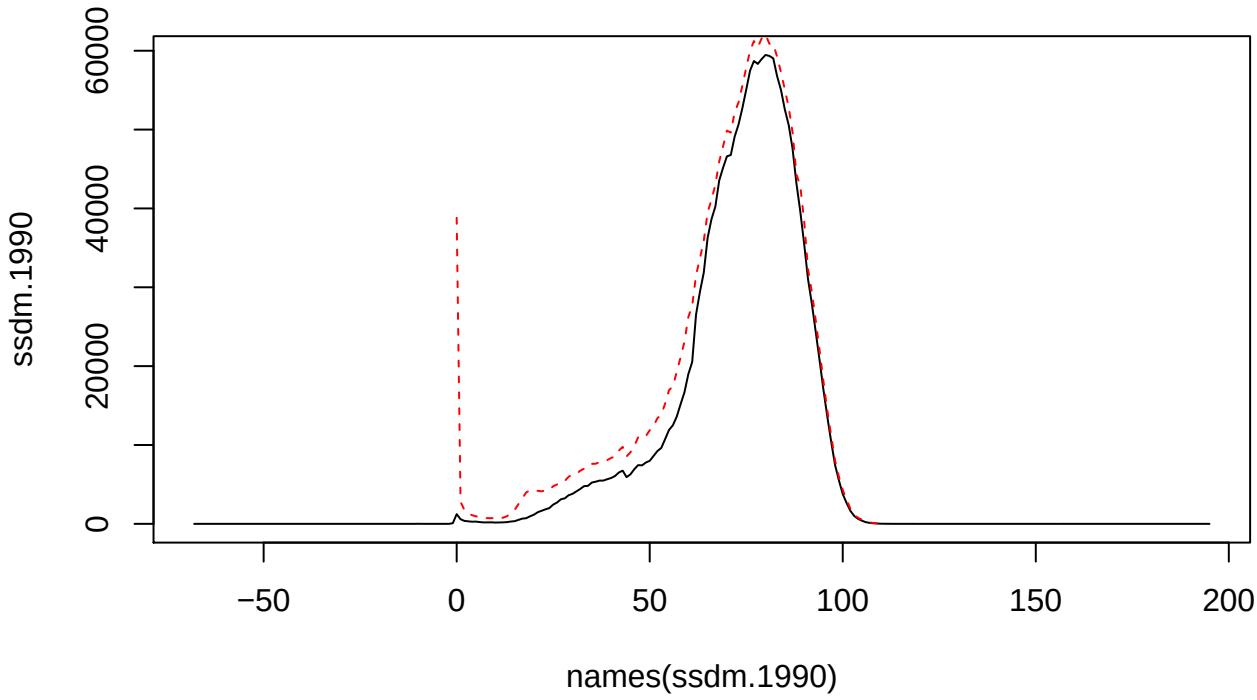
```
## compare total deaths (both sexes) in ssdm with hmd
library(data.table)

## need to get the HMD file (I transferred hear via /90days)
hmd <- fread("Deaths_1x1.txt", skip = 2)

## ssdm table
out <- load("death_freq_tab.RData")
d.tab <- death.freq.tab

## now turn HMD into table format, rows are ages , col are year
year.vec <- names(table(hmd$Year))
age.vec <- 0:110
hmd.tab <- matrix(NA, length(age.vec), length(year.vec))
dimnames(hmd.tab) <- list(age.vec, year.vec)
for (i in 1:length(year.vec))
    {
        this.year <- year.vec[i]
        hmd.tab[,i] <- hmd$Total[hmd$Year == this.year]
    }


## ok, now let's compare in the year 2000

hmd.2000 <- hmd.tab[,"2000"]
ssdm.2000 <- d.tab[,"2000"]
ratio.2000 <- ssdm.2000[names(ssdm.2000) %in% age.vec]/hmd.2000
par(mfrow = c(2,1))
plot(names(ssdm.2000), ssdm.2000, type = "l")
lines(names(hmd.2000), hmd.2000, lty = 2, col = "red")
plot(age.vec, ratio.2000, type = "l")
abline(h = 1, col = "grey")
```

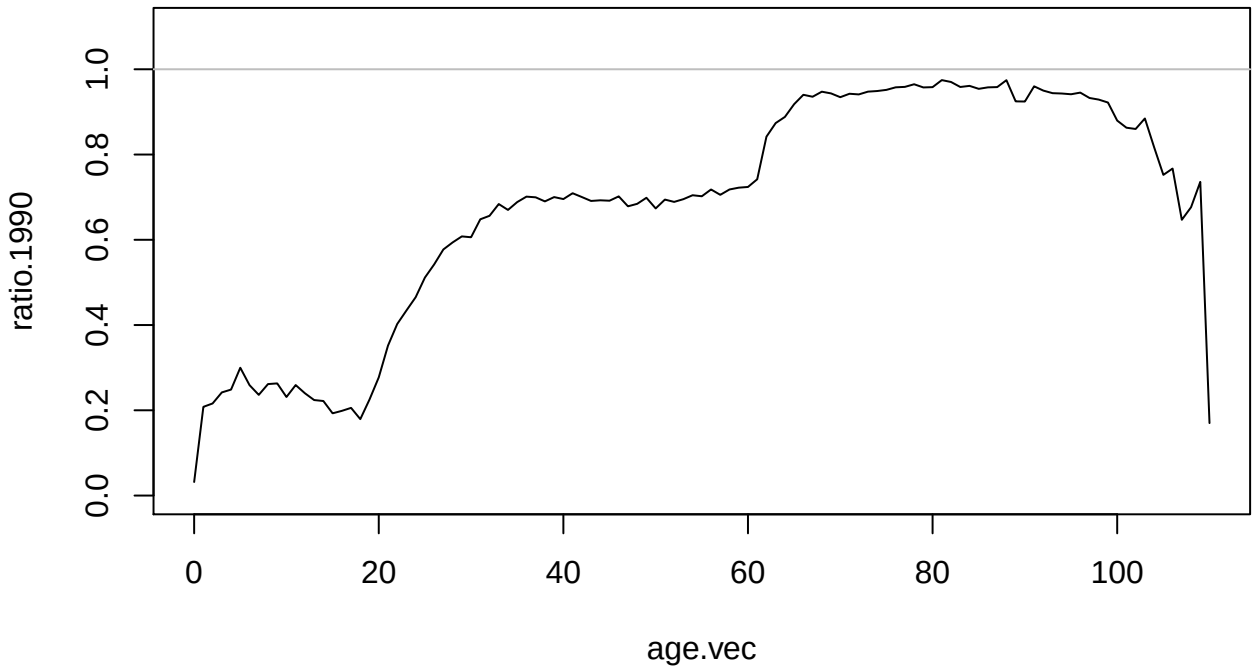





```
## we see here that SSDM has some weird ages of death (e.g -30, or
## 150), but very few cases.  HMD has many more deaths for infants,
## also for young adults, and also at peak ages


hmd.1990 <- hmd.tab[,"1990"]
```

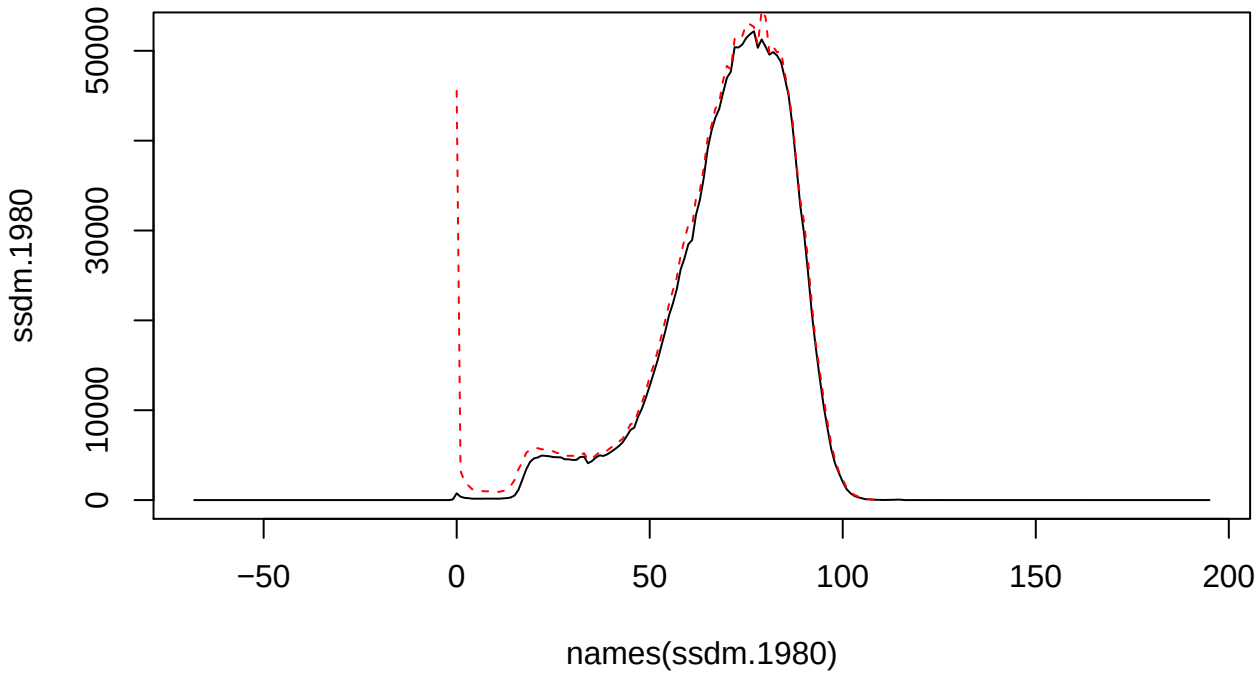
```r
ssdm.1990 <- d.tab[,"1990"]
ratio.1990 <- ssdm.1990[names(ssdm.1990) %in% age.vec]/hmd.1990
par(mfrow = c(2,1))
plot(names(ssdm.1990), ssdm.1990, type = "l")
lines(names(hmd.1990), hmd.1990, lty = 2, col = "red")
plot(age.vec, ratio.1990, type = "l", ylim = c(0, 1.1))
abline(h = 1, col = "grey")
```

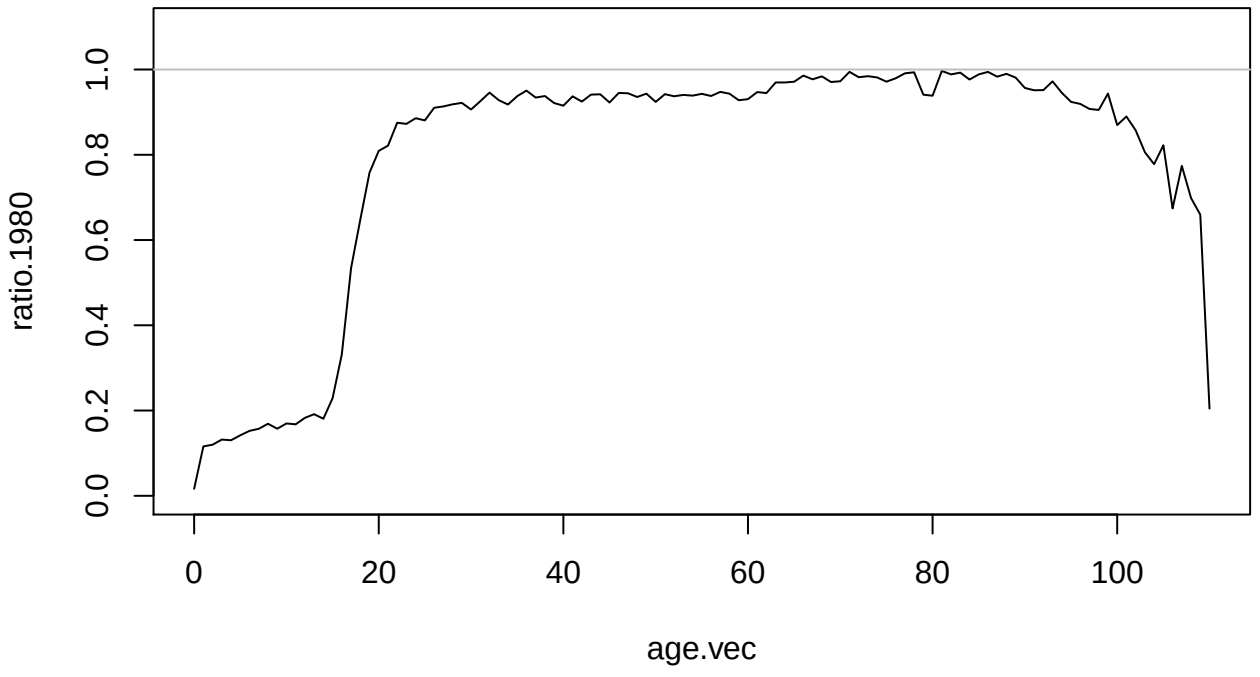





```
## 1990 is worse than 2000

hmd.1980 <- hmd.tab[,"1980"]
ssdm.1980 <- d.tab[,"1980"]
ratio.1980 <- ssdm.1980[names(ssdm.1980) %in% age.vec]/hmd.1980
```

```r
par(mfrow = c(2,1))
plot(names(ssdm.1980), ssdm.1980, type = "l")
lines(names(hmd.1980), hmd.1980, lty = 2, col = "red")
plot(age.vec, ratio.1980, type = "l", ylim = c(0, 1.1))
abline(h = 1, col = "grey")
```



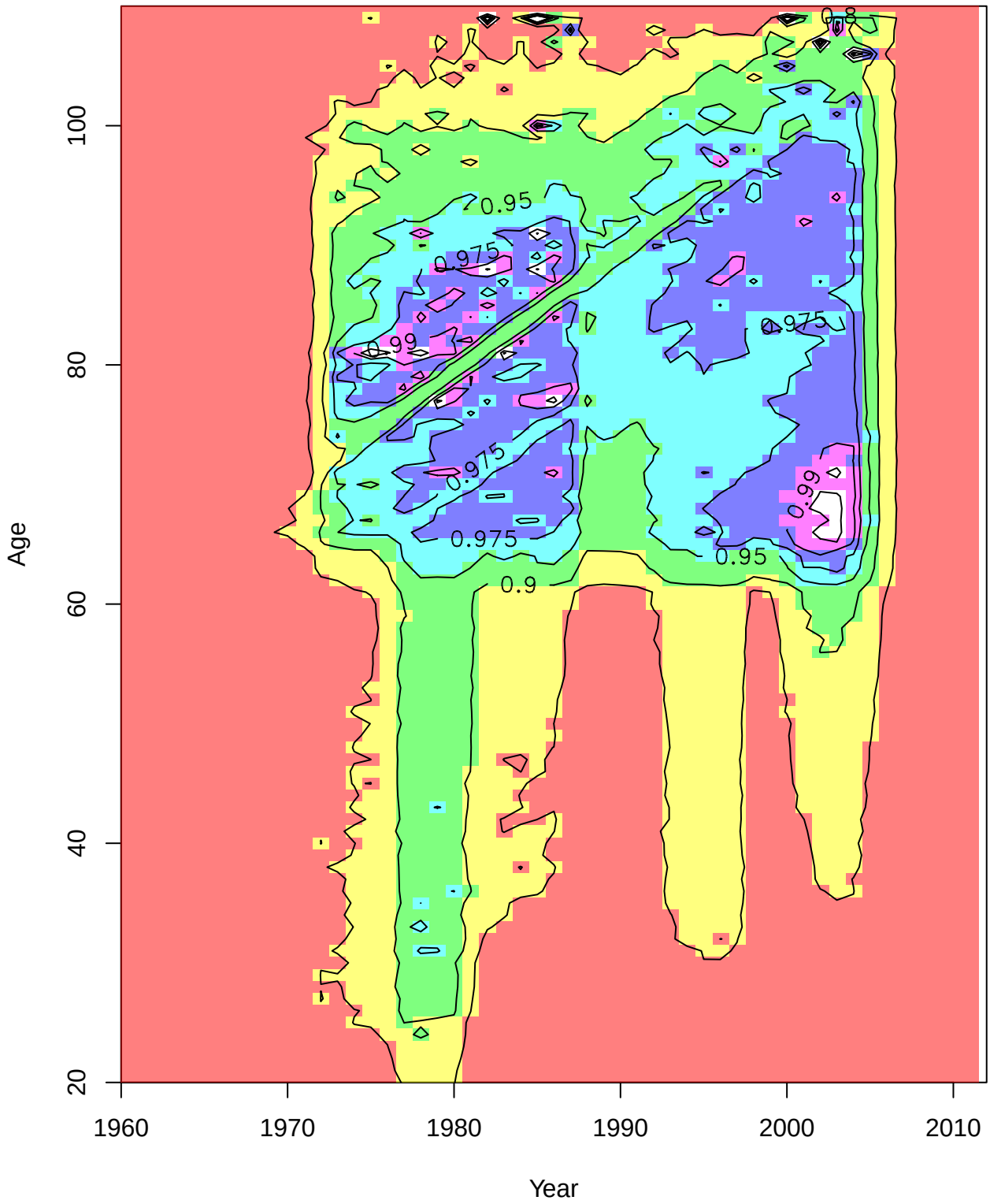


```
## 1980 is again pretty good

## ok, let's try the whole surface

d.tab.conform <- d.tab[rownames(d.tab) %in% age.vec,
                       colnames(d.tab) %in% colnames(hmd.tab)]

ratio.mat <- d.tab.conform / hmd.tab[, colnames(hmd.tab) %in% colnames(d.tab.conform)]

ratio.mat[,"2000"]
```

```
##         0         1         2         3         4         5         6
## 0.1849258 0.6095343 0.6023452 0.6179263 0.6644306 0.6257401 0.6354213
##         7         8         9        10        11        12        13
## 0.6629231 0.6862584 0.6428583 0.7042036 0.6611837 0.7043299 0.7051539
##        14        15        16        17        18        19        20
## 0.6857743 0.7221282 0.7186166 0.7002689 0.6598073 0.6475839 0.6678769
##        21        22        23        24        25        26        27
## 0.6458264 0.6787839 0.6866490 0.6967407 0.7031091 0.7315505 0.7102558
##        28        29        30        31        32        33        34
## 0.7435897 0.7212585 0.7483088 0.7556657 0.7645277 0.7753623 0.7518839
##        35        36        37        38        39        40        41
## 0.7793727 0.7748726 0.7704778 0.7948228 0.7754714 0.7641403 0.7720185
##        42        43        44        45        46        47        48
## 0.7766560 0.7847119 0.7698487 0.7848910 0.7874521 0.7890462 0.7903915
##        49        50        51        52        53        54        55
## 0.7929473 0.7826407 0.8043708 0.7993013 0.8078325 0.8051058 0.8064379
##        56        57        58        59        60        61        62
## 0.8170028 0.8179510 0.8221726 0.8255051 0.8308680 0.8293453 0.9221554
##        63        64        65        66        67        68        69
## 0.9484397 0.9560179 0.9835684 0.9896368 0.9919605 0.9882201 0.9934953
##        70        71        72        73        74        75        76
## 0.9839363 0.9823185 0.9822846 0.9789593 0.9749844 0.9716590 0.9774838
##        77        78        79        80        81        82        83
## 0.9755829 0.9746554 0.9737380 0.9699022 0.9735853 0.9713556 0.9741978
##        84        85        86        87        88        89        90
## 0.9709199 0.9756560 0.9787746 0.9770949 0.9843761 0.9796333 0.9828404
##        91        92        93        94        95        96        97
## 0.9853494 0.9845880 0.9813586 0.9840617 0.9766743 0.9761145 0.9761862
##        98        99       100       101       102       103       104
## 0.9751963 0.9454338 0.9434879 0.9633953 0.9503582 0.9642075 0.9117385
##       105       106       107       108       109       110
## 0.9839213 0.8673159 0.9151377 0.8333333 1.0555556 0.3411765
```

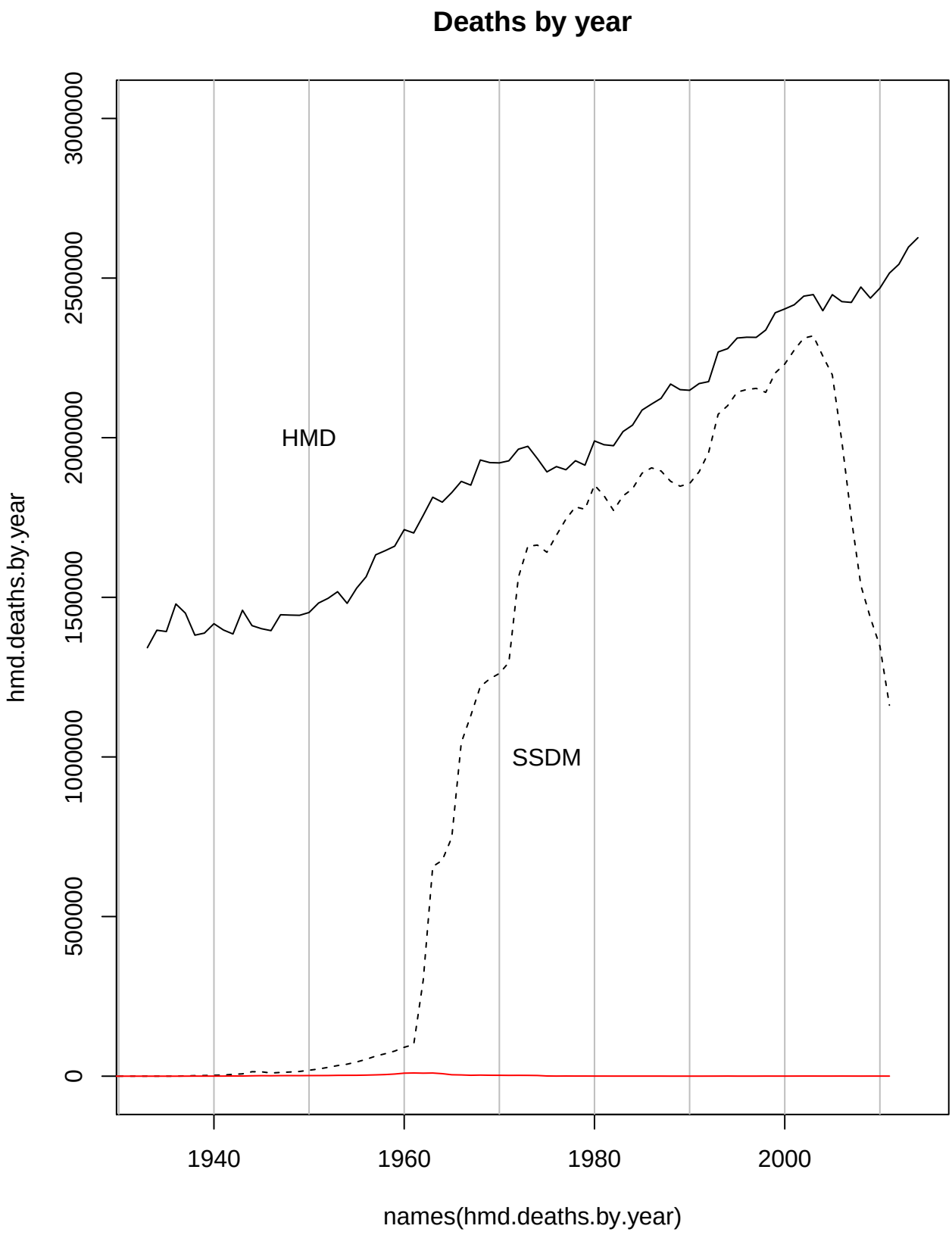
```
my.breaks <- c(0, .8, .9, .95, .975, .99, 1)
my.col <- c("blue", "red", "yellow", "green", "purple", "orange")
par(mfrow = c(1,1))
image(as.numeric(rownames(t(ratio.mat))),
      as.numeric(colnames(t(ratio.mat))),
      t(ratio.mat),
      col = rainbow(n = length(my.breaks) - 1, alpha = .5),
##       col = topo.colors(n = length(my.breaks) - 1, alpha = .5),
##            col = my.col,
      breaks = my.breaks,
```

```
     useRaster = T,
     xlab = "Year", ylab = "Age",
     xlim = c(1960, 2012),
     ylim = c(20, 110))
contour(as.numeric(rownames(t(ratio.mat))),
      as.numeric(colnames(t(ratio.mat))),
      t(ratio.mat),
      labcex = 1,
      vfont = c("sans serif", "bold"),
      add = T,
      levels = my.breaks)
```



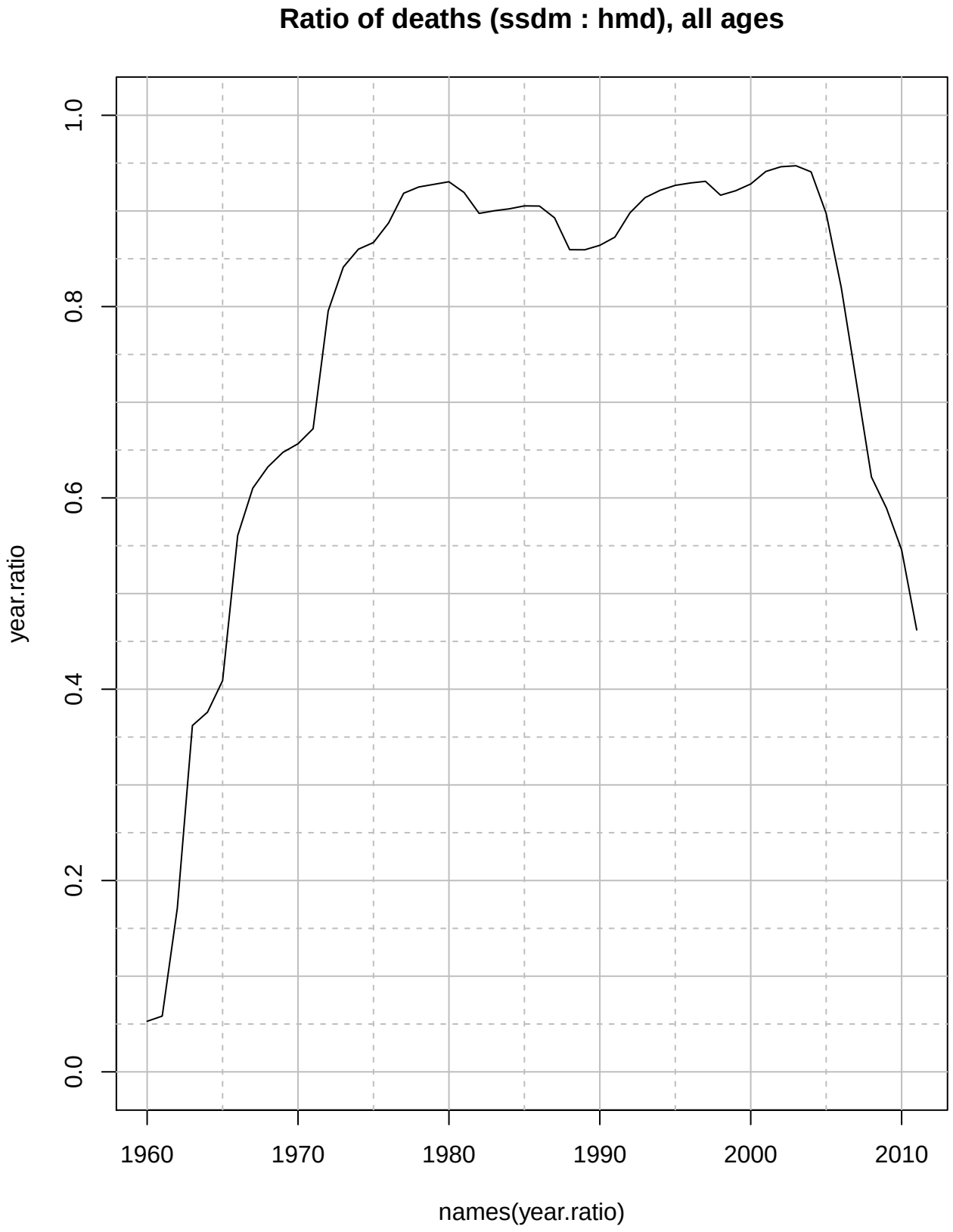

```
## We see that we have coverage over 90% and mostly over 95% for ages
## 65+ between 1975 and 2005. This is very good news.



############ now let's do total deaths by year
```

```r
hmd.deaths.by.year <- apply(hmd.tab,2, sum)
ssdm.deaths.by.year <- apply(d.tab,2, sum)

plot(names(hmd.deaths.by.year), hmd.deaths.by.year, type = "n",
     ylim = c(1, 3*10^6), log = "")
abline(v = seq(1900, 2010, 10), col = "grey")
lines(names(hmd.deaths.by.year), hmd.deaths.by.year)
lines(names(ssdm.deaths.by.year), ssdm.deaths.by.year, lty = 2)
title("Deaths by year")
text(1950, 2 * 10^6, "HMD")
text(1975, 1 * 10^6, "SSDM")

## Bad ages by year

ssdm.age.vec <- as.numeric(rownames(d.tab))
s <- ssdm.age.vec > 110 | ssdm.age.vec < 0 | is.na(ssdm.age.vec)
bad.ssdm.deaths.by.year <- apply(d.tab[s,], 2, sum)
lines(names(bad.ssdm.deaths.by.year),
      bad.ssdm.deaths.by.year, col = "red")
```



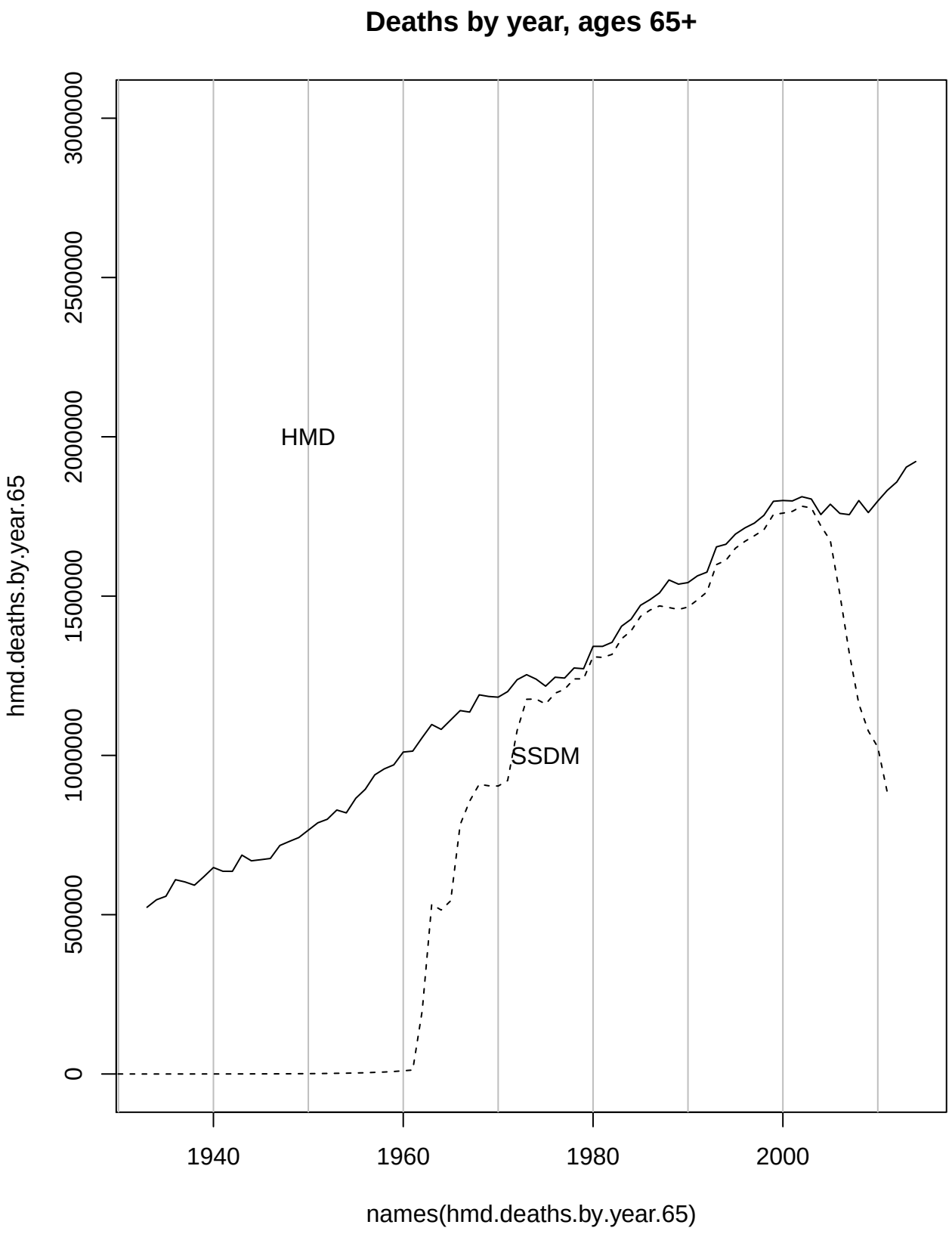

```
## We see that total coverage is good after 1975 and also that "bad
## ages" make no difference at all.
```

```
## now plot 1960 to 2011

year.ratio <- ssdm.deaths.by.year[paste(1960:2011)]/
    hmd.deaths.by.year[paste(1960:2011)]

plot(names(year.ratio), year.ratio, type = "n",
     ylim = c(0, 1))
abline(h = seq(0, 1, .1), col = "grey")
abline(h = seq(0, 1, .1) + .05, col = "grey", lty = 2)
abline(v = seq(1960, 2010, 10), col = "grey")
abline(v = seq(1960, 2010, 10) + 5, col = "grey", lty = 2)
lines(names(year.ratio), year.ratio)
title("Ratio of deaths (ssdm : hmd), all ages")
```



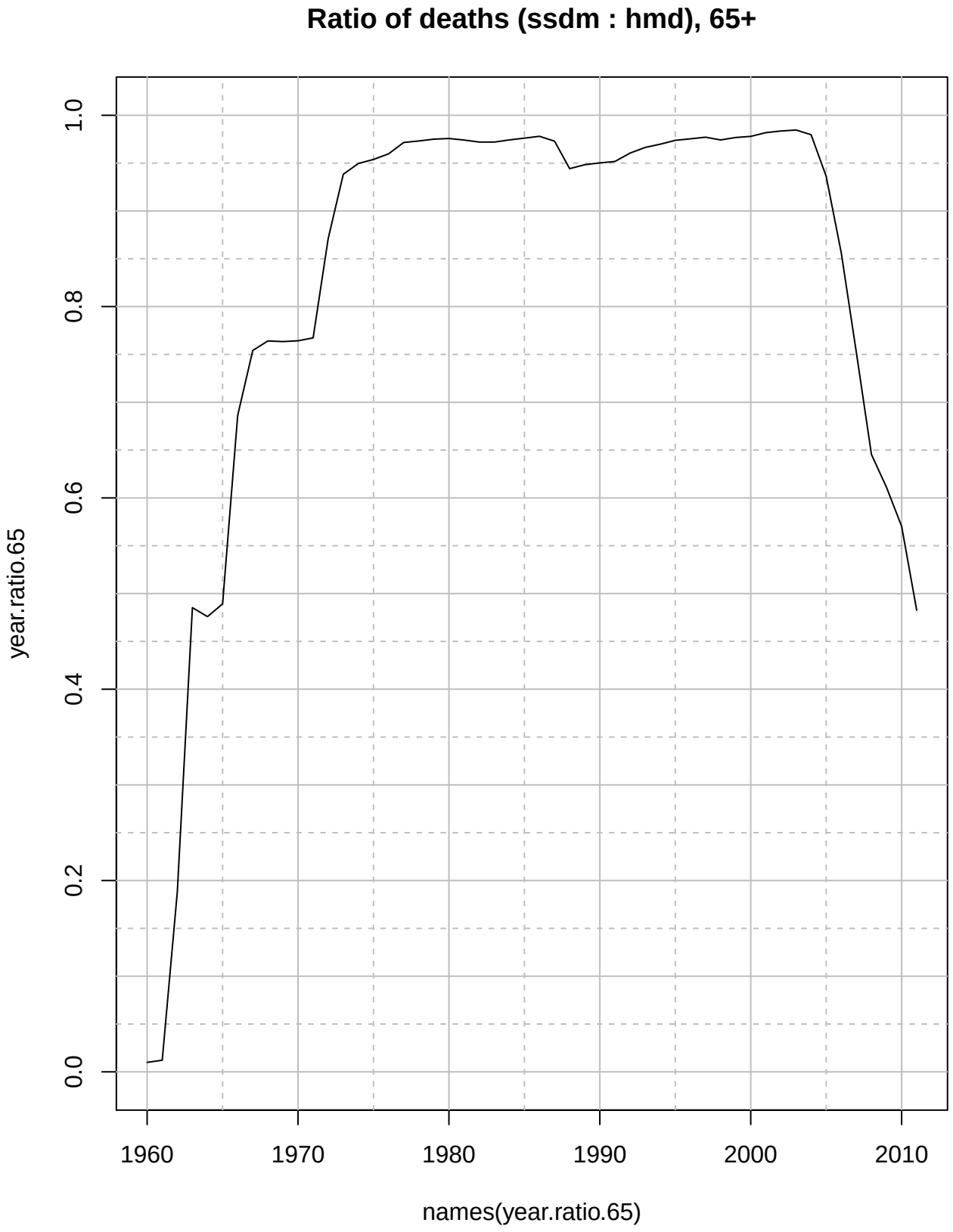

```
########### now let's redo only for deaths over age 65

hmd.tab.65 <- hmd.tab[paste(65:110),]
```

```r
d.tab.65 <- d.tab[paste(65:110),]

hmd.deaths.by.year.65 <- apply(hmd.tab.65, 2, sum)
ssdm.deaths.by.year.65 <- apply(d.tab.65,2, sum)

plot(names(hmd.deaths.by.year.65), hmd.deaths.by.year.65, type = "n",
     ylim = c(1, 3*10^6), log = "")
abline(v = seq(1900, 2010, 10), col = "grey")
lines(names(hmd.deaths.by.year.65), hmd.deaths.by.year.65)
lines(names(ssdm.deaths.by.year.65), ssdm.deaths.by.year.65, lty = 2)
title("Deaths by year, ages 65+")
text(1950, 2 * 10^6, "HMD")
text(1975, 1 * 10^6, "SSDM")
```




```
## now plot 1960 to 2011

year.ratio.65 <- ssdm.deaths.by.year.65[paste(1960:2011)]/
```

```
    hmd.deaths.by.year.65[paste(1960:2011)]

plot(names(year.ratio.65), year.ratio.65, type = "n",
     ylim = c(0, 1))
abline(h = seq(0, 1, .1), col = "grey")
abline(h = seq(0, 1, .1) + .05, col = "grey", lty = 2)
abline(v = seq(1960, 2010, 10), col = "grey")
abline(v = seq(1960, 2010, 10) + 5, col = "grey", lty = 2)
lines(names(year.ratio.65), year.ratio.65)
title("Ratio of deaths (ssdm : hmd), 65+")
```




```
## Bottom line is that coverage is 95% + between 1975 and 2005
```

Bottom line is that coverage is above 95% between 1975 and 2005.

One could extend this by imputing sex to names, and breaking down coverage by male and female.

One should also be slightly concerned about cohort artifacts (it looks like there is some diagonal striping in the figure.)