Cleaning_IDEIA

Natalia Tosi

10/20/2021

1. Cleaning Brazil Data

```
#Importing raw data
d <- read_csv("DATA/Agregador_popularity_BR_IDEIA.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   day = col_double(),
##   year = col_double(),
##   Date = col_date(format = ""),
##   Positive = col_double(),
##   Negative = col_number(),
##   `DK-Neutral` = col_number(),
##   Neutral = col_number(),
##   Sum = col_double(),
##   Validation_Institute = col_logical()
## )
## i Use `spec()` for the full column specifications.
```

```
# d <- read_sheet("https://docs.google.com/spreadsheets/d/1ryF9r-kQdu3QRbRynJuI6xHOKofkHb6cZlWrngZiHzU/

dd <- subset(d, President != "Figueiredo")

#Get rid of spaces in pollster names
d$Institute <- gsub("\\s","\\.", d$Institute, perl = T)

#Use short presidential names, and rder factors cronologically
d$PresidentS <- factor(toupper(d$President),levels = c("FIGUEIREDO","SARNEY",
                                                       "COLLOR", "FRANCO", "CARDOSO",
                                                       "LULA", "DILMA", "TEMER",
                                                       "BOLSONARO"))

#Check for missing data in relevant vars
tmp <- apply(is.na(subset(d, select = c(Date,President,PresidentS,Positive,Institute))),
             2,sum)
if(sum(tmp) > 0){cat("Attention! Data missing in source\n"); print(tmp)}

d$Date <- as_date(d$Date)
```

```
d <- d[sort(as.character(d$Date), index.return = TRUE)$ix,] #sort by date
d$Q <- paste(substr(d$Date,1,5), quarters(d$Date), sep="") #quarter indicator
d$Q <- gsub("Q", "", d$Q)
d$M <- substr(d$Date, 1, 7)#month indicator
d$raw.date <- NULL

# Save table
save(d, file = "R/popularity_raw_BR.RData")
```

2. Functions

```
observations <- wcalcdiagnosticsQ <-  wcalcdiagnosticsM <- list()

#source("R/Extract.r") #load stimsons "extract" function (downloaded from internet)

source("https://raw.githubusercontent.com/nataliatosi/nataliatosi/main/Agregador_BR/Formulas_Agregador.
#source("https://raw.githubusercontent.com/nataliatosi/nataliatosi/main/Agregador_BR/R/Extract.r")
```

3. Setting

3.1. Create empty dataframe for merging results later

```
#### Merge estimates into d dataset for plotting ####
Ms <- expand.grid(1980:2017, sprintf("%02.0f", 1:12))
Ms <- sort(paste(Ms[,1],Ms[,2], sep = "-"))
Qs <- gsub("-01|-02|-03","-1",
           gsub("-04|-05|-06","-2",
                gsub("-07|-08|-09","-3",
                     gsub("-10|-11|-12","-4", Ms))))
MQ <- data.frame(M = Ms,Q = Qs)

#define a month
m1 <- 365/12
```

4. Prepare data from stimsons' wcalc

```
### The simple averaging approach ###################
ms <- my.averageM(d)  #by month
qs <- my.averageQ(d)  #by quarter

#For WCalc, start by saving the info of which pres to use into data
#this is to make sure only "incoming" presidents are used when
#there are more than one per time period
#this should not affect monthly latent estimates
#but definitely affects quarterly
```

Table 1: Observation by pollster in dataset

Varname	n
Datafolha	207
IBOPE	172
IDEIA	86
Sensus	76
Gallup	63
DataPoder360	48
Vox	46
IPSOS	45
IPESPE	42
MDA	31
Atlas	21
Quaest	10
IBPAD	7
Parana	7
Offerwise	6
FSB/Veja	4

```
d <- merge(d, subset(ms, select = c(M,presUsed)), by = "M", all.x = T)
d$useM <- d$PresidentS == d$presUsed

d$Varname <-  gsub("\\s","", d$Institute)
d$Index <- d$Positive
```

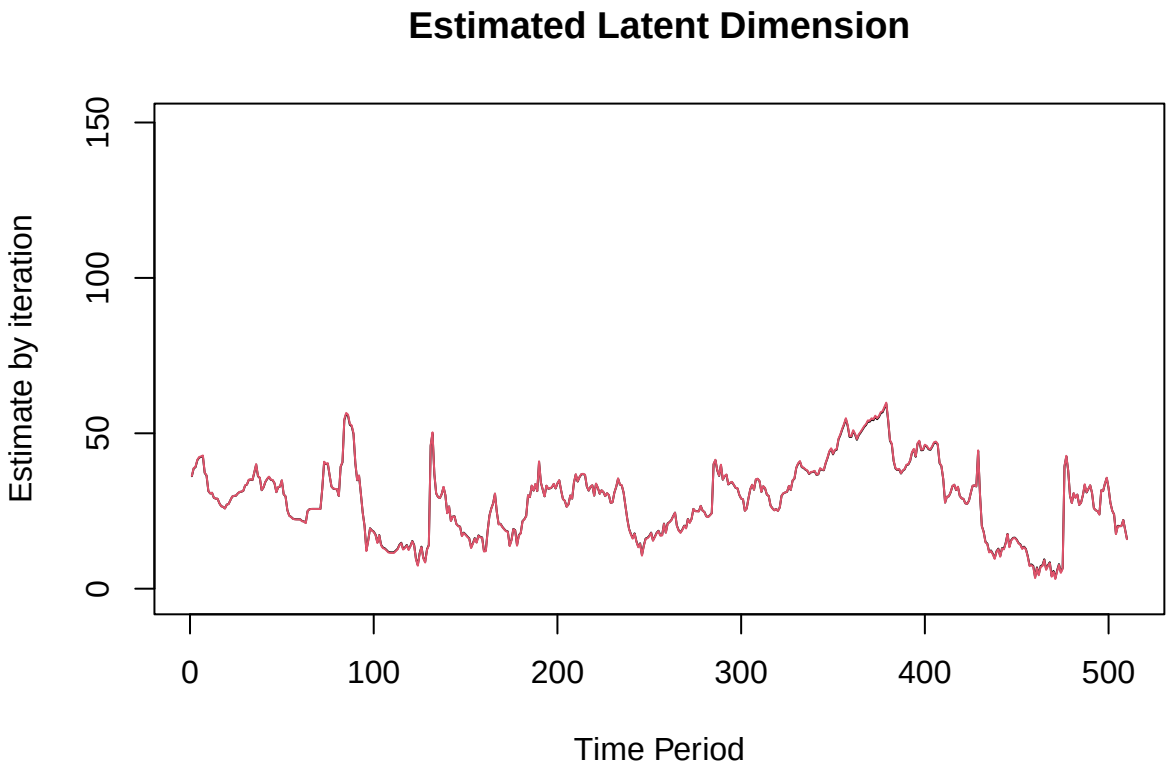
```
ds <- d %>%
  select (Varname, Date, Index, PresidentS, Q, M, useM)

obs_pollster <- ds %>%
  count(Varname) %>%
  arrange(desc(n))

kable(obs_pollster, caption = "Observation by pollster in dataset\n")
```

5. WCALC Monthly

```
ds2 <- ds %>%
  filter(useM == T)

wcalc.Mraw <- extract(varname = ds2$Varname,
                      date = ds2$Date,
                      index = ds2$Index,
                      unit = "M")
```

```
## [1] "Estimation report:"
## [1] "Period: 1979 5  to 2021 10 510  time points"
## [1] "Number of series:  16"
```

```
## [1] "Number of usable series:  16"
## [1] "Exponential smoothing:  TRUE"
## [1] "Iteration history: Dimension  1"
## [1] " "
## [1] "Iter Convergence Criterion Reliability Alphaf Alphab"
```




```
## [1] "1         0.0181     0.001      0.934 0.7311 0.8381"
## [1] "2         0.0019     0.001      0.934 0.7303 0.84"
## [1] "3         3e-04     0.001      0.935 0.7301 0.8401"
## [1] " "
## [1] "Eigen Estimate  1.28  of possible  1.39"
## [1] "  Percent Variance Explained:  92.06"
## [1] " "
## [1] "Final Weighted Average Metric:  Mean:  28.72  St. Dev:  11.9"
```

```
wcalcdiagnosticsM[["brazil"]] <- summary(wcalc.Mraw)
```

```
## Variable Loadings and Descriptive Information: Dimension 1
## Variable Name Cases Loading    Mean  Std Dev
##         Atlas     17    0.5537 28.17647  4.07626
##     Datafolha    174    0.9853 33.38218 18.59332
##  DataPoder360     27    0.9679 22.48148 12.72318
##      FSB/Veja      4    0.5527 32.50000  2.29129
##        Gallup     63    0.9903 32.66667 17.33608
```

```
##         IBOPE    143    0.9848 36.45221 18.16710
##         IBPAD      6   -0.6674 34.66667  3.19722
##         IDEIA     32    0.9379 32.88021  6.08294
##        IPESPE     33    0.9700 28.77273 10.86772
##         IPSOS     45    0.9763  5.55556  5.00469
##           MDA     29    0.9717 25.34483 16.42176
##     Offerwise      6    0.0613 34.00000  2.30940
##        Parana      7    0.5842 34.14286  2.79942
##        Quaest      7    0.7464 24.00000  3.50510
##        Sensus     71    0.9930 41.92958 17.27517
##           Vox     45    0.9785 24.13333 13.45214
```

```
wcalc.M <- data.frame(M = gsub("\\.", "-", wcalc.Mraw$period, perl = T),
                      latentM = wcalc.Mraw$latent1)
wcalc.M$M <- gsub("-1$", "-10", wcalc.M$M)


### Merge WCALC, and averaging estimates:
### Raw estimates no longer saved (look at raw file, instead)
dm <- merge(ms, wcalc.M, by = c("M"), all=T)

### Fill in missing presidents names (for those months for which we had not data)
### This is based on dates, so first impute day of month for missing observations
### For mnth, take center of month, doesn't matter because never two presidents

dm$Date  <- as.Date(paste(dm$M ,"-15",sep=""))

#Enter the dates of presidencies#######
pres.dates <- c(
  as.Date(c(
    "1979-03-15",#start of Figueiredo, prior to start of data
    "1985-03-15",#start of Sarney
    "1990-03-15",#start of collor
    "1992-10-02",#start of Franco
    "1995-01-01",#start of FHC
    "2003-01-01",#start of Lula
    "2011-01-01",#start of Dilma
    "2016-08-31",#start of Temer
    "2019-01-01")), #start of Bolsonaro
  Sys.Date())

dm$PresidentS <- dm$presUsed
missing.pres <- dm$PresidentS[is.na(dm$PresidentS)]
missing.dates <- dm$Date[is.na(dm$PresidentS)]
dm$PresidentS[is.na(dm$PresidentS)] <- ifelse(
  missing.dates < pres.dates[2],"FIGUEIREDO",
  ifelse(missing.dates >= pres.dates[2] & missing.dates < pres.dates[3],"SARNEY",
  ifelse(missing.dates >= pres.dates[3] & missing.dates < pres.dates[4],"COLLOR",
  ifelse(missing.dates >= pres.dates[4] & missing.dates < pres.dates[5],"FRANCO",
  ifelse(missing.dates >= pres.dates[5] & missing.dates < pres.dates[6],"CARDOSO",
  ifelse(missing.dates >= pres.dates[6] & missing.dates < pres.dates[7],"LULA",
  ifelse(missing.dates >= pres.dates[7] & missing.dates < pres.dates[8],"DILMA",
  ifelse(missing.dates >= pres.dates[8] & missing.dates < pres.dates[9],"TEMER",
  ifelse(missing.dates >= pres.dates[9],"BOLSONARO", NA)))))))))
```

```r
dm$PresidentS <- factor(dm$PresidentS,levels = c("FIGUEIREDO","SARNEY","COLLOR",
                                                 "FRANCO", "CARDOSO", "LULA",
                                                 "DILMA", "TEMER", "BOLSONARO"))


## This is the same for both datasets (record presidents that finished term, etc)
elected.pres <- levels(dm$PresidentS)[-c(1,2,4)]
concluded.pres <- levels(dm$PresidentS)[-c(3,7)]

### Add linear interpolations for average approach
### We do this by president so as not interpolate at end and start
### At end and start, repeat first or last obser
allpres <- levels(dm$PresidentS)
dm$popM.li <- NA
for(pp in allpres){
  l <- min(which(is.na(dm$popM) == F & dm$PresidentS == pp))
  h <- max(which(is.na(dm$popM) == F & dm$PresidentS == pp))
  dm$popM.li[l:h] <- data.frame(dm$popM[l:h],
          approx(dm$popM[l:h],method = "linear", n = length(dm$popM[l:h])))$y

  hh <-  max(which(dm$PresidentS == pp))
  ll <-  min(which(dm$PresidentS == pp))
  if(hh > h){#if there is missing at the end of term, impute average of last values
    # (and project in LatenM)
    dm$popM.li[(h+1):hh] <-mean(dm$popM.li[(h-2):h])
    m.to.fill <- length((h+1):hh) ##and for LatentM->linearly project from last 3 points
    dm$latentM[(h+1):hh] <-approx(dm$latentM[(h-2):h], n=(3+m.to.fill))$y[-c(1:3)]
  }
  if(ll<l){#if there is missing at  start of term, impute average of first values
    dm$popM.li[ll:(l-1)] <- mean(dm$popM.li[l:(l+1)])
    dm$latentM[ll:(l-1)] <- mean(dm$latentM[l:(l+1)])}
}

#Compute the counter for months in the term for each observation
dm$minterm <- round(as.numeric(dm$Date-pres.dates[as.numeric(dm$PresidentS)])/m1,1)

#Honey moon indicator, but only for elected presidents
dm$hm <- ifelse(dm$minterm<=4 &
                  is.element(dm$PresidentS, elected.pres),T,F)
dm$hmc <- ifelse(dm$minterm<=6 &
                   is.element(dm$PresidentS, elected.pres),abs(dm$minterm-6),0)


#Compute months left in term
dm$mleft <- round(
  ifelse(as.numeric(dm$PresidentS) == max(as.numeric(dm$PresidentS)),
         NA,#last president, can't compute months left in term
         as.numeric(pres.dates[1+ as.numeric(dm$PresidentS)]-dm$Date)/m1),1)

#Compute lame duck indicator
dm$ld <- ifelse(is.na(dm$mleft),F, #last president is NA
                dm$mleft<=4&is.element(dm$PresidentS,concluded.pres))
```

6. Summary statistics

```
Nall <- nrow(dd)

d <- subset(d, President != "Figueiredo")
dm <- subset(dm, PresidentS != "FIGUEIREDO")

Nused <- sum(is.na(d$Positive) != T) #total number of raw observations used
Npollsters <- length(unique(d$Institute))
Nmonths <- nrow(dm) #months spanned by the monthly dataset
Nmonthsdata <- nrow(ms) #months in which there was some observation
Nimp <- sum(is.na(dm$popM))
N1 <- min(dm$M)
NN <- max(dm$M)

obs <- data.frame(Nall, Nused, Npollsters, Nmonths, Nimp,
                  First = N1, Last = NN)

obs <- list(summary = obs, by.pollster = as.matrix(table(ds2$Varname)))
save(obs,file="DATA/obs_BR.RData")
write.csv(obs,'DATA/obs_BR-M.csv')
```

7. Save the datasets

```
dm <- dm %>%
  select(Date, M, PresidentS, minterm, hm, hmc, ld,
         popM, popM.li, latentM, instituteM)

dm$term <- as.character(dm$PresidentS)
dm$term[which(dm$Date>as.Date("1999-01-01")&dm$PresidentS=="CARDOSO")] <- "CARDOSO II"
dm$term[which(dm$Date>as.Date("2007-01-01")&dm$PresidentS=="LULA")] <- "LULA II"
dm$term[which(dm$Date>as.Date("2015-01-01")&dm$PresidentS=="DILMA")] <- "DILMA II"
dm$country<-"Brazil"
save(dm,file="DATA/data_BR-M.RData")
write.csv(dm,'DATA/data_BR-M.csv')

cat("\nCorrelation between MONTHLY linear imputed and Wcalc:\n")
```

```
##
## Correlation between MONTHLY linear imputed and Wcalc:
```

```
print(cor.test(dm$popM.li,dm$latentM))
```

```
##
##  Pearson's product-moment correlation
##
## data:  dm$popM.li and dm$latentM
## t = 113.1, df = 438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.9798984 0.9861412
## sample estimates:
##       cor
## 0.9833067
```

8. Plotting

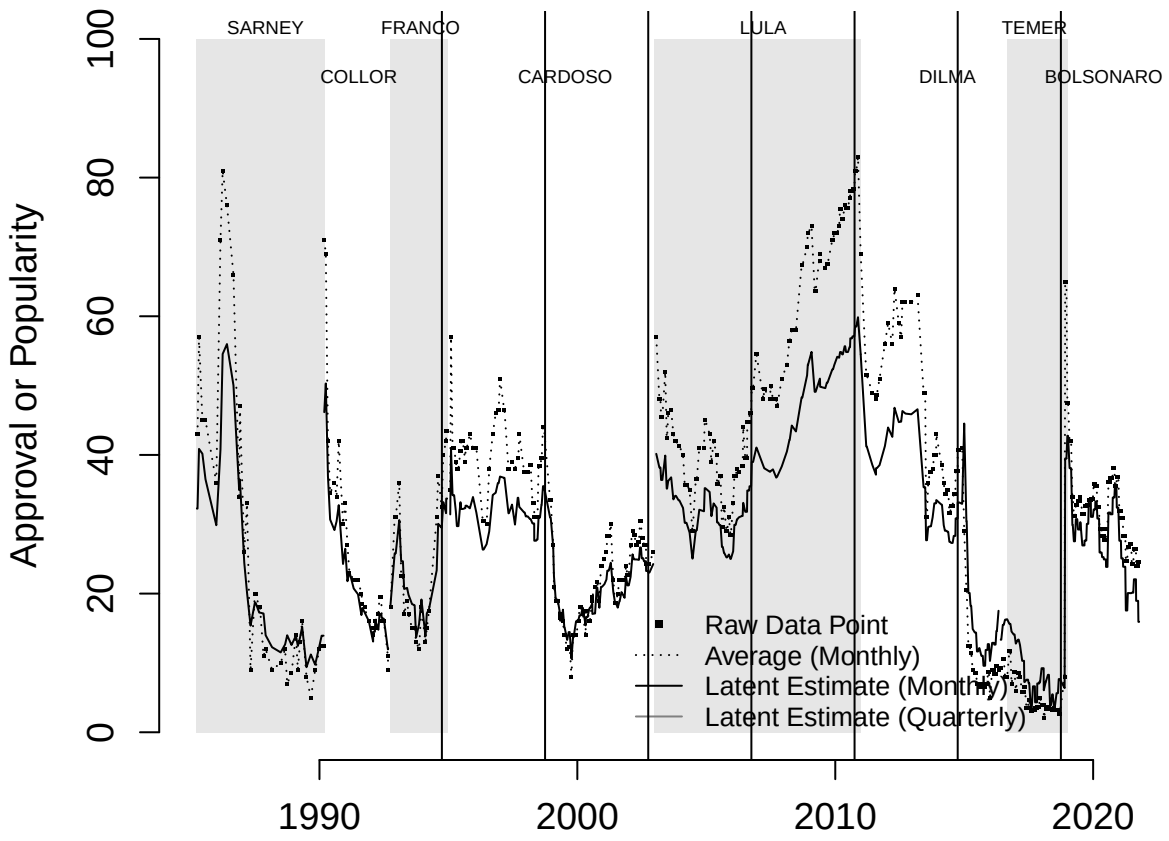
8.1. Merge estimates into d dataset for plotting

```
d <- merge(d, subset(dm, select = c(M, popM.li, latentM)), by = "M", all = T)
d <- d[order(d$M),]#make sure data are ordered
save(d, file = "DATA/data_BR-D.RData")

## Save popularity at election time
elec.date <- as.Date(c("1988-11-15","1989-11-15","1990-03-10",
                       "1992-10-03","1994-10-03","1996-10-03",
                       "1998-10-04","2000-10-01","2002-10-06",
                       "2004-10-03","2006-10-01",
                       "2008-10-05","2010-10-03",
                       "2012-10-07","2014-10-05","2016-10-01",
                       "2018-10-07"))
pop.elec <- data.frame(matrix(NA, nrow = 2, ncol = length(elec.date),
                              dimnames = list(c("popM.li","latentM"),
                                              c(as.character(elec.date)))))
#popularity of presidents close to election

for(i in 1:length(elec.date)){
  pop.elec[1,i] <- d$popM.li[which.min(abs(as.numeric(d$Date-elec.date[i])))]
  pop.elec[2,i] <- d$latentM[which.min(abs(as.numeric(d$Date-elec.date[i])))]
}
pop.elec <- t(pop.elec)
save(pop.elec,file="DATA/data_BR_elections.RData")
write.csv(pop.elec,'DATA/data_BR-elections.csv')
```

8.2. Plot

```
par(mar = c(2.5,5.5,.5,.5))
min.y <- 0
max.y <- 100
plot(d$Date, d$Positive, type = "n",
     ylab = "Approval or Popularity",
     xlab = "Year", bty = "n",
     cex.axis = 1.2, cex.lab = 1.2, ylim = c(min.y,max.y))

polygon(x = c(min(d$Date), pres.dates[1], pres.dates[1], min(d$Date)),
        y = c(min.y, min.y, max.y, max.y), border = NA, col = gray(0.9))

for(i in seq(2, length(pres.dates), by = 2)){
```

```
  polygon(x = c(pres.dates[i], pres.dates[i+1], pres.dates[i+1], pres.dates[i]),
          y = c(min.y, min.y, max.y, max.y), border = NA, col=gray(0.9))
  }

points(d$Date, d$popM, pch = ".", cex = 2)
alt <- -1
for(i in levels(d$PresidentS)){
  text(mean(d$Date[d$PresidentS == i], na.rm = T), max.y-2, labels = i, cex = 0.6,
       pos = 2 + alt)
  lines(d$Date[which(d$PresidentS==i)], d$latentM[which(d$PresidentS==i)], col=gray(0))
  lines(d$Date[which(d$PresidentS==i)], d$popM.li[which(d$PresidentS==i)], col=1, lty=3)
  alt <- alt * -1 #to alternate position of name
}
legend(x = as.Date("2002-01-01"), y = 20,
       legend = c("Raw Data Point","Average (Monthly)","Latent Estimate (Monthly)",
                  "Latent Estimate (Quarterly)"),
       cex = 0.8,
       lty = c(NA,3,1,1),
       col = c(1, gray(0), gray(0), gray(.5)),
       pch = c(".",NA,NA,NA), pt.cex = 4, bty = "n")

#abline(h = 33,lty = 2)
abline(v = c(as.Date("1994-10-01"),
           as.Date("1998-10-01"),
           as.Date("2002-10-01"),
           as.Date("2006-10-01"),
           as.Date("2010-10-01"),
           as.Date("2014-10-01"),
           as.Date("2018-10-01")))
```



8.3. Save Plot

```
pdf(file = "FIGURES/fig-popBR.pdf", width = 8, height = 6)
par(mar = c(2.5,5.5,.5,.5))
min.y <- 0
max.y <- 100
plot(d$Date, d$Positive, type = "n",
     ylab = "Approval or Popularity",
     xlab = "Year", bty = "n",
     cex.axis = 1.2, cex.lab = 1.2, ylim = c(min.y,max.y))

polygon(x = c(min(d$Date), pres.dates[1], pres.dates[1], min(d$Date)),
        y = c(min.y, min.y, max.y, max.y), border = NA, col = gray(0.9))

for(i in seq(2, length(pres.dates), by = 2)){
  polygon(x = c(pres.dates[i], pres.dates[i+1], pres.dates[i+1], pres.dates[i]),
          y = c(min.y, min.y, max.y, max.y), border = NA, col=gray(0.9))
  }

points(d$Date, d$popM, pch = ".", cex = 2)
alt <- -1
for(i in levels(d$PresidentS)){
  text(mean(d$Date[d$PresidentS == i], na.rm = T), max.y-2, labels = i, cex = 0.6,
       pos = 2 + alt)
  lines(d$Date[which(d$PresidentS==i)], d$latentM[which(d$PresidentS==i)], col=gray(0))
  lines(d$Date[which(d$PresidentS==i)], d$popM.li[which(d$PresidentS==i)], col=1, lty=3)
  alt <- alt * -1 #to alternate position of name
```

```
}
legend(x = as.Date("2002-01-01"), y = 20,
       legend = c("Raw Data Point","Average (Monthly)","Latent Estimate (Monthly)",
                  "Latent Estimate (Quarterly)"),
       cex = 0.8,
       lty = c(NA,3,1,1),
       col = c(1, gray(0), gray(0), gray(.5)),
       pch = c(".",NA,NA,NA), pt.cex = 4, bty = "n")

#abline(h = 33,lty = 2)
abline(v = c(as.Date("1994-10-01"),
           as.Date("1998-10-01"),
           as.Date("2002-10-01"),
           as.Date("2006-10-01"),
           as.Date("2010-10-01"),
           as.Date("2014-10-01"),
           as.Date("2018-10-01")))
```