MLB eligible population

Daniel J. Eck

In this document we estimate the MLB eligible population factoring in the changing levels of interest in the game. This document will frequently reference Baseball Reference, the US Census, Statistics Canada, Gallup survey data, and Wikipedia. This document will closely follow the calculations made in Eck [2020a] and Eck [2020b]. However, we will take a more exacting look at the MLB eligible population and historical baseball inference.

In this document the eligible MLB population will follow Baseball Reference's page listing a player's birthplace. Most countries that have produced two or more players will be added to the eligible MLB population. The countries that will form our eligible MLB population will be Aruba, Australia, Bahamas, Brazil, Canada, Colombia, Cuba, Curaçao, Dominican Republic, Jamaica, Japan, Mexico, Nicaragua, Panama, Puerto Rico, South Korea, Taiwan, the United States, the United States Virgin Islands, and Venezuela.

Latin American countries' populations will often be added four years before their first MLB player reached the MLB unless otherwise noted. This is to reflect the interest in baseball that must precede the arrival of an MLB player. The same will not apply to Asian players. This reflects the defacto ban that was in place prior to 1994 which limited Japanese players from entering the MLB.

We first obtain the the estimated MLB eligible population for the United States and Canada before 1960. Links to sources are commented below.

```
years <- 1871:2021

### Canadian population before 1960
#https://www65.statcan.gc.ca/acyb02/1907/acyb02_1907001701a-eng.htm
Can1871 <- 0.16 + 0.13
Can1881 <- 0.21 + 0.17
Can1891 <- 0.24 + 0.19
Can1901 <- 0.26 + 0.22

#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
CanM1871 <- 1.87
CanM1881 <- 2.19
CanM1891 <- 2.16
CanM1901 <- 2.75

# estimated proportion of males aged 20-30 with respect to males of
# all ages
propCan2030 <- (Can1871 + Can1881 + Can1891 + Can1901) /
  (CanM1871 + CanM1881 + CanM1891 + CanM1901)

#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
Can1911 <- 3.82 * propCan2030
Can1921 <- 4.53 * propCan2030
Can1931 <- 5.37 * propCan2030
Can1941 <- 5.90 * propCan2030

#https://www65.statcan.gc.ca/acyb02/1967/acyb02_19670194010-eng.htm
```

```
Can1951 <- 7.09 * propCan2030

### US population before 1960
#https://www2.census.gov/library/publications/decennial/1870/vital-statistics/1870b-33.pdf
US1870 <- 1.45 + 1.52
#https://www2.census.gov/library/publications/decennial/1880/vol-01-population/1880_v1-15.
US1880 <- 2.22 + 1.84
#https://www2.census.gov/library/publications/decennial/1900/volume-2/volume-2-p5.pdf
US1900 <- 2.73 + 2.37
US1890 <- mean(c(US1880,US1900))
#https://www2.census.gov/library/publications/decennial/1910/volume-1/volume-1-p6.pdf
US1910 <- 4.07 + 3.79
#https://www2.census.gov/library/publications/decennial/1920/volume-2/41084484v2ch03.pdf
US1920 <- 4.02 + 4.09
#https://www2.census.gov/library/publications/decennial/1930/population-volume-2/16440598v
US1930 <- 4.69 + 4.25
#https://www2.census.gov/library/publications/decennial/1940/population-volume-4/33973538v
US1940 <- 1.08 + 1.06 + 1.01 + 1.00 + 1.01 +
  1.01 + 0.99 + 0.98 + 0.97 + 0.95
#https://www2.census.gov/library/publications/decennial/1950/population-volume-2/21983999v
US1950 <- 5.00 + 5.30

Can <- c(Can1871, Can1881, Can1891, Can1901, Can1911, Can1921, Can1931, Can1941, Can1951)
US <- c(US1870, US1880, US1890, US1900, US1910, US1920, US1930, US1940, US1950)
datUSCan7050 <- data.frame(year = 187:195 * 10, US, Can)
```

We now provide estimates for the US and Canada populations for 1960 until the present, and we combine those estimates with the previous 1870-1950 estimates.

```
########################################
### Load in population data
########################################
library(tidyverse)
population_data <- read.csv("population.csv", header = TRUE)[, -1] %>%
  mutate(age20 = age20 / 1e3, age25 = age25 / 1e3) %>%
    mutate(pop = age20 + age25) %>%
    select("region", "year", "pop") %>%
  filter(year >= 1950)
population_data$region <- as.factor(population_data$region)
population_data$region <- recode_factor(
    population_data$region, WORLD = "world", Canada = "Can",
    "United States of America" = "US")

foo <- population_data %>% filter(region %in% c("US", "Can")) %>%
    filter(year >= 1960)
datUSCan6010 <- spread(foo, region, pop)[, c(1,3,2)]
bar  <- rbind(datUSCan7050, datUSCan6010) %>%
    mutate(pop = US + Can) %>%
    select(year, pop)

#https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluatio
#https://www.worldometers.info/world-population/us-population/
# US population in 2021 divided by US population in 2010
# population estimate taken October 11, 2021
```

```
bar[16, ] <- c(2021, 333474263 / 308745538  * bar[15, 2])
baz <- approx(x = bar$year, y = bar$pop, xout = years, n = length(years))
datUSCan <- data.frame(year = baz$x, pop = baz$y)
```

We now estimate baseball interest starting with Gallup polling data on the US population's favorite sport to watch, and we will assume the same interest levels for Canada. We shift historical interest by 13 years, so that the interest level for eligible MLB players aged 20-29 reflects when these players grew up (age 7-16).

```
#https://news.gallup.com/poll/4735/sports.aspx
interest <- data.frame(
  year = c(1937, 1948, 1960, 1972, 1981, 1990, 1994, 1998, 2000, 2001, 2002,
                 2003, 2004, 2005, 2006, 2007, 2008, 2013, 2017) + 13,
  interest = c(0.34, 0.39, 0.34, 0.22, 0.16, 0.16, 0.18, 0.12, 0.13, 0.12, 0.12,
                    0.10, 0.10, 0.12, 0.11, 0.13, 0.10, 0.14, 0.09)
)
foo <- approx(x = interest$year, y = interest$interest, xout = 1950:2024)
dat_interest <- data.frame(year = foo$x, interest = foo$y)
```

It is important to note that the Gallup survey does not parse out football and basketball into professional and college leagues while baseball is mostly only the MLB. Active MLB fans account for roughly 45% of the population as of 2017.

We also note that baseball participation outpaces interest presented by Gallup. This survey shows that the number of participants in baseball in the United States from 2006 to 2017 ranges from 12.5 to 16 million people. This survey collected information on individuals age 6 and older. In 2010 there were roughly 14.6 Americans aged 6 or older participating in baseball and there are roughly 84.3 million American males aged 5-49. Therefore participation is roughly 15.5% where this approximation assumes that participation stops after age 49 and that female participation in baseball is sparse. The interpolation of the (shifted) Gallup data shows that baseball interest is lower

```
dat_interest %>% filter(year == 2010 + 13)
```

```
##   year interest
## 1 2023    0.116
```

We therefore will not allow the minimum interest in baseball to be lower than 15.5%.

```
dat_interest$interest <- unlist(lapply(dat_interest$interest,
  function(x) max(c(x,0.155))))
```

We now estimate what baseball interest would be from 1920 to 1937. Before 1920, there was no radio broadcasting games and Babe Ruth had not yet started hitting home runs. We do think that extrapolation before 1920 will hold up. We first obtain MLB attendance data and the eligible MLB population weighted by interest at any set time.

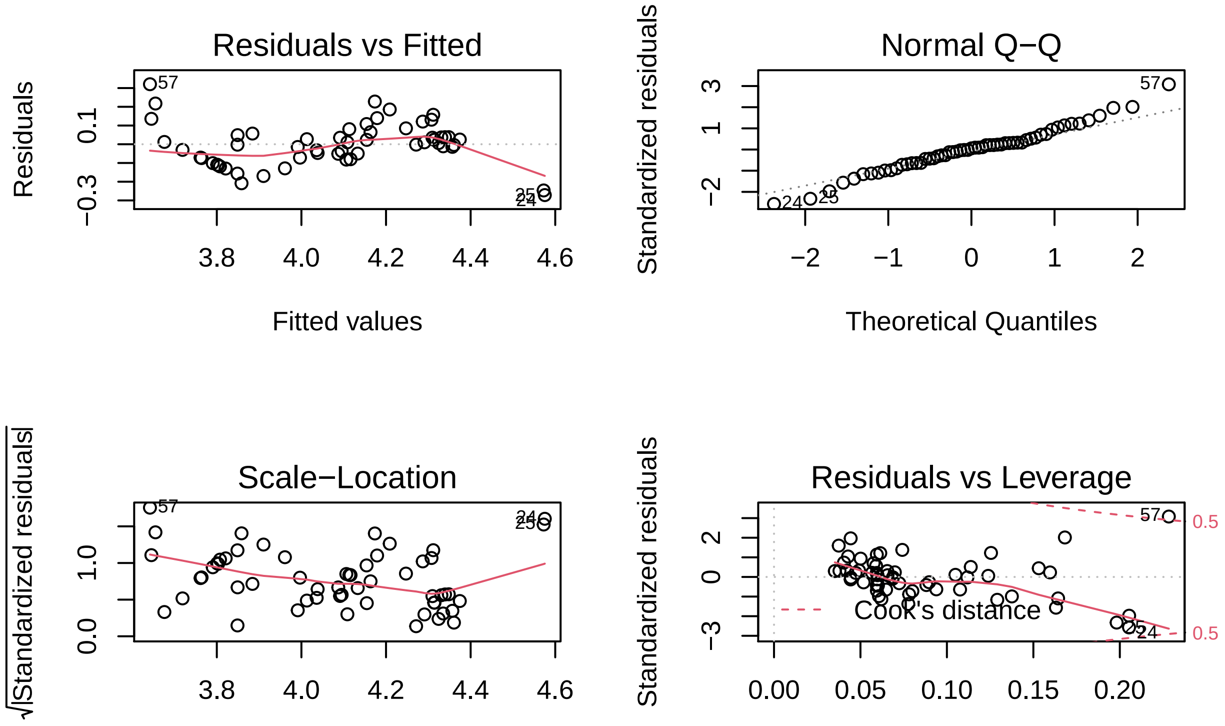
```
## obtain attendance data
MLBattendance <- read.csv("MLBattendance.csv") %>%
    rename(year = "Year")
MLBattendance[, 3] <- as.numeric(gsub(",", "", as.character(MLBattendance[, 3]))) / 1000
MLBattendance <- MLBattendance %>% rename(attend = "Attend.G")

## obtain Gallup interest data
#MLBinterest <- data.frame(
#  year = c(1937, 1948, 1960, 1972, 1981, 1990, 1994, 1998, 2000, 2001, 2002,
#                2003, 2004, 2005, 2006, 2007, 2008, 2013, 2017),
#  interest = c(0.34, 0.39, 0.34, 0.22, 0.16, 0.16, 0.18, 0.12, 0.13, 0.12, 0.12,
#                   0.10, 0.10, 0.12, 0.11, 0.13, 0.10, 0.14, 0.09) * multiple
#)

MLBinterest <- data.frame(
  year = c(1937, 1948, 1960, 1972, 1981, 1990, 1994, 1998, 2000, 2001, 2002,
                 2003, 2004, 2005, 2006, 2007, 2008, 2013, 2017),
  interest = c(0.34, 0.39, 0.34, 0.22, 0.16, 0.16, 0.18, 0.12, 0.13, 0.12, 0.12,
                    0.10, 0.10, 0.12, 0.11, 0.13, 0.10, 0.14, 0.09)
)
MLBinterest$interest <- unlist(lapply(MLBinterest$interest,
  function(x) max(c(x,0.155))))


## interpolate interest and combine with attendance
foo2 <- approx(x = MLBinterest$year, y = MLBinterest$interest, xout = 1937:2017)
dat_MLBinterest <- data.frame(year = foo2$x, interest = foo2$y)
bar <- merge(dat_MLBinterest, datUSCan) %>% mutate(pop = pop * interest)
attend <- (MLBattendance %>% filter(year >= 1937, year <= 2017) %>%
                        select(attend))
attend <- as.numeric(unlist(attend))
bar$attend <- attend
dat <- bar %>% filter(year < 1994) %>% mutate(year = year / 1000)
```

We fit a regression model with interest weighted MLB population as the response and quadratic terms for attendance and year.

```
## fit is good
m1 <- lm(pop ~ attend + I(attend^2) +
                 year + I(year^2), data = dat)
summary(m1)
```

```
##
## Call:
## lm(formula = pop ~ attend + I(attend^2) + year + I(year^2), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27150 -0.07156  0.00622  0.04861  0.32048
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.841e+03  5.093e+02  -3.614 0.000680 ***
## attend        2.633e-02  4.763e-02   0.553 0.582669
## I(attend^2)   6.132e-04  9.076e-04   0.676 0.502279
## year          1.865e+03  5.219e+02   3.573 0.000771 ***
## I(year^2)    -4.715e+02  1.335e+02  -3.530 0.000879 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1184 on 52 degrees of freedom
## Multiple R-squared:  0.8184, Adjusted R-squared:  0.8044
## F-statistic: 58.57 on 4 and 52 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(m1)
```



We now predict interest weighted MLB population for years 1920 to 1937 and compute interest.

```
baz <- data.frame(year = 1920:1936 / 1000,
                                                  attend = MLBattendance %>%
                                                      filter(year >= 1920, year <= 1936) %>%
                                                      select(attend))
dat_pop2036 <- data.frame(year = 1920:1936,
  pop = predict(m1, newdata = baz),
  tot = as.numeric(unlist(datUSCan %>% filter(year %in% 1920:1936) %>%
    select(pop))))
interest2036 <- dat_pop2036[, 2] / dat_pop2036[, 3]
interest2036
```

```
##  [1] 0.2170553 0.2186101 0.2210116 0.2217107 0.2273008 0.2345829 0.2424344
##  [8] 0.2434432 0.2443628 0.2489684 0.2508728 0.2515523 0.2529215 0.2514416
## [15] 0.2528053 0.2533605 0.2554885
```

We now shift interest as before and add it to our data frame on interest.

```
dat_interest <- rbind(data.frame(year = 1933:1949, interest = interest2036),
                                                      dat_interest)
```

Prior to Babe Ruth and the radio attendance was lower (World War 1 and the Spanish flu were also occurring during this time). We see that the average attendance from 1917 to 1919 is about 60% of the average attendance from 1920 to 1922

```
perc <- mean((MLBattendance %>% filter(year %in% 1917:1919))$attend) /
    mean((MLBattendance %>% filter(year %in% 1920:1922))$attend)
perc
```

```
## [1] 0.6069719
```

So we will assume that 1917 (shifted from 1930) has 60% of the interest as the average interest from 1920 to 1922, and we will linearly interpolate the in between values.

```
interest1720 <- approx(x = c(1930,1933),
             y = c(mean((dat_interest %>% filter(year %in% 1933:1935))$interest) * perc,
                        0.2615039), xout = 1930:1932)
dat_interest <- rbind(data.frame(year = 1930:1932, interest = interest1720$y),
                                            dat_interest)
```

There is not much concrete information for estimating baseball interest prior to what have estimated so far. Schmidt and Berri [2005] stated that at the onset of the 20th century, participation in MLB was confined to White Americans, originating primarily in the Northeastern states. We will suppose that baseball interest was 12% in 1871 and we will interpolate every value between 1871 and 1916 (shifted from 1929).

```
interest7116 <- approx(x = c(1871,1929),
             y = c(0.12, 0.1634787), xout = 1871:1929)
dat_interest <- rbind(data.frame(year = 1871:1929, interest = interest7116$y),
                                            dat_interest)
```

We now obtain the eligible MLB population for US and Canada after factoring in the changing interest in baseball.

```
datUSCan$pop <- datUSCan$pop * dat_interest$interest[1:nrow(datUSCan)]
head(datUSCan)
```

```
##   year       pop
## 1 1871 0.4053600
## 2 1872 0.4221407
## 3 1873 0.4390983
## 4 1874 0.4562329
## 5 1875 0.4735443
## 6 1876 0.4910327
```

```
tail(datUSCan)
```

```
##     year      pop
## 146 2016 4.020099
## 147 2017 4.048145
## 148 2018 4.076191
## 149 2019 4.104237
## 150 2020 4.132284
## 151 2021 4.160330
```

Other countries

The early Latin American countries On the onset of integration players from Cuba, Dominican Republic, Puerto Rico, and Panama populated the Negro leagues and the MLB [Burgos, 2007, page 112]. We will first investigate these countries continued contributions to the MLB eligible talent pool. These countries will be added to the MLB eligible talent pool starting in 1947.

Cuba Cuba continued to provide a steady stream of MLB talent after integration. Cubans have historically been among the most fervent of baseball fans, and have long been demanding of high caliber play [Baird, 2005]. Moreover, the prospect of beating Americans at their own game has long held special appeal to Cuba's leaders [Echevarría, 1999, page 354]. Baseball is the most popular sport in Cuba and 62% play it, and baseball is the official sport of Cuba. We will use the figure of 62% as a measure of Cuba's historic interest in baseball.

```
foo <- population_data %>% filter(region %in% c("Cuba")) %>%
    mutate(pop = pop * 0.62)
#https://www.worldometers.info/world-population/cuba-population/
# Cuban population in 2021 divided by Cuba population in 2010
# population estimate taken October 11, 2021
```

```
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 11.317806/11.23)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Cuba, xout = 1950:2021, n = 70)
datCuba <- data.frame(year = baz$x, pop = baz$y)
head(datCuba)
```

```
##   year       pop
## 1 1950 0.2971617
## 2 1951 0.3044215
## 3 1952 0.3116814
## 4 1953 0.3189413
## 5 1954 0.3262012
## 6 1955 0.3334611
```

```
tail(datCuba)
```

```
##    year       pop
## 67 2016 0.4900839
## 68 2017 0.4904308
## 69 2018 0.4907776
## 70 2019 0.4911245
## 71 2020 0.4914714
## 72 2021 0.4918183
```

This approximation is comparable to previous estimates. By 1977 Ron Fimrite reported that: An estimated 493,000 Cubans out of a total population of about nine million play some form of organized baseball [Fimrite, 1977]. Our estimate of the eligible MLB population (aged 20-29 males) is

```
(datCuba %>% filter(year == 1977) %>% select(pop)) * 1e6
```

```
##        pop
## 1 469380.2
```

Of course these estimates are for slightly different quantities. We target interest among aged 20-29 males. While Fimrite reported totals for overall participation in organized baseball, either in school or in the complicated system established by the government's Cuban Baseball Federation [Fimrite, 1977]. Both are estimates, it is encouraging to see that our estimate is slightly lower and is close to Firmrite's estimate. Also, we desire to weight by interest. So it is likely that our estimate underestimates overall interest.

Dominican Republic The Dominican Republic continued to provide a steady stream of MLB talent after integration. A figure quantifying baseball interest in the Dominican Republic is not easily obtainable. Considering that 62% of Cubans play baseball and that the Dominican Republic produces more MLB talent, we will place the interest level for the Dominican Republic at 85%.

```
population_data$region <- recode_factor(
    population_data$region, "Dominican Republic" = "DR")
foo <- population_data %>% filter(region %in% c("DR")) %>%
    mutate(pop = pop * 0.85)
#https://www.worldometers.info/world-population/dominican-republic-population/
# DR population in 2021 divided by DR population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 10.986358 / 9.7)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$DR, xout = 1950:2021, n = 70)
datDR <- data.frame(year = baz$x, pop = baz$y)
head(datDR)
```

```
##   year       pop
## 1 1950 0.1619904
## 2 1951 0.1674376
## 3 1952 0.1728847
## 4 1953 0.1783319
## 5 1954 0.1837790
## 6 1955 0.1892261
```

```
tail(datDR)
```

```
##    year       pop
## 67 2016 0.7791490
## 68 2017 0.7879087
## 69 2018 0.7966684
## 70 2019 0.8054280
## 71 2020 0.8141877
## 72 2021 0.8229474
```

Panama Panama continued to provide a steady stream of MLB talent after integration with a pronounced blip between 1979 and 1986. We will reduce the eligible MLB population by 50% over that span. A figure quantifying baseball interest in Panama is not easily obtainable. Considering that 62% of Cubans play baseball and that Panama produces comparable MLB talent given the differences in population, we will place the interest level for Panama at 62%.

```
population_data$region <- recode_factor(
    population_data$region, "Panama" = "Pan")
foo <- population_data %>% filter(region %in% c("Pan")) %>%
    mutate(pop = pop * 0.62)
#https://www.worldometers.info/world-population/panama-population/
# Pan population in 2021 divided by Pan population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 4.401139 / 3.64)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Pan, xout = 1950:2021, n = 70)
datPan <- data.frame(year = baz$x, pop = baz$y)
datPan[datPan$year %in% 1979:1986, ]$pop <-
    0.5 * datPan[datPan$year %in% 1979:1986, ]$pop
head(datPan)
```

```
##   year        pop
## 1 1950 0.04283766
## 2 1951 0.04378006
## 3 1952 0.04472246
## 4 1953 0.04566486
## 5 1954 0.04660726
## 6 1955 0.04754966
```

```
tail(datPan)
```

```
##    year       pop
## 67 2016 0.2034357
## 68 2017 0.2069069
## 69 2018 0.2103782
## 70 2019 0.2138495
## 71 2020 0.2173208
## 72 2021 0.2207921
```

Puerto Rico Puerto Rico continued to provide a steady stream of MLB talent after integration. A figure quantifying baseball interest in Puerto Rico is not easily obtainable. Considering that 62% of Cubans play baseball and that Puerto Rico produces comparable MLB talent with less people, we will place the interest level for Puerto Rico at 85%.

```
population_data$region <- recode_factor(
    population_data$region, "Puerto Rico" = "PR")
foo <- population_data %>% filter(region %in% c("PR")) %>%
    mutate(pop = pop * 0.85)
#https://www.worldometers.info/world-population/panama-population/
# PR population in 2021 divided by PR population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 2.765747 / 3.7)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$PR, xout = 1950:2021, n = 70)
datPR <- data.frame(year = baz$x, pop = baz$y)
head(datPR)
```

```
##   year       pop
## 1 1950 0.1420350
## 2 1951 0.1397568
## 3 1952 0.1374787
## 4 1953 0.1352005
## 5 1954 0.1329223
## 6 1955 0.1306442
```

```
tail(datPR)
```

```
##    year       pop
## 67 2016 0.2025148
## 68 2017 0.1971237
## 69 2018 0.1917325
## 70 2019 0.1863413
## 71 2020 0.1809502
## 72 2021 0.1755590
```

We will extrapolate the tallies of Cuba, Dominican Republic, Puerto Rico, and Panama to 1948, the first complete season of post-integration baseball. We will also combine these tallies to those computed for the US and Canada.

```
MLBpop <- datUSCan
MLBpop[datUSCan$year %in% 1948:2021, ]$pop <-
    datUSCan[datUSCan$year %in% 1948:2021, ]$pop +
    c(0.62, 0.63, datCuba$pop + datDR$pop + datPan$pop + datPR$pop)
```

Additional Latin American countries

Aruba Aruba has recently produced a steady stream MLB players beginning in 1996. Baseball is one of the most popular sports on the island, with a number of Aruban citizens having gone to the United States to play professionally. Considering that 62% of Cubans play baseball and that Aruba produces comparable MLB talent given the differences in population, we will place the interest level for Panama at 62%. We will add Aruba to the eligible MLB population starting in 1992, four years before their first player made the MLB.

```
foo <- population_data %>% filter(region %in% c("Aruba")) %>%
    mutate(pop = pop * 0.62)
#https://www.worldometers.info/world-population/aruba-population/
# Aruba population in 2021 divided by Aruba population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 1073.43 / 1017)
```

```
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Aruba, xout = 1992:2021)
datAruba <- data.frame(year = baz$x, pop = baz$y)
head(datAruba)
```

```
##   year         pop
## 1 1992 0.003143028
## 2 1993 0.003172292
## 3 1994 0.003201556
## 4 1995 0.003230820
## 5 1996 0.003260084
## 6 1997 0.003289348
```

```
tail(datAruba)
```

```
##    year         pop
## 25 2016 0.004209459
## 26 2017 0.004230069
## 27 2018 0.004250678
## 28 2019 0.004271288
## 29 2020 0.004291898
## 30 2021 0.004312508
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1992, ]$pop <-
    MLBpop[MLBpop$year >= 1992, ]$pop + datAruba$pop
```

Bahamas The Bahamas has fielded an MLB player consistently since integration consistently since 1932. A figure quantifying baseball interest in the Bahamas is not easily obtainable. Considering that 62% of Cubans play baseball and that the Bahamas produces fewer MLB talent given the differences in population, we will place the interest level for the Bahamas at 20%.

```
population_data$region <- recode_factor(
    population_data$region, "Bahamas" = "BA")
foo <- population_data %>% filter(region %in% c("BA")) %>%
    mutate(pop = pop * 0.20)
#https://www.worldometers.info/world-population/bahamas-population/
# Bahamas population in 2021 divided by Bahamas population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 398027 / 354942)
foo <- foo %>% arrange(year)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$BA, rule = 2,
                            xout = 1930:2021, n = length(1930:2021))
datBA <- data.frame(year = baz$x, pop = baz$y)
head(datBA)
```

```
##   year       pop
## 1 1930 0.0011458
## 2 1931 0.0011458
## 3 1932 0.0011458
## 4 1933 0.0011458
## 5 1934 0.0011458
## 6 1935 0.0011458
```

```
tail(datBA)
```

```
##    year         pop
## 87 2016 0.006266119
## 88 2017 0.006330973
## 89 2018 0.006395826
## 90 2019 0.006460679
## 91 2020 0.006525532
## 92 2021 0.006590386
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1930, ]$pop <-
    MLBpop[MLBpop$year >= 1930, ]$pop + datBA$pop
```

Brazil Brazil has recently produced a handful of MLB players beginning in 2012. That being said, baseball is not very popular in Brazil. We will place the interest level for Brazil at 2%.

```
population_data$region <- recode_factor(
    population_data$region, "Brazil" = "BR")
foo <- population_data %>% filter(region %in% c("BR")) %>%
    mutate(pop = pop * 0.02)
#https://www.worldometers.info/world-population/brazil-population/
# Brazil population in 2021 divided by Brazil population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 214.483151 / 195.7)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$BR, xout = 2012:2021)
datBR <- data.frame(year = baz$x, pop = baz$y)
head(datBR)
```

```
##   year       pop
## 1 2012 0.3546074
## 2 2013 0.3576484
## 3 2014 0.3606894
## 4 2015 0.3637304
## 5 2016 0.3667714
## 6 2017 0.3698125
```

```
tail(datBR)
```

```
##    year       pop
## 5  2016 0.3667714
## 6  2017 0.3698125
## 7  2018 0.3728535
## 8  2019 0.3758945
## 9  2020 0.3789355
## 10 2021 0.3819765
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 2012, ]$pop <-
    MLBpop[MLBpop$year >= 2012, ]$pop + datBR$pop
```

Colombia An influx of Colombian players began in the mid-1990s. We include Colombia into the eligible MLB talent pool starting in 1992, four years before the first player of this influx, Edgar Renteria, entered the MLB. Before 1992

Colombia produced only two MLB players. Baseball is not a major sport in Colombia. Although a figure quantifying baseball interest in Colombia is not easily obtainable, we will place interest at 5%.

```
population_data$region <- recode_factor(
    population_data$region, "Colombia" = "Col")
foo <- population_data %>% filter(region %in% c("Col")) %>%
    mutate(pop = pop * 0.05)
#https://www.worldometers.info/world-population/colombia-population/
# Colombia population in 2021 divided by Colombia population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 51.573067 / 45.22)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Col, xout = 1992:2021)
datCol <- data.frame(year = baz$x, pop = baz$y)
head(datCol)
```

```
##   year       pop
## 1 1992 0.1633703
## 2 1993 0.1645845
## 3 1994 0.1657986
## 4 1995 0.1670128
## 5 1996 0.1682270
## 6 1997 0.1694411
```

```
tail(datCol)
```

```
##    year       pop
## 25 2016 0.2146239
## 26 2017 0.2171700
## 27 2018 0.2197161
## 28 2019 0.2222622
## 29 2020 0.2248082
## 30 2021 0.2273543
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1992, ]$pop <-
    MLBpop[MLBpop$year >= 1992, ]$pop + datCol$pop
```

Curaçao Per Baseball Reference and MLB.com, fifteen Curaçao natives have played in the Major Leagues making for a rate of roughly one per every 10,000 residents that far surpasses the yield of any other country in the world. We are going to place baseball interest in Curaçao at 100% (probably should be higher!). We will add Curaçao's tally to the overall population beginning in 1985, four years before the first player from Curaçao joined the MLB. Roughly 6% of Curaçao's population are males aged 20-29 according to their 2011 Census.

```
foo <- approx(x = c(1981, 1992, 2001, 2011, 2021),
                            y = c(.148, .144, .131, .151, .165) * .06, xout = 1985:2021)
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1985, ]$pop <-
    MLBpop[MLBpop$year >= 1985, ]$pop + foo$y
```

Jamaica Jamaica has produced a handful of MLB players beginning recently with Chili Davis in 1981. Baseball is a new, non-traditional sport in Jamaica. We will add Jamaica to the eligible MLB population starting in 1977, four years before their first player made the MLB. We will place Jamaican interest in baseball at 5%.

```
foo <- population_data %>% filter(region %in% c("Jamaica")) %>%
    mutate(pop = pop * 0.05)
#https://www.worldometers.info/world-population/jamaica-population/
# Jamaica population in 2021 divided by Jamaica population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 2.977628 / 2.81)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Jam, xout = 1977:2021)
datJam <- data.frame(year = baz$x, pop = baz$y)
head(datJam)
```

```
##   year         pop
## 1 1977 0.007453505
## 2 1978 0.007754020
## 3 1979 0.008054535
## 4 1980 0.008355050
## 5 1981 0.008589590
## 6 1982 0.008824130
```

```
tail(datJam)
```

```
##    year        pop
## 40 2016 0.01080748
## 41 2017 0.01086424
## 42 2018 0.01092100
## 43 2019 0.01097777
## 44 2020 0.01103453
## 45 2021 0.01109129
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1977, ]$pop <-
    MLBpop[MLBpop$year >= 1977, ]$pop + datJam$pop
```

Mexico Mexico has produced a steady strem of MLB talent. However, the eligible MLB population hailing from Mexico is hard to quantify.

From this article in 2018:

> Attendance at professional baseball games in Mexico has been increasing steadily. In its heyday in 1979, the Mexican League drew 4.6 million fans. From 1998 – 2003, attendance fell flat, averaging approximately 2.3 million tickets sold each year as soccer ticket sales rose 27 per cent in the same period, to more than 4.9 million annually. As Mexicans became more affluent, they spent more on soccer than ever before. However, that trend is beginning to reverse. The Mexican League nearly matched their 1979 high, having sold slightly more than 4 million tickets in 2017, despite losing many dates to an extraordinarily active hurricane season. Total attendance has topped 3 million in all but one season since 2003. Although attendance numbers are healthy, at least half of Mexican League teams are bordering on bankruptcy and off-season franchise relocations are common.

Although baseball interest is high in Mexico, soccer is the dominant sport. There is also a concern as to why so (relatively) few baseball players come from Mexico. In 1992 Ranieri [1992] stated that:

> Mexico also produces players for US professional baseball. Their entrance into this country, however, is limited by two factors: individual Mexican ball players are not allowed to negotiate contracts with US teams, and the US Department of Labor places a quota on the number of temporary worker certifications for foreigners, mainly Latin Americans, allowed into the baseball industry. In 1991, the number was 585. As a result, very few Mexicans ever get an opportunity to exhibit their talents in this country. As the baseball

> season ended, there were only eight Mexican-born players in the major leagues and only 30 more in the minor leagues. In comparison, there were over 50 major league players from the Dominican Republic and almost 400 Dominican players in the minors in the United States. This despite the fact Mexico has a population 10 times that of the Dominican Republic.

We see that there are limitations in place which limit how many Mexican players enter the MLB. In light of these challenges, and with no hard figures of baseball interest, we will suppose that baseball interest in Mexico is at 15% in the height of its popularity in 1979. We will also suppose that baseball interest in Mexico was at 5% in 1930. We will use an interpolation based on the attendance data to estimate the interest for other time points. First the raw population totals of aged 20-29 Mexican males.

```
foo <- population_data %>% filter(region %in% c("Mexico"))
#https://www.worldometers.info/world-population/mexico-population/
# Mexico population in 2021 divided by Mexico population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 130.656659 / 114.1)
bar <- spread(foo, region, pop)
# an estimate of the number of Mexican males aged 20-29: 16.55 * 0.06
baz <- approx(x = c(1930, bar$year),
                            y = c(16.55 * 0.06, bar$Mexico), xout = 1930:2021)
datMexico <- data.frame(year = baz$x, pop = baz$y)
```

We will now interpolate interest.

```
foo <- approx(y = c(4.6, 2.3, 2.3, 4), x = c(1979, 1998, 2003, 2021),
                            xout = 1979:2021)
attend <-  data.frame(year = foo$x, attend = foo$y)
interest7919 <- data.frame(year = 1979:2021,
                                                     interest = 0.15 * attend$attend / max
bar <- approx(x = c(1930, 1979:2021),
                            y = c(0.05, interest7919$interest), xout = 1930:2021)
Mexico_interest <- data.frame(year = 1930:2021, interest = bar$y)
datMexico$pop <- datMexico$pop * Mexico_interest$interest
head(datMexico)
```

```
##   year        pop
## 1 1930 0.04965000
## 2 1931 0.05491578
## 3 1932 0.06043562
## 4 1933 0.06620952
## 5 1934 0.07223747
## 6 1935 0.07851949
```

```
tail(datMexico)
```

```
##    year      pop
## 87 2016 1.138731
## 88 2017 1.183509
## 89 2018 1.229033
## 90 2019 1.275301
## 91 2020 1.322315
## 92 2021 1.370075
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1930, ]$pop <-
    MLBpop[MLBpop$year >= 1930, ]$pop + datMexico$pop
```

Nicaragua Nicaragua has produced a steady stream of baseball players since 1976. We include Nicaragua into the eligible MLB talent pool starting in 1972, four years before the their first player, Dennis Martinez, entered the MLB. Considering that 62% of Cubans play baseball and that Nicaragua produces fewer MLB talent given the differences in population, we will place the interest level for Nicaragua at 20%.

```
population_data$region <- recode_factor(
    population_data$region, "Nicaragua" = "NC")
foo <- population_data %>% filter(region %in% c("NC")) %>%
    mutate(pop = pop * 0.20)
#https://www.worldometers.info/world-population/nicaragua-population/
# NC population in 2021 divided by NC population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 6.724836 / 5.824)
foo <- foo %>% arrange(year)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$NC, xout = 1972:2021)
datNC <- data.frame(year = baz$x, pop = baz$y)
head(datNC)
```

```
##   year       pop
## 1 1972 0.0382550
## 2 1973 0.0403497
## 3 1974 0.0424444
## 4 1975 0.0445391
## 5 1976 0.0466338
## 6 1977 0.0487285
```

```
tail(datNC)
```

```
##    year       pop
## 45 2016 0.1174393
## 46 2017 0.1189622
## 47 2018 0.1204851
## 48 2019 0.1220080
## 49 2020 0.1235309
## 50 2021 0.1250538
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1972, ]$pop <-
    MLBpop[MLBpop$year >= 1972, ]$pop + datNC$pop
```

US Virgin Islands The US Virgin Islands has produced a steady stream of baseball players since 1945. Considering that 62% of Cubans play baseball and that the US Virgin Islands produces comparable MLB talent given the differences in population, we will place the interest level for Panama at 85%.

```
population_data$region <- recode_factor(
    population_data$region, "United States Virgin Islands" = "USVI")
foo <- population_data %>% filter(region %in% c("USVI")) %>%
    mutate(pop = pop * 0.85)
#https://www.worldometers.info/world-population/united-states-virgin-islands-population/
# USVI population in 2021 divided by USVI population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 104229 / 108358)
# USVI population in 1940 set to USVI population in 1950
foo[9, 1] <- foo[7, 1]; foo[9, 2:3] <- c(1940, foo[1,3])
foo <- foo %>% arrange(year)
```

```
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$USVI, xout = 1945:2021)
datUSVI <- data.frame(year = baz$x, pop = baz$y)
head(datUSVI)
```

```
##   year        pop
## 1 1945 0.00147815
## 2 1946 0.00147815
## 3 1947 0.00147815
## 4 1948 0.00147815
## 5 1949 0.00147815
## 6 1950 0.00147815
```

```
tail(datUSVI)
```

```
##    year         pop
## 72 2016 0.005698984
## 73 2017 0.005678824
## 74 2018 0.005658663
## 75 2019 0.005638502
## 76 2020 0.005618341
## 77 2021 0.005598180
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1945, ]$pop <-
    MLBpop[MLBpop$year >= 1945, ]$pop + datUSVI$pop
```

Venezuela An influx of Venezuelan players began in 1967 with a boom in the 1980s and early 1990s. Before that there was scattering of baseball players tracing back to 1950 and the 1940s. It is therefore likely that baseball interest has been changing in Venezuela over time. Recent surveys show that 75 percent of criollos — native-born Venezuelans — selected baseball as their favorite sport, and no one in Venezuela ever gets bored watching or talking about it [Jamail, 2008]. We will therefore assign a 75% interest level to Venezuela from 1967 to present, and we will assign a 40% interest level prior to 1967.

```
population_data$region <- recode_factor(
    population_data$region, "Venezuela (Bolivarian Republic of)" = "Ven")
foo <- population_data %>% filter(region %in% c("Ven"))
#https://www.worldometers.info/world-population/venezuela-population/
# Ven population in 2021 divided by Ven population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 28.333252 / 28.44)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Ven, xout = 1950:2021)
datVen <- data.frame(year = baz$x, pop = baz$y)
datVen[datVen$year >= 1967, 2] <- 0.75 * datVen[datVen$year >= 1967, 2]
datVen[datVen$year < 1967, 2] <- 0.40 * datVen[datVen$year < 1967, 2]
head(datVen)
```

```
##   year       pop
## 1 1950 0.1691824
## 2 1951 0.1758800
## 3 1952 0.1825775
## 4 1953 0.1892751
## 5 1954 0.1959726
## 6 1955 0.2026702
```

```
tail(datVen)
```

```
##    year      pop
## 67 2016 1.942531
## 68 2017 1.941867
## 69 2018 1.941203
## 70 2019 1.940538
## 71 2020 1.939874
## 72 2021 1.939210
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1950, ]$pop <-
    MLBpop[MLBpop$year >= 1950, ]$pop + datVen$pop
```

Asia

Japan An influx of Japanese players began arriving to the MLB in 1995. Baseball is a major sport in Japan. Data published in June, 2018 by Japan's Central Research Services showed baseball was comfortably the most popular sport in the country, with 48.1 percent of respondents naming it as their favorite sport.

```
foo <- population_data %>% filter(region %in% c("Japan")) %>%
    mutate(pop = pop * 0.481)
#https://www.worldometers.info/world-population/japan-population/
# Japan population in 2021 divided by Japan population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 125.982980 / 128.1)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Japan, xout = 1995:2021)
datJapan <- data.frame(year = baz$x, pop = baz$y)
head(datJapan)
```

```
##   year      pop
## 1 1995 4.260279
## 2 1996 4.298528
## 3 1997 4.336776
## 4 1998 4.375025
## 5 1999 4.413274
## 6 2000 4.451522
```

```
tail(datJapan)
```

```
##    year      pop
## 22 2016 3.435980
## 23 2017 3.430771
## 24 2018 3.425562
## 25 2019 3.420353
## 26 2020 3.415144
## 27 2021 3.409935
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1995, ]$pop <-
    MLBpop[MLBpop$year >= 1995, ]$pop + datJapan$pop
```

South Korea An influx of South Korean players began arriving to the MLB in 1994. Baseball is the favorite sport of South Koreans with 62% of respondents listing it as a sport they follow in a 2017 survey. This survey records overall fan interest and not favorite sport. We will therefore state that the overall interest is 48.1% percent, same as Japan, to stay consistent with the Gallup survey that was used for US baseball interest.

```
population_data$region <- recode_factor(
    population_data$region, "Republic of Korea" = "SK")
foo <- population_data %>% filter(region %in% c("SK")) %>%
    mutate(pop = pop * 0.481)
#https://www.worldometers.info/world-population/south-korea-population/
# SK population in 2021 divided by SK population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 51.325401 / 49.55)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$SK, xout = 1994:2021)
datSK <- data.frame(year = baz$x, pop = baz$y)
head(datSK)
```

```
##   year      pop
## 1 1994 2.060841
## 2 1995 2.044908
## 3 1996 2.028976
## 4 1997 2.013043
## 5 1998 1.997110
## 6 1999 1.981178
```

```
tail(datSK)
```

```
##    year      pop
## 23 2016 1.728357
## 24 2017 1.733879
## 25 2018 1.739401
## 26 2019 1.744923
## 27 2020 1.750445
## 28 2021 1.755967
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1994, ]$pop <-
    MLBpop[MLBpop$year >= 1994, ]$pop + datSK$pop
```

Taiwan A steady stream of Taiwanese players began arriving to the MLB in 2002. Baseball is one of the country's most popular spectator sports, and it is commonly considered the national sport in Taiwan. The original reference to tabulate the Taiwan population of males aged 20-29 used in Eck [2020b] has been taken down. From that reference Eck [2020b] approximated the population of males aged 20-29 size to be a constant 1.5 million people. Rich survey data is not easily obtainable for Taiwan. We will estimate the interest at 40%, placing it a little lower than that of Japan and South Korea.

```
datTai <- data.frame(year = 2002:2021, pop = 1.5 * 0.40)
head(datTai)
```

```
##   year pop
## 1 2002 0.6
## 2 2003 0.6
## 3 2004 0.6
## 4 2005 0.6
## 5 2006 0.6
## 6 2007 0.6
```

```
tail(datTai)
```

```
##    year pop
## 15 2016 0.6
## 16 2017 0.6
## 17 2018 0.6
## 18 2019 0.6
## 19 2020 0.6
## 20 2021 0.6
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 2002, ]$pop <-
    MLBpop[MLBpop$year >= 2002, ]$pop + datTai$pop
```

Oceania

Australia An influx of Australian players began in the early-1990s. We include Australia into the eligible MLB talent pool starting in 1982, four years before the first player of this influx, Craig Shipley, entered the MLB. Before 1982 Australia produced only one MLB player, Joe Quinn who played in the late nineteenth century. Figures for baseball popularity in Australia are hard to find, but we know that it is not the most popular sport. In 2003, there were roughly 57,000 Australians playing baseball in around 5000 teams. We will estimate the overall interest at 5%.

```
population_data$region <- recode_factor(
    population_data$region, "Australia" = "Aus")
foo <- population_data %>% filter(region %in% c("Aus")) %>%
    mutate(pop = pop * 0.05)
#https://www.worldometers.info/world-population/australia-population/
# Aus population in 2021 divided by Aus population in 2010
# population estimate taken October 11, 2021
foo[8, 1] <- foo[7, 1]; foo[8, 2:3] <- c(2021, foo[7,3] * 25.876349 / 22.03)
bar <- spread(foo, region, pop)
baz <- approx(x = bar$year, y = bar$Aus, xout = 1994:2021)
datAus <- data.frame(year = baz$x, pop = baz$y)
head(datAus)
```

```
##   year        pop
## 1 1994 0.06990605
## 2 1995 0.06976523
## 3 1996 0.06962440
## 4 1997 0.06948358
## 5 1998 0.06934275
## 6 1999 0.06920193
```

```
tail(datAus)
```

```
##    year        pop
## 23 2016 0.09137933
## 24 2017 0.09270362
## 25 2018 0.09402791
## 26 2019 0.09535219
## 27 2020 0.09667648
## 28 2021 0.09800077
```

We add this population to the overall tally of the eligible MLB population.

```
MLBpop[MLBpop$year >= 1994, ]$pop <-
    MLBpop[MLBpop$year >= 1994, ]$pop + datAus$pop
```

World Wars

Baseball Reference lists the period of 1918-1919 and 1942-1946 as world war years. We will adopt their convention and apply an interest reduction multiplier to the eligible MLB population over those seasons. We will also allow these reductions to gradually tail off in the years after the world wars to reflect the decimation of the talent pool caused by war.

According to historian Jim Leeke, author of "From the Dugouts to the Trenches: Baseball During the Great War," approximately 38 percent of active Major League players went on to serve in World War I, and eight current or former players were either killed in action or died of illness during the war.

World War II was more devastating than World War I. A paragraph from this source lists how much World War II disrupted the MLB:

> More than 500 major league players swapped flannels for khakis during World War II, and such well-known players as Stan Musial, Joe DiMaggio and Ted Williams served their nation off the diamond. The minor leagues, formerly a veritable oasis of baseball talent, were seriously affected by the manpower shortage with 4,076 players seeing military service. On a daily basis, talent was drained from the game as promising young athletes who had spent summers developing their athletic skills were plucked from baseball diamonds all across the country and taught to fly planes, shoot weapons and maneuver tanks. No more than 12 minor leagues survived during the war years compared to 44 circuits that operated in 1940 [27%].

With these accounts we will reduce the eligible MLB population by 38% during World War I and by 70% for the peak of World War II. We will allow for World War II to be more gradual.

```
# WWI
MLBpop[MLBpop[, 1] %in% c(1918:1921), ]$pop <-
    MLBpop[MLBpop[, 1] %in% c(1918:1921), ]$pop * c(0.75,0.62,0.75,0.90)

# WWII
MLBpop[MLBpop[, 1] %in% c(1941:1952), ]$pop <-
    MLBpop[MLBpop[, 1] %in% c(1941:1952), ]$pop *
    c(0.80,0.50,0.30,0.30,0.35,0.40,0.45,0.50,0.60,0.70,0.80,0.90)
```

AL and NL integrate at different rates

The American League (AL) and the National League (NL) integrated at different rates. To incorporate this information we need to obtain the tallies of nonwhite Americans from 1947-1975 and isolate the international population over this span. The span 1948-1975 coincides with production differences between blacks in the two leagues before the reserve clause was abolished.

We first isolate the US and Canada populations and adjust them to account for the world wars.

```
# US and Canada populations during WWI
 datUSCan[ datUSCan[, 1] %in% c(1918:1921), ]$pop <-
     datUSCan[ datUSCan[, 1] %in% c(1918:1921), ]$pop * c(0.75,0.62,0.75,0.90)

# US and Canada populations during WWII
 datUSCan[ datUSCan[, 1] %in% c(1941:1952), ]$pop <-
    datUSCan[datUSCan[, 1] %in% c(1941:1952), ]$pop *
    c(0.80,0.50,0.30,0.30,0.35,0.40,0.45,0.50,0.60,0.70,0.80,0.90)
```

We now estimate the nonwhite population of US males aged 20-29 from 1948-1975. We first obtain the tallies of nonwhite males for the 1950 Census, pages 1-91, 1-92, and 1-94. First the 1940 nonwhite US males ages 20-29 adjusted for interest

```
nonwhiteUS1940 <- round((578750 + 558649) *
    dat_interest[dat_interest$year == 1940, 2])
nonwhiteUS1940
```

```
## [1] 276892
```

Here is the tally of 1950 nonwhite US males aged 20-29 adjusted for interest and war effort

```
nonwhiteUS1950 <- round((603511 + 622371) *
    dat_interest[dat_interest$year == 1950, 2] * 0.70)
nonwhiteUS1950
```

```
## [1] 291760
```

The following tally is of 1960 nonwhite US males aged 20-29 adjusted for interest. This tally is taken from the 1960 US Census. This tally is the multiple of interest adjusted total aged 20-29 males and the fraction of nonwhite males in the population.

```
nonwhiteUS1960 <- round((5272440 + 5333075) *
    dat_interest[dat_interest$year == 1960, 2] *
        9964345 / 88331494)
nonwhiteUS1960
```

```
## [1] 461146
```

The following tally is of 1970 nonwhite US males aged 20-29 adjusted for interest. This tally is taken from the 1970 US Census. This tally is the average of the 15-24 and 25-34 tallies of black Americans.

```
nonwhiteUS1970 <- round(mean(c(1234855, 2047791)) *
    dat_interest[dat_interest$year == 1970, 2])
nonwhiteUS1970
```

```
## [1] 578566
```

We now approximate the tally of 1975 nonwhite US males aged 20-29 adjusted for interest.

```
nonwhiteUS1975 <- round(mean(c(nonwhiteUS1970 / nonwhiteUS1960, 1)) *
  nonwhiteUS1970 /
    dat_interest[dat_interest$year == 1970, 2] *
    dat_interest[dat_interest$year == 1975, 2])
```

We now interpolate the US nonwhite populations from 1948 to 1975.

```
foo <- approx(x = c(1940,1950,1960,1970,1975),
                           y = c(nonwhiteUS1940, nonwhiteUS1950,
                                        nonwhiteUS1960, nonwhiteUS1970,
                                        nonwhiteUS1975),
                           xout = 1948:1975)
dat_nonwhite <- data.frame(year = 1948:1975,
  pop = round(foo$y / 1e6, 6))
```

We now parse out the international population from 1948 to 1975, and construct the US nonwhite and white populations from 1948 to 1975.

```
MLBpop_nonwhite <- MLBpop[MLBpop$year %in% 1948:1975, ]
MLBpop_nonwhite$pop <- MLBpop_nonwhite$pop -
    datUSCan[datUSCan$year %in% 1948:1975, ]$pop +
    dat_nonwhite$pop
```

```
MLBpop_white <- datUSCan[datUSCan$year %in% 1948:1975, ]
MLBpop_white$pop <- MLBpop_white$pop - dat_nonwhite$pop
```

We now obtain reductions for the AL which correspond to the slower rate of integration in that league. This figure is a balance between the difference in win shares and Hall of Famers between the two leagues.

```
foo <- approx(x = c(1948, 1950, 1952, 1955, 1960, 1966, 1971, 1975),
                           y = c(0.95, 0.90, 0.80, 0.65, 0.35, 0.35, 0.90, 1),
                           xout = 1948:1975)
AL_interest <- data.frame(year = 1948:1975, interest = foo$y)
```

We now obtain the eligible MLB populations for both the NL and the AL.

```
MLBpop_NL <- MLBpop_AL <- MLBpop
MLBpop_AL[MLBpop_AL$year %in% 1948:1975, ]$pop <-
    MLBpop_white$pop + MLBpop_nonwhite$pop *
    AL_interest$interest
MLBpop <- data.frame(year = MLBpop_NL$year,
  NL_pop = round(MLBpop_NL$pop * 1e6),
  AL_pop = round(MLBpop_AL$pop * 1e6))
MLBpop
```

```
##      year   NL_pop   AL_pop
## 1    1871   405360   405360
## 2    1872   422141   422141
## 3    1873   439098   439098
## 4    1874   456233   456233
## 5    1875   473544   473544
## 6    1876   491033   491033
## 7    1877   508698   508698
## 8    1878   526540   526540
## 9    1879   544559   544559
## 10   1880   562755   562755
## 11   1881   573351   573351
## 12   1882   584032   584032
## 13   1883   594799   594799
## 14   1884   605651   605651
## 15   1885   616588   616588
## 16   1886   627611   627611
## 17   1887   638720   638720
## 18   1888   649913   649913
## 19   1889   661193   661193
## 20   1890   672558   672558
## 21   1891   684008   684008
## 22   1892   695543   695543
## 23   1893   707165   707165
## 24   1894   718871   718871
## 25   1895   730663   730663
## 26   1896   742541   742541
## 27   1897   754504   754504
## 28   1898   766552   766552
## 29   1899   778686   778686
## 30   1900   790906   790906
## 31   1901   837164   837164
## 32   1902   883864   883864
## 33  1903   931008   931008
## 34  1904   978594   978594
## 35  1905  1026623  1026623
## 36  1906  1075094  1075094
## 37  1907  1124009  1124009
## 38  1908  1173366  1173366
## 39  1909  1223165  1223165
## 40  1910  1273408  1273408
## 41  1911  1285430  1285430
## 42  1912  1297508  1297508
## 43  1913  1309642  1309642
## 44  1914  1321833  1321833
## 45  1915  1334080  1334080
## 46  1916  1346383  1346383
## 47  1917  1358742  1358742
## 48  1918  1028368  1028368
## 49  1919   857850   857850
## 50  1920  1047118  1047118
## 51  1921  1276412  1276412
## 52  1922  1440461  1440461
## 53  1923  1462833  1462833
## 54  1924  1485352  1485352
## 55  1925  1508017  1508017
## 56  1926  1530829  1530829
## 57  1927  1553787  1553787
## 58  1928  1576893  1576893
## 59  1929  1600144  1600144
## 60  1930  1364249  1364249
## 61  1931  1814754  1814754
## 62  1932  2275918  2275918
## 63  1933  2292149  2292149
## 64  1934  2340639  2340639
## 65  1935  2398643  2398643
## 66  1936  2439416  2439416
## 67  1937  2533120  2533120
## 68  1938  2646801  2646801
## 69  1939  2768756  2768756
## 70  1940  2816921  2816921
## 71  1941  2276737  2276737
## 72  1942  1458251  1458251
## 73  1943   887221   887221
## 74  1944   895475   895475
## 75  1945  1057711  1057711
## 76  1946  1210262  1210262
## 77  1947  1377835  1377835
## 78  1948  1854585  1819997
## 79  1949  2258928  2199998
## 80  1950  3461974  3361463
## 81  1951  4051961  3879942
## 82  1952  4666942  4408057
## 83  1953  5307297  4945279
## 84  1954  5430369  4982068
## 85  1955  5554706  5015406
## 86  1956  5680309  5029379
## 87  1957  5807177  5038884
## 88  1958  5935312  5043873
## 89  1959  6064712  5044303
## 90  1960  6195378  5040124
## 91  1961  6468909  5279807
## 92  1962  6628739  5405572
## 93  1963  6785149  5527700
## 94  1964  6938139  5646191
## 95  1965  7087709  5761046
## 96  1966  7233859  5872264
## 97  1967  7640656  6337688
## 98  1968  7788170  6723698
## 99  1969  7932264  7119956
## 100 1970  8072937  7526572
## 101 1971  8310781  8040010
## 102 1972  8582724  8368771
## 103 1973  8814351  8666236
## 104 1974  8927904  8851086
## 105 1975  9030398  9030398
## 106 1976  9121833  9121833
## 107 1977  9209665  9209665
## 108 1978  9279284  9279284
## 109 1979  9288148  9288148
## 110 1980  9300767  9300767
## 111 1981  9148554  9148554
## 112 1982  8995012  8995012
## 113 1983  8840142  8840142
## 114 1984  8683942  8683942
## 115 1985  8535206  8535206
## 116 1986  8453202  8453202
## 117 1987  8434685  8434685
## 118 1988  8351942  8351942
## 119 1989  8267853  8267853
## 120 1990  8182416  8182416
## 121 1991  8016379  8016379
## 122 1992  8017135  8017135
## 123 1993  7852834  7852834
## 124 1994  9819558  9819558
## 125 1995 14049153 14049153
## 126 1996 14055176 14055176
## 127 1997 14059660 14059660
## 128 1998 14062602 14062602
## 129 1999 14099888 14099888
## 130 2000 14137174 14137174
## 131 2001 14106406 14106406
## 132 2002 14675844 14675844
## 133 2003 14645283 14645283
## 134 2004 14757944 14757944
## 135 2005 14873698 14873698
## 136 2006 14992545 14992545
## 137 2007 15114485 15114485
## 138 2008 14755161 14755161
## 139 2009 14508949 14508949
## 140 2010 14504848 14504848
## 141 2011 14585671 14585671
## 142 2012 15021811 15021811
## 143 2013 15107129 15107129
## 144 2014 15193193 15193193
## 145 2015 15280003 15280003
## 146 2016 15367557 15367557
## 147 2017 15455857 15455857
## 148 2018 15544903 15544903
## 149 2019 15634693 15634693
## 150 2020 15725229 15725229
## 151 2021 15816510 15816510
```

```
write_csv(MLBpop, file = "datMLBpop.csv")
```

Negro Leagues

Major League Baseball (MLB) elevates Negro Leagues to Major League status between 1920-1948.

We now tally the Negro league eligible population. We first obtain Census data for the population of black/non-white Americans from 1920-1948. We will also need population totals for baseball playing Latin American countries such as Cuba, Dominican Republic, Puerto Rico, and Panama. These countries participated in the Negro leagues [Burgos, 2007, page 112].

US Census data Taken from the 1950 Census, pages 1-91, 1-92, and 1-94. Website Source.

1920 nonwhite males ages 20-29:

```
nonwhiteUS1920 <- 508469 + 443932
nonwhiteUS1920
```

```
## [1] 952401
```

1930 nonwhite males ages 20-29:

```
nonwhiteUS1930 <- 590023 + 535866
nonwhiteUS1930
```

```
## [1] 1125889
```

The 1930 and 1940 tallies are included above. We now interpolate the number of nonwhite US males aged 20-29 from 1900 to 1950 adjusted for interest.

```
foo <- approx(x = c(1920,1930,1940,1950),
  y = c(nonwhiteUS1920 * dat_interest[dat_interest$year == 1920,]$interest,
            nonwhiteUS1930 * dat_interest[dat_interest$year == 1930,]$interest,
            nonwhiteUS1940, nonwhiteUS1950),
  xout = 1920:1948)
datUSnonwhite <- data.frame(year = foo$x, popUSnonwhite = foo$y)
```

Latin American population data The following notes come from pages 111 and 112 Burgos [2007]:

> From the turn of the early twentieth century in Jim Crow America, before there were formally organized Negro leagues, to after the Negro leagues had disintegrated in the 1950s, African American players were sought after and secured for their skills by Latin American baseball entrepreneurs in Cuba, Mexico, Puerto Rico, and else-where in the Spanish-speaking Americas. Conversely, the professional teams and leagues of black baseball extended a much kinder welcome to Latinos than did the white baseball establishment.... As they traveled the black baseball circuit, Latinos would encounter directly the brutal reality of Jim Crow segregation.... As the owner of Negro-league teams and as a scout for the Giants, Alex Pompez introduced the greatest number of talented Latinos to U.S. professional baseball. He also opened up the Negro leagues

> to talent from outside **Cuba**, introducing the first players from the **Dominican Republic**, **Puerto Rico**, and **Panama**.

We are going to define the Negro league eligible population as population of males aged 20-29 that are US nonwhite as tabulated by the Census or from Cuba, Dominican Republic, Panama, or Puerto Rico. There were not many players from Mexico participating in the Negro leagues according to baseball reference.

Cuba Cuban population of males ages 20-24 (% of males) in 1960: 8.84% World bank

Cuban population of males ages 25-29 (% of males) in 1960: 7.51% World bank

Cuban male population in 1960: 3648440 World Bank

Cuban population 1960: 7141241 World Bank

Percentage of Cuban males in 1960:

```
perCuba1960 <- 3648440 / 7141241
perCuba1960
```

```
## [1] 0.5108972
```

total males age 20-29 in 1960:

```
cuba1960 <- round(3648440 * (0.0884 + 0.0751) / 1000) * 1000
cuba1960
```

```
## [1] 597000
```

Estimated Cuban population in 1900: 1600000 Wikipedia

Estimated Cuban male population aged 20-29 in 1900:

```
cuba1900 <- round(1600000 * (0.0884 + 0.0751) * perCuba1960 / 1000) * 1000
cuba1900
```

```
## [1] 134000
```

We now interpolate the interest adjusted number of Cuban males aged 20-29 from 1920 to 1950.

```
foo <- approx(x = c(1900,1960), y = c(cuba1900, cuba1960), n = 61)
dat <- data.frame(year = foo$x, popCuba = foo$y * 0.62)
datCuba <- dat %>% filter(year >= 1920, year <= 1948)
```

Dominican Republic Dominican Republic population of males ages 20-24 (% of males) in 1960: 7.93% World bank

Dominican Republic population of males ages 25-29 (% of males) in 1960: 7.01% World bank

Dominican Republic male population in 1960: 1667973 World Bank

Dominican Republic population 1960: 3294222 World Bank

Percentage of Dominican Republic males in 1960:

```
perDR1960 <- 1667973 / 3294222
perDR1960
```

```
## [1] 0.5063329
```

total males age 20-29 in 1960:

```
dr1960 <- round(1667973 * (0.0793 + 0.0701) / 1000) * 1000
dr1960
```

```
## [1] 249000
```

Estimated Dominican Republic population in 1900: 600000 Wikipedia

Estimated Dominican Republic male population aged 20-29 in 1900:

```
dr1900 <- round(600000 * (0.0793 + 0.0701) * perDR1960 / 1000) * 1000
dr1900
```

```
## [1] 45000
```

We now interpolate the interest adjusted number of Dominican Republic males aged 20-29 from 1920 to 1950.

```
foo <- approx(x = c(1900,1960), y = c(dr1900, dr1960), n = 61)
dat <- data.frame(year = foo$x, popDR = foo$y * 0.85)
datDR <- dat %>% filter(year >= 1920, year <= 1948)
```

Panama Panama population of males ages 20-24 (% of males) in 1960: 8.43% World bank

Panama population of males ages 25-29 (% of males) in 1960: 7.13% World bank

Panama male population in 1960: 578268 World Bank

Panama population 1960: 1133005 World Bank

Percentage of Panama males in 1960:

```
perPanama1960 <- 578268 / 1133005
perPanama1960
```

```
## [1] 0.5103843
```

total males age 20-29 in 1960:

```
panama1960 <- round(578268 * (0.0843 + 0.0713) / 1000) * 1000
panama1960
```

```
## [1] 90000
```

Estimated Panama population in 1900: 263000 Wikipedia

Estimated Panama male population aged 20-29 in 1900:

```
panama1900 <- round(263000 * (0.0843 + 0.0713) * perPanama1960 / 1000) * 1000
panama1900
```

```
## [1] 21000
```

We now interpolate the interest adjusted number of Dominican Republic males aged 20-29 from 1920 to 1950.

```
foo <- approx(x = c(1900,1960), y = c(panama1900, panama1960), n = 61)
dat <- data.frame(year = foo$x, popPanama = foo$y * 0.62)
datPanama <- dat %>% filter(year >= 1920, year <= 1948)
```

Puerto Rico Puerto Rican population of males ages 20-24 (% of males) in 1960: 6.67% World bank

Puerto Rican population of males ages 25-29 (% of males) in 1960: 5.09% World bank

Puerto Rican male population in 1960: 1163631 World Bank

Puerto Rican population 1960: 2358000 World Bank

Percentage of Puerto Rican males in 1960:

```
perPR1960 <- 1163631 / 2358000
perPR1960
```

```
## [1] 0.4934822
```

total males age 20-29 in 1960:

```
pr1960 <- round(1163631 * (0.0667 + 0.0509) / 1000) * 1000
pr1960
```

```
## [1] 137000
```

Estimated Puerto Rican population in 1900: 986000 Wikipedia

Estimated Puerto Rican male population aged 20-29 in 1900:

```
pr1900 <- round(986000 * (0.0667 + 0.0509) * perPR1960 / 1000) * 1000
pr1900
```

```
## [1] 57000
```

We now interpolate the interest adjusted number of Puerto Rican males aged 20-29 from 1920 to 1950.

```
foo <- approx(x = c(1900,1960), y = c(pr1900, pr1960), n = 61)
dat <- data.frame(year = foo$x, popPR = foo$y * 0.85)
datPR <- dat %>% filter(year >= 1920, year <= 1948)
```

Eligible Negro League population We now tabulate the Negro league eligible population by adding up the estimating US nonwite, Cuban, Puerto Rican, Dominican Republic, and Panama tallies of males aged 20-29. We adjust these totals to account for the effect of WWI and WWII.

```
# total
datNegroLeague <- data.frame(year = datDR[, 1], pop =
  round( (datUSnonwhite[, 2] + datCuba[, 2] + datPR[, 2] +
                    datDR[, 2] + datPanama[, 2]) / 1000) * 1000)

# adjust for WWI
datNegroLeague[datNegroLeague$year %in% 1920:1921, ]$pop  <-
    datNegroLeague[datNegroLeague$year %in% 1920:1921, ]$pop *
    c(0.75,0.90)

# adjust for WWII
datNegroLeague[datNegroLeague$year %in% 1941:1948, ]$pop  <-
    datNegroLeague[datNegroLeague$year %in% 1941:1948, ]$pop *
    c(0.80,0.50,0.30,0.30,0.35,0.40,0.45,0.50)

# final tally
datNegroLeague$pop <- round(datNegroLeague$pop)
datNegroLeague
```

```
##    year    pop
## 1  1920 391500
## 2  1921 478800
## 3  1922 542000
## 4  1923 551000
## 5  1924 561000
## 6  1925 570000
## 7  1926 580000
## 8  1927 589000
## 9  1928 599000
## 10 1929 608000
## 11 1930 618000
## 12 1931 640000
## 13 1932 663000
## 14 1933 685000
## 15 1934 707000
## 16 1935 729000
## 17 1936 752000
## 18 1937 774000
## 19 1938 796000
## 20 1939 818000
## 21 1940 841000
## 22 1941 681600
## 23 1942 431500
## 24 1943 262200
## 25 1944 265500
## 26 1945 313600
## 27 1946 362800
## 28 1947 413100
## 29 1948 464500
```

```
write_csv(datNegroLeague, file = "datNegroLeague.csv")
```

References

Katherine E Baird. Cuban baseball: Ideology, politics, and market forces. *Journal of Sport and Social Issues*, 29(2): 164–183, 2005.

Adrian Burgos. *Playing America's Game*. University of California Press, 2007.

Roberto González Echevarría. *The pride of Havana: A history of Cuban baseball*. Oxford University Press, USA, 1999.

Daniel J Eck. Challenging nostalgia and performance metrics in baseball. *Chance*, 33(1):16–25, 2020a.

Daniel J Eck. Supporting data analysis for "challenging nostalgia and performance metrics in baseball". 2020b.

Ron Fimrite. In cuba, it's viva el grand old game. *Sports Illustrated*, 6:68–80, 1977.

Milton H. Jamail. *Venezuelan Bust, Baseball Boom: Andrés Reiner and Scouting on the New Frontier*. University of Nebraska Press, 2008.

Steven Ranieri. Put baseball on the free trade agenda. 1992.

Martin B Schmidt and David J Berri. Concentration of playing talent: evolution in major league baseball. *Journal of Sports Economics*, 6(4):412–419, 2005.