Quantifying NBA Home Court Advantage and the Effect of Travel Distance on it Since 1970

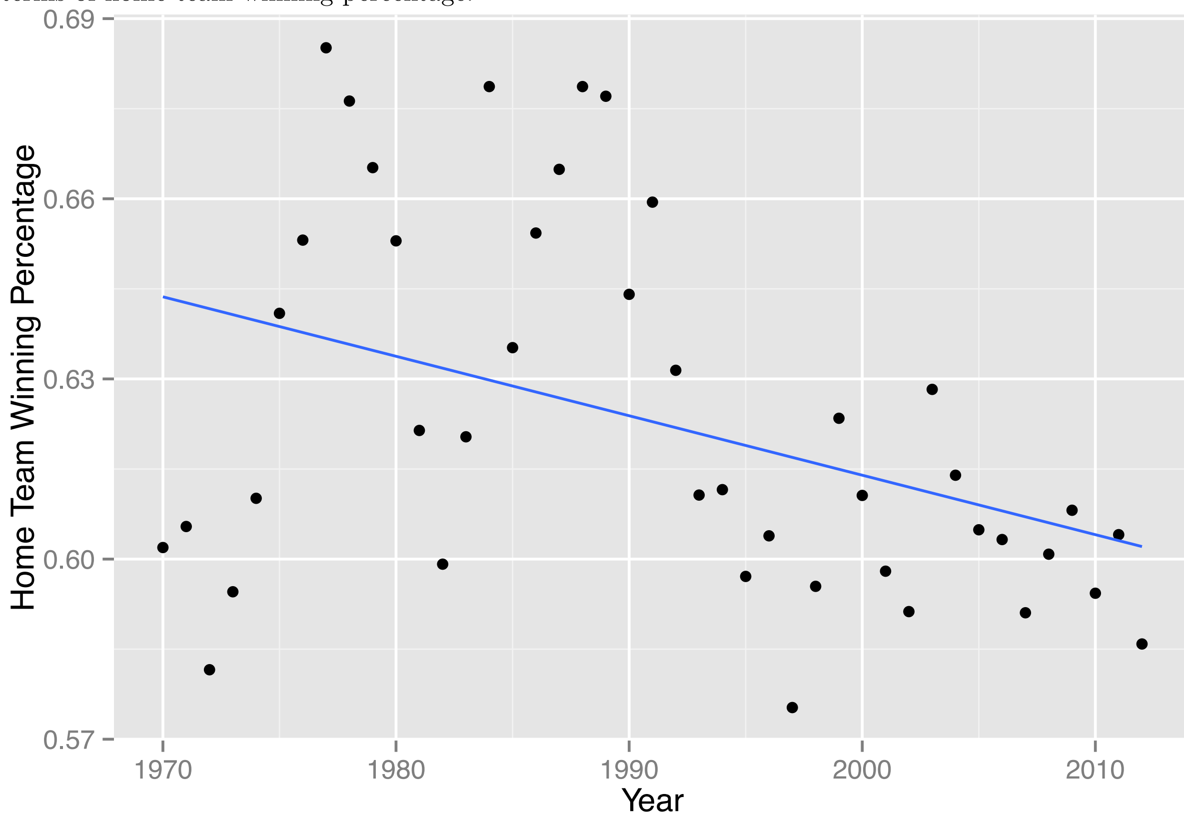
Shane Fenske

May 4, 2015

Abstract: ESPN National Basketball Association writer Tom Haberstroh recently explored the idea becoming popular in NBA circles that home-court advantage is becoming less and less of an advantage. His article's title–"Home-court advantage? Not so much."–implies that playing at home is no longer an advantage. The article quantified advantage only in terms of home team winning percentage, a metric that could be skewed by an increase in league-wide parity. This paper attempts to quantify home-court advantage in a way which takes team strength into account by constructing a linear model for each NBA season from 1970 to 2012. Each model attempts to quantify home-court advantage in terms of extra points the home team is predicted to score in an individual game compared to if the same game were to be played at a neutral site. Additionally, this paper explores travel distance's effect on home-court advantage with the idea being that an improvement in travel accommodations could be partially responsible for the phenomena Haberstroh observed. The paper confirms Haberstroh's conclusion of a decrease in home court advantage over the past four decades of NBA play, but finds that playing at home is still a significant advantage. Furthermore, this paper does not find convincing evidence that the effect of opponents' travel distance on home-court advantage has decreased in the 43 year timespan studied.

Introduction

In recent months, the idea that home-court advantage in the NBA is becoming less and less of an advantage has become a popular conversation topic in NBA circles. ESPN writer, Tom Haberstroh explored it statistically in his article "Home-court advantage? Not so much." The article quantified advantage only in terms of home team winning percentage.



Team strengths did not play a role in his analysis, leaving increased league parity to possibly explain what could be only an illusion of decreasing home-court advantage. It would be better to quantify home-court advantage in a way which takes team strength into account. We will do this by evaluating home-court advantage in terms of extra points the home team is predicted to score in an individual game, compared to if the same game were to be played at a neutral site. This will allow us to better see the change in the advantage home teams have had over the 43 years studied and to determine if playing at home has really become a "disadvantage" as implied by Haberstroh's choice of article title and his use of "#HomeCourtDisadvantage" on Twitter.

At the 2015 MIT Sloan Sports Analytics Conference, former NBA head coach Mike D'Antoni answered a question regarding concerns of travel fatigue on NBA players with: "Are you worried [their] shrimp cocktail on the plane is going to be a little bit warmer?" The implication of this answer was that, in D'Antoni's opinion, the luxurious modes of travel employed by NBA teams today make traveling a minimal hindrance to performance. Travel was much different for NBA players of the 1970's who frequently flew business class. This change in travel accommodations is something not considered in Haberstroh's article. We will explore it in this paper by examining the effect of miles traveled by the away team on a home team's advantage over the course of the 43 years studied.

Data Cleaning
The dataset I started my analyis with contained the box score of every NBA game played between the 1970 and 2012 seasons. It can be found in Christopher Long's GitHub at: https://github.com/octonion/basketball-public/blob/master/bbref/csv/team_logs_1970-2012.csv. I modified the dataset by labeling each column, removing the columns which I did not need, removing the duplicate entry of each game, and adding latitude and longitude coordinates for both the home and away team involved in each game. The code for this initial data cleaning can be found in the appendix of this paper.

The latitude and longitude coordinates I added are utilized here, as I use Vincenty's great-circle distance formula (as it is implemented in the package "geosphere") to add a column that contains the distance in miles between the two cities involved in each game in the dataset.

```
for(i in 1:nrow(data)) {
  home_loc <- c(data$H_long[i],data$H_lat[i])
  away_loc <- c(data$A_long[i],data$A_lat[i])
  dist <- distVincentyEllipsoid(home_loc, away_loc, a=6378137, b=6356752.3142,
                                f=1/298.257223563)*0.000621371
  data$travel_dist[i] = dist
}
data <- subset(data, select=-c(A_long, A_lat, H_long, H_lat))
data$diff <-   data$home_PTS - data$away_PTS
```

This results in the final dataset I used, as sampled here.

```
head(data,3)
```

```
##   year away_team home_team away_result away_PTS home_PTS travel_dist diff
## 1 1970       PHI       CIN           W      134      123   502.73615  -11
## 4 1970       PHI       BAL           W      129      105    90.00508  -24
## 9 1970       PHI       DET           L      128      134   442.44829    6
```

Methods and Results

In order to quantify home-court advantage in a way other than raw percentage of home team wins, we will build a model for each year of games in our data set. Each model will predict scoring differential based on a matrix constructed with as many rows as there were games during that season. Each row (representing a single game) is all 0's aside from a -1 in the away team's column and a 1 in the home team's. Additionally, a differential column holds the result of the game in the form of home team's score minus the away team's score.

```
j <- 1
models <- list()
for(year in 1970:2012) {
  y <- data[data$year==year,]
  tot_teams <- length(unique(y$home_team))
  z <- matrix(0, nrow=nrow(y), ncol=tot_teams)
  colnames(z) <- sort(unique(y$home_team))

  for (i in 1:nrow(y)) {
    z[i,y$home_team[i]==colnames(z)] <- 1
    z[i,y$away_team[i]==colnames(z)] <- -1
  }
  z <- as.data.frame(z)
  z$diff <- y$diff
  z <- z[,-1]
  lm.1 <- lm(diff ~ ., data=z)
  models[[j]] <- lm.1
  j <- j+1
}
```

2012 data and portion of 2012 matrix to illustrate:

```
head(data[data$year==2012,],4)[,-7]
```

```
##       year away_team home_team away_result away_PTS home_PTS diff
## 85008 2012       DEN       DAL           W      115       93  -22
## 85010 2012       DEN       POR           L      102      111    9
## 85011 2012       DEN       LAL           L       89       92    3
## 85015 2012       DEN       NOH           W       96       88   -8
```

```
head(z,4)[,c(5:7,12,18,24,30)]
```

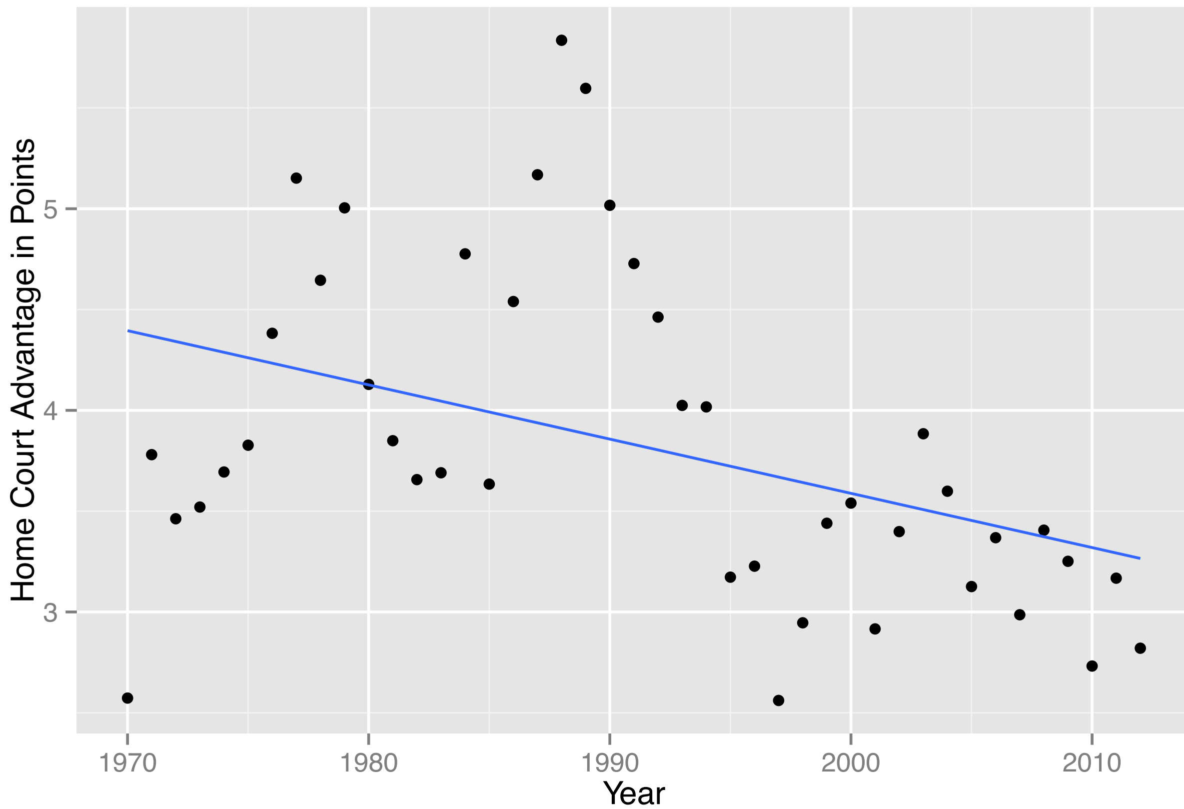
```
##   DAL DEN DET LAL NOH POR diff
## 1   1  -1   0   0   0   0  -22
## 2   0  -1   0   0   0   1    9
## 3   0  -1   0   1   0   0    3
## 4   0  -1   0   0   1   0   -8
```

Each model contains coefficients for each team's strength (relative to the first alphabetical team, Atlanta, which has team strength of 0). Each coefficient can be thought of as the score differential if the team were to play Atlanta at a neutral site. To calculate the predicted differential of a neutral site game played between teams A and B, take A's coefficient and subtract B's from it.
Here is the 1970 model:

```r
summary(models[[1]])
```

```
##
## Call:
## lm(formula = diff ~ ., data = z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.936  -7.870   0.960   8.733  34.380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5728     0.5388   4.775 2.36e-06 ***
## BAL           0.8270     1.9541   0.423   0.6723
## BOS          -2.7824     1.9373  -1.436   0.1516
## CHI          -2.8629     2.0128  -1.422   0.1555
## CIN          -3.0637     2.0125  -1.522   0.1285
## DET          -3.7056     1.9851  -1.867   0.0625 .
## LAL           0.7421     1.9132   0.388   0.6983
## MIL           3.2585     1.9553   1.667   0.0962 .
## NYK           7.8720     1.9316   4.075 5.33e-05 ***
## PHI           2.2352     1.9430   1.150   0.2505
## PHO          -2.9935     1.9662  -1.523   0.1285
## SDR          -3.3015     1.9483  -1.695   0.0908 .
## SEA          -3.8161     1.9540  -1.953   0.0514 .
## SFW          -4.5482     1.9426  -2.341   0.0196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.27 on 506 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.1249
## F-statistic: 6.697 on 13 and 506 DF,  p-value: 6.572e-12
```

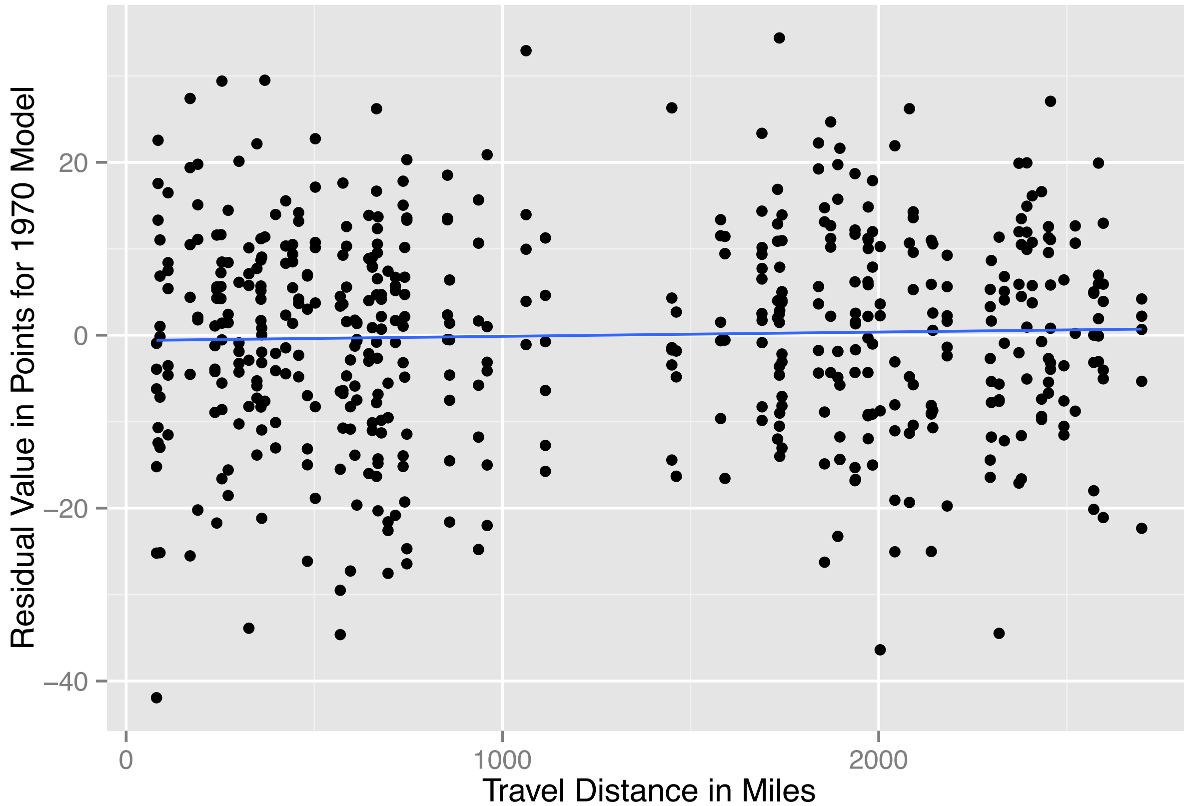
With this thought in mind it is easy to see the intercept of each model is then what is predicted to be the point advantage for a team playing a given game at home. A plot of these intercepts is revealing:



This confirms Haberstroh's hypothesis that home-court advantage is declining and, with team strength accounted for, rules out the possibility that the decline was an allusion based on increased parity. However, this allows us to see that contrary to the recently popular Twitter hashtag, "#HomeCourtDisadvantage", playing at home is still nearly a three point advantage to the home team.

Next we will explore the question of improved travel conditions. Before we can explore whether traveling has become less of a burden on away teams, we must find a way quantify travel distance's effect.

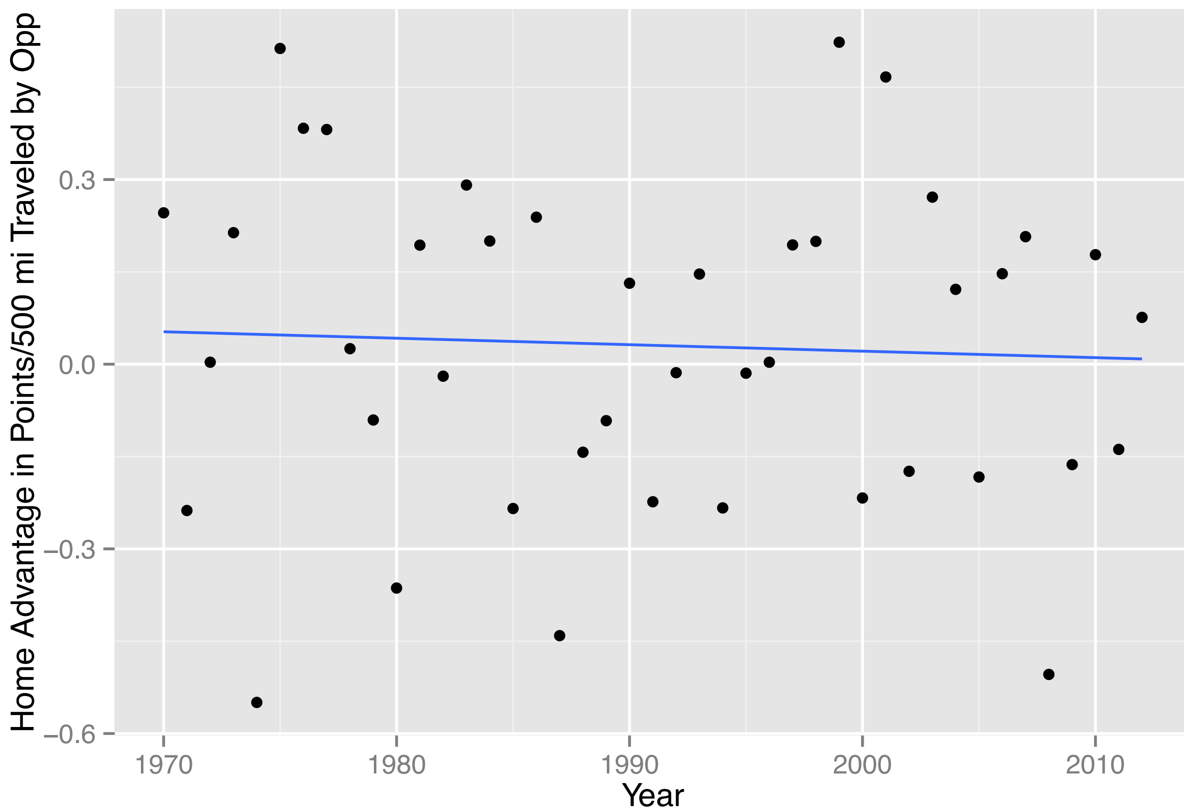
If we plot each game's residual in the 1970 model against the travel distance of the away team in said game, we notice a slightly positive sloping relation; leading us to believe that perhaps travel distance would be wise to include in a future model.



The slope of this regression line can be used as a quantification in points of the effect of each mile travel by the away team on the home team's expected points for games played in the given year.

Let us find this slope for each year in the dataset, multiply each of these slopes by 500 in order to use more consequential numbers, and then plot the result.

```
j <- 1
data$trav_coeffs <-rep(0,nrow(data))
for(year in 1970:2012) {
  lm.1 <- lm(summary(models[[j]])$residuals ~ data[data$year==year,]$travel_dist)
  trav_coeff <- summary(lm.1)$coefficients[2]
  data[data$year==year,]$trav_coeffs <- trav_coeff
  j <- j+1
}
x$trav <- unique(data$trav_coeffs) * 500
ggplot(x, aes(yr,trav)) + geom_point() +  labs(x="Year",y = "Home Advantage in Points/500 mi Traveled by
```



Despite it being negative, the small magnitude of the slope of this plot's regression line means there is no conclusive evidence to say that a mile traveled by an away team is less of a burden now than it was 40 years ago. Furthermore, the random scattering of points and the large number of them that are negative, would suggest that perhaps travel distance of the opponent has always had very little effect on a team's home court advantage.

Appendix

The column names were not included in the dataset, but I was able to gleen what each column was by comparing data entries to the corresponding game box score on basketballreference.com.

```
col_names = c("year","away_team", "gm_num_a", "gm_num_b", "date", "at", "home_team", "away_result",
              "MP","away_FG","away_FGA","away_3P","away_3PA", "away_FT", "away_FTA",
              "away_ORB", "away_TRB",  "away_AST", "away_STL","away_BLK", "away_TOV",
              "away_PF",  "away_PTS","home_FG","home_FGA", "home_3P","home_3PA",
              "home_FT", "home_FTA","home_ORB", "home_TRB",  "home_AST", "home_STL",
              "home_BLK", "home_TOV",  "home_PF",   "home_PTS")

data.0 <- read.csv("team_logs_1970-2012.csv", as.is = T,col.names = col_names)
```

This data set includes every game played twice (once for each team). I'll fix this by subsetting the data to only include the away team's entry for each game. Additionally, I will remove all games played at neutral sites.

```
data.0 <- data.0[data.0$at=="@",]
```

This subsetting makes the game number columns and "at" columns inaccurate and not useful, respectively. I will remove them. I will also remove all the colums that will not be used in our analysis.

```
data <- subset(data.0, select=-c(gm_num_a, gm_num_b, at))
data <- subset(data, select=c(year, away_team, home_team, away_result, away_PTS,
                              home_PTS))
```

Now I will add the latitude and longitude of each city by hand.

```
data$A_long <- rep(0,43311)
data$A_lat <- rep(0,43311)
data$H_long <- rep(0,43311)
data$H_lat <- rep(0,43311)
data$travel_dist <- rep(0,43311)

data[data$away_team=="ATL",]$A_long <- 84.3900
data[data$away_team=="ATL",]$A_lat <- 33.7550
data[data$home_team=="ATL",]$H_long <-84.3900
data[data$home_team=="ATL",]$H_lat <-33.7550

data[data$away_team=="BAL",]$A_long <- 76.6167
data[data$away_team=="BAL",]$A_lat <- 39.2833
data[data$home_team=="BAL",]$H_long <-76.6167
data[data$home_team=="BAL",]$H_lat <-39.2833

data[data$away_team=="BOS",]$A_long <- 71.0589
data[data$away_team=="BOS",]$A_lat <- 42.3601
data[data$home_team=="BOS",]$H_long <-71.0589
data[data$home_team=="BOS",]$H_lat <-42.3601

data[data$away_team=="BUF",]$A_long <- 78.8494
```

```
data[data$away_team=="BUF",]$A_lat <- 42.9047
data[data$home_team=="BUF",]$H_long <-78.8494
data[data$home_team=="BUF",]$H_lat <-42.9047

data[data$away_team=="CAP",]$A_long <- 77.0164
data[data$away_team=="CAP",]$A_lat <- 38.9047
data[data$home_team=="CAP",]$H_long <-77.0164
data[data$home_team=="CAP",]$H_lat <-38.9047

data[data$away_team=="CHA",]$A_long <- 80.8433
data[data$away_team=="CHA",]$A_lat <- 35.2269
data[data$home_team=="CHA",]$H_long <-80.8433
data[data$home_team=="CHA",]$H_lat <-35.2269

data[data$away_team=="CHH",]$A_long <- 80.8433
data[data$away_team=="CHH",]$A_lat <- 35.2269
data[data$home_team=="CHH",]$H_long <- 80.8433
data[data$home_team=="CHH",]$H_lat <- 35.2269

data[data$away_team=="CHI",]$A_long <- 87.6847
data[data$away_team=="CHI",]$A_lat <- 41.8369
data[data$home_team=="CHI",]$H_long <-87.6847
data[data$home_team=="CHI",]$H_lat <-41.8369

data[data$away_team=="CIN",]$A_long <- 84.5167
data[data$away_team=="CIN",]$A_lat <- 39.1
data[data$home_team=="CIN",]$H_long <-84.5167
data[data$home_team=="CIN",]$H_lat <-39.1

data[data$away_team=="CLE",]$A_long <- 81.6697
data[data$away_team=="CLE",]$A_lat <- 41.4822
data[data$home_team=="CLE",]$H_long <-81.6697
data[data$home_team=="CLE",]$H_lat <-41.4822

data[data$away_team=="DAL",]$A_long <- 96.7970
data[data$away_team=="DAL",]$A_lat <- 32.7767
data[data$home_team=="DAL",]$H_long <-96.7970
data[data$home_team=="DAL",]$H_lat <-32.7767

data[data$away_team=="DEN",]$A_long <- 104.9903
data[data$away_team=="DEN",]$A_lat <- 39.7392
data[data$home_team=="DEN",]$H_long <-104.9903
data[data$home_team=="DEN",]$H_lat <-39.7392

data[data$away_team=="DET",]$A_long <- 83.0458
data[data$away_team=="DET",]$A_lat <- 42.3314
data[data$home_team=="DET",]$H_long <-83.0458
data[data$home_team=="DET",]$H_lat <-42.3314

data[data$away_team=="GSW",]$A_long <- 122.2708
data[data$away_team=="GSW",]$A_lat <- 37.8044
data[data$home_team=="GSW",]$H_long <-122.2708
data[data$home_team=="GSW",]$H_lat <-37.8044
```

```
data[data$away_team=="HOU",]$A_long <- 95.3698
data[data$away_team=="HOU",]$A_lat <- 29.7604
data[data$home_team=="HOU",]$H_long <- 95.3698
data[data$home_team=="HOU",]$H_lat <-29.7604

data[data$away_team=="IND",]$A_long <- 86.1480
data[data$away_team=="IND",]$A_lat <- 39.7910
data[data$home_team=="IND",]$H_long <-86.1480
data[data$home_team=="IND",]$H_lat <-39.7910

data[data$away_team=="KCK",]$A_long <- 94.5783
data[data$away_team=="KCK",]$A_lat <- 39.0997
data[data$home_team=="KCK",]$H_long <-94.5783
data[data$home_team=="KCK",]$H_lat <-39.0997

data[data$away_team=="KCO",]$A_long <- 94.5783
data[data$away_team=="KCO",]$A_lat <- 39.0997
data[data$home_team=="KCO",]$H_long <-94.5783
data[data$home_team=="KCO",]$H_lat <-39.0997

data[data$away_team=="LAC",]$A_long <- 118.2500
data[data$away_team=="LAC",]$A_lat <- 34.0500
data[data$home_team=="LAC",]$H_long <-118.2500
data[data$home_team=="LAC",]$H_lat <-34.0500

data[data$away_team=="LAL",]$A_long <- 118.2500
data[data$away_team=="LAL",]$A_lat <- 34.0500
data[data$home_team=="LAL",]$H_long <-118.2500
data[data$home_team=="LAL",]$H_lat <-34.0500

data[data$away_team=="MEM",]$A_long <- 89.9711
data[data$away_team=="MEM",]$A_lat <- 35.1174
data[data$home_team=="MEM",]$H_long <-89.9711
data[data$home_team=="MEM",]$H_lat <-35.1174

data[data$away_team=="MIA",]$A_long <- 80.2089
data[data$away_team=="MIA",]$A_lat <- 25.7753
data[data$home_team=="MIA",]$H_long <-80.2089
data[data$home_team=="MIA",]$H_lat <-25.7753

data[data$away_team=="MIL",]$A_long <- 87.9500
data[data$away_team=="MIL",]$A_lat <- 43.0500
data[data$home_team=="MIL",]$H_long <-87.9500
data[data$home_team=="MIL",]$H_lat <-43.0500

data[data$away_team=="MIN",]$A_long <- 93.2650
data[data$away_team=="MIN",]$A_lat <- 44.9778
data[data$home_team=="MIN",]$H_long <-93.2650
data[data$home_team=="MIN",]$H_lat <-44.9778

data[data$away_team=="NJN",]$A_long <- 74.1726
data[data$away_team=="NJN",]$A_lat <- 40.7242
data[data$home_team=="NJN",]$H_long <-74.1726
```

```
data[data$home_team=="NJN",]$H_lat <-40.7242

data[data$away_team=="NOH",]$A_long <- 90.0667
data[data$away_team=="NOH",]$A_lat <- 29.9500
data[data$home_team=="NOH",]$H_long <-90.0667
data[data$home_team=="NOH",]$H_lat <-29.9500

data[data$away_team=="NOJ",]$A_long <- 90.0667
data[data$away_team=="NOJ",]$A_lat <- 29.9500
data[data$home_team=="NOJ",]$H_long <-90.0667
data[data$home_team=="NOJ",]$H_lat <-29.9500

data[data$away_team=="NOK",]$A_long <- 97.5350
data[data$away_team=="NOK",]$A_lat <- 35.4822
data[data$home_team=="NOK",]$H_long <-97.5350
data[data$home_team=="NOK",]$H_lat <-35.4822

data[data$away_team=="NYK",]$A_long <- 74.0059
data[data$away_team=="NYK",]$A_lat <- 40.7127
data[data$home_team=="NYK",]$H_long <-74.0059
data[data$home_team=="NYK",]$H_lat <-40.7127

data[data$away_team=="NYN",]$A_long <- 73.3
data[data$away_team=="NYN",]$A_lat <- 40.8
data[data$home_team=="NYN",]$H_long <-73.3
data[data$home_team=="NYN",]$H_lat <-40.8

data[data$away_team=="OKC",]$A_long <- 97.5350
data[data$away_team=="OKC",]$A_lat <- 35.4822
data[data$home_team=="OKC",]$H_long <-97.5350
data[data$home_team=="OKC",]$H_lat <-35.4822

data[data$away_team=="ORL",]$A_long <- 81.2989
data[data$away_team=="ORL",]$A_lat <- 28.4158
data[data$home_team=="ORL",]$H_long <-81.2989
data[data$home_team=="ORL",]$H_lat <-28.4158

data[data$away_team=="PHI",]$A_long <- 75.1667
data[data$away_team=="PHI",]$A_lat <- 39.9500
data[data$home_team=="PHI",]$H_long <-75.1667
data[data$home_team=="PHI",]$H_lat <-39.9500

data[data$away_team=="PHO",]$A_long <- 112.0667
data[data$away_team=="PHO",]$A_lat <- 33.4500
data[data$home_team=="PHO",]$H_long <-112.0667
data[data$home_team=="PHO",]$H_lat <-33.4500

data[data$away_team=="POR",]$A_long <- 122.6819
data[data$away_team=="POR",]$A_lat <- 45.5200
data[data$home_team=="POR",]$H_long <-122.6819
data[data$home_team=="POR",]$H_lat <-45.5200

data[data$away_team=="SAC",]$A_long <- 121.4689
```

```
data[data$away_team=="SAC",]$A_lat <- 38.5556
data[data$home_team=="SAC",]$H_long <-121.4689
data[data$home_team=="SAC",]$H_lat <-38.5556

data[data$away_team=="SAS",]$A_long <- 98.5000
data[data$away_team=="SAS",]$A_lat <- 29.4167
data[data$home_team=="SAS",]$H_long <-98.5000
data[data$home_team=="SAS",]$H_lat <-29.4167

data[data$away_team=="SDC",]$A_long <- 117.1625
data[data$away_team=="SDC",]$A_lat <- 32.7150
data[data$home_team=="SDC",]$H_long <-117.1625
data[data$home_team=="SDC",]$H_lat <-32.7150

data[data$away_team=="SDR",]$A_long <- 117.1625
data[data$away_team=="SDR",]$A_lat <- 32.7150
data[data$home_team=="SDR",]$H_long <-117.1625
data[data$home_team=="SDR",]$H_lat <-32.7150

data[data$away_team=="SEA",]$A_long <- 122.3331
data[data$away_team=="SEA",]$A_lat <- 47.6097
data[data$home_team=="SEA",]$H_long <-122.3331
data[data$home_team=="SEA",]$H_lat <-47.6097

data[data$away_team=="SFW",]$A_long <- 122.4167
data[data$away_team=="SFW",]$A_lat <- 37.7833
data[data$home_team=="SFW",]$H_long <-122.4167
data[data$home_team=="SFW",]$H_lat <-37.7833

data[data$away_team=="TOR",]$A_long <- 79.4000
data[data$away_team=="TOR",]$A_lat <- 43.7000
data[data$home_team=="TOR",]$H_long <-79.4000
data[data$home_team=="TOR",]$H_lat <-43.7000

data[data$away_team=="UTA",]$A_long <- 111.5000
data[data$away_team=="UTA",]$A_lat <- 39.5000
data[data$home_team=="UTA",]$H_long <-111.5000
data[data$home_team=="UTA",]$H_lat <-39.5000

data[data$away_team=="VAN",]$A_long <- 123.1207
data[data$away_team=="VAN",]$A_lat <- 49.2827
data[data$home_team=="VAN",]$H_long <-123.1207
data[data$home_team=="VAN",]$H_lat <-49.2827

data[data$away_team=="WAS",]$A_long <- 77.0164
data[data$away_team=="WAS",]$A_lat <- 38.9047
data[data$home_team=="WAS",]$H_long <-77.0164
data[data$home_team=="WAS",]$H_lat <- 38.9047

data[data$away_team=="WSB",]$A_long <- 77.0164
data[data$away_team=="WSB",]$A_lat <- 38.9047
data[data$home_team=="WSB",]$H_long <-77.0164
data[data$home_team=="WSB",]$H_lat <-38.9047
```

```
home_loc <- c(0,0)
away_loc <-c(0,0)
```

These are the libraries used in my paper.

```
library(geosphere)
library(ggplot2)
library(gridExtra)
```

This code was used in making the plot in the Introduction.

```
adv_per <- rep(0,43)
i <- 1
for(year in 1970:2012) {
  y <-  data[data$year==year,]$away_result == "L"
  home_win <- sum(y) / length(y)
  adv_per[i] <-home_win
  i <- i+1
}

yr <- unique(data$year)
x <- data.frame(yr, adv_per)
```

This code was used in making the plot of home-court advantage coefficients.

```
j <- 1
data$coeffs <-rep(0,nrow(data))
for(year in 1970:2012) {
  coeff <- summary(models[[j]])$coefficients[1]
  data[data$year==year,]$coeffs <- coeff
  j <- j+1
}
x$adv <- unique(data$coeffs)
```