Technical Write Up

Andy Ki, Shu Amano

2018/12/18

Contents

Document Options 1
Load Packages 1

Abstract **2**

Introduction **2**

Data **2**
Load Datasets 3
Wrangling elevation data 3
Combine countrycode and temphistory data 4
Clean up countrytemp 4
Clean up dates of olymweat 4
Combine athletes and region 5
Combine olymweat2 and athletes and filter out indoor sports 5
Combine olympics1 with countrytemp 5
Calculate difference between host city temperature and home country temperature 6
Combine elevation data with olympics3 and compute elevation difference 7
Clean up medal data 7
Create dataset that contains total medals for each olympics 7
Calculate percentage of medals won by each country 7
Get rid of distinctions between each athletes and focus on country level performance 8
Combined join medalratio data onto temperature/elevation data 8

Analysis **8**
Plots exploring effects of elevation/temperature difference on medal ratios 8

Results **9**
Create a linear mixed model 9

Conclusion **11**

Document Options

Load Packages

```r
require(mosaic)
require(tidyverse)
require(lubridate)
require(rvest)
require(lme4)
```

Abstract

In the Olympics, regulations and rules define the games, facilitating a fair competitive environment. In this investigation, we explore uncontrollable environmental factors (temperature & elevation) that might influence an athlete's performance. Specifically, we assess performance on a country-level scale, looking at how a country's proportion of medals taken from one year's games fluctuates depending on different Olympic host cities. In our initial modeling attempt, multiple linear regression failed the independence condition, forcing us to undertake linear mixed modelling. From this new model, we found no significant relationship between environmental factors (temperature, elevation) and country's performance. However, our model uses only Summer Olympics data and a small sample size, which limit the model's conclusions. Adding Winter Olympics data and other world-class competitions to our model would be a promising future step in the topic.

Introduction

With our Olympic athlete data and country-level weather data (plus weather data specific to host cities), we looked to answer whether a difference between a host city's and participating countries' elevations/temperatures had a significant positive or negative relationship with the country's medal ratio (proportion of medals won by that country for that game) for that year's games.

With the 2020 Tokyo Olympics quickly approaching, our team looked back to Tokyo's most recent summer and its record-breaking temperatures. Because the Olympics is highly regulated and standardized, we hypothesized that factors that cannot be controlled such as extreme temperatures would have an adverse effect on a country's medal performance. In addition to temperature, we found elevation to be a common environmental factor that elite athletes take into account, especially in a new trend of 'altitude training'. Through altitude training, athletes simulate low oxygen supply while practicing, in an effort to stress-test their performance. Mexico City's 1968 games are an example where the high altitude and low oxygen content disrupted many athletes from performing their best.

We found that a basic multiple linear regression model cannot take into account the inherent complexities of the Olympics. Within our observations, there were correlations by country, which is expected since athletes from the same country are generally expected to perform similarly. Thus, we moved forward to use linear mixed models (LMM) to account the correlation among observations within the same country.

In our LMM, which models medal ratio using temperature difference and elevation difference, we found neither predictors to have significant explanatory power for our response variable.

Data

In this section, we will describe the process we employed to create the dataset used in the final analysis. We began this process by deciding that the final dataset will contain following variables in each row: Country, Year, Elevation Difference, Temperature Difference and Medal Ratio. This decision was made by thinking which data would most effectively answer our research question.

We started off with 7 different datasets: - regions dataset (NOC_Region) - Contained the NOC country code and name of the country. - Collected from Kaggle page "Olympic history data: thorough analysis" by Randi H. Griffin - athletes dataset (Olympics_Athletes) - Contained the NOC country code, Year, Season, Host City, Team, Sports and Medal("Bronze", "Silver", "Gold", "NA") for every athletes who competed in any Olympics before and including the 2016 Summer Olympics. - Collected from Kaggle page "Olympic History Data: Thorough Analysis" by Randi H. Griffin - countryCode dataset (ISO-CountryCode) - Contained the ISO country code and country name for every country. - Collected from https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv - temphistory dataset (Historical_Temp_Data) - Contained the ISO country code and historic average temperature (degree

Celsius) for each month over the past 38 years. - Collected from the World Bank's Climate Change Knowledge Portal
http://sdwebx.worldbank.org/climateportal/index.cfm?page=country_historical_climate - olymwweather dataset (Olympics_Weather_Data) - Contained year, city, season(Summer of Winter), starting date, ending date, number of days, altitude of host city(m), average monthly temperature in the host city for the durations of the olympics(degree Celsius). If the olympic was across two months, the dataset contained average temperature for both months. - Collected from Berkley Earth, a climate change research institution http://berkeleyearth.lbl.gov/city-list/ - Indoor/Outdoor dataset (indooroutdoorClass) - Contained binary variable for whether a sport is played indoor or outdoor for every sports in the athletes dataset - We created this dataset by looking at each sports in wikipedia. If a sport was played both indoor and outdoor depending on competition, we labelled the sport as outdoor. - Elevation dataset - Contained average elevation for every country in meters - Scraped from Wikipedia "List of Countries by Average Elevation" page

Load Datasets

```r
regions <- read_csv("Data/NOC_Region.csv")
athletes <- read_csv("Data/Olympics_Athletes.csv")
countrycode <- read_csv("Data/ISOCountryCode.csv")
temphistory <- readxl::read_xlsx("Data/Historical_Temp_Data.xlsx", sheet = 2)
olymweat <- readxl::read_xlsx("Data/Olympics_Weather_Data.xlsx")
inout <- Class.csv <- read_csv("Data/indooroutdoorClass.csv")
```

Wrangling elevation data

```r
# Create function to extract elevation data from wikipedia
wikiTable <- function(source) {
  read_html(source)%>%
    html_nodes("table.wikitable") %>%
    html_table(fill=T)%>%
    magrittr::extract2(1)
}

# Read in wikipedia table for average elevation per country
elevation <- wikiTable("https://en.wikipedia.org/wiki/List_of_countries_by_average_elevation#cite_note-

# Clean up format of elevation value
elelist <- strsplit(elevation$Elevation, "m")
vec <- c()
for (i in 1:length(elelist)){
  vec[i] <- elelist[[i]][1]
}
elevation$Elevation <- vec
elevation <- elevation %>%
  mutate(Elevation = parse_number(Elevation))

# Change format of name of countries
pos <- which(elevation$Country %in% c("Trinidad and Tobago","United Kingdom","United States"))
elevation[pos,1] <- c("Trinidad","UK", "USA")
```

Combine countrycode and temphistory data

```
# Join countrycode and temphistory
countrytemp <- left_join(temphistory, countrycode, by = c("ISO_3DIGIT" = "alpha-3")) %>%
  select(-c(16:24), -`ISO_3DIGIT`) %>%

# Rename each month column so that wrangling is easier later on
  plyr::rename(c("Jan_Temp" = "1", "Feb_temp" = "2",
                 "Mar_temp" = "3", "Apr_Temp" = "4",
                 "May_temp" = "5", "Jun_Temp" = "6",
                 "July_Temp" = "7", "Aug_Temp" = "8",
                 "Sept_temp" = "9", "Oct_temp" = "10",
                 "Nov_Temp" = "11", "Dec_temp" = "12"))
```

Joined country code and temperature history by ISO, so that each average monthly temperature was linked to a country name instead of the ISO country code. We reformatted some of the the names used in the the created dataset to match the country names used in other datasets. An example would be reformatting "United States of America" to USA. We called this new dataset "countrytemp".

Clean up countrytemp

```
# Rename some countries to standardize country name format
countrytemp <- countrytemp %>%
  mutate(name = ifelse(name == "Bolivia (Plurinational State of)", "Bolivia", name),
         name = ifelse(name == "Czechia", "Czech Republic", name),
         name = ifelse(name == "Korea (Republic of)", "South Korea", name),
         name = ifelse(name == "Russian Federation", "Russia", name),
         name = ifelse(name == "Trinidad and Tobago", "Trinidad", name),
         name = ifelse(name == "United Kingdom of Great Britain and Northern Ireland", "UK", name),
         name = ifelse(name == "United States of America", "USA", name),
         name = ifelse(name == "Venezuela (Bolivarian Republic of)", "Venezuela", name),
         name = ifelse(name == "Viet Nam", "Vietnam", name))
```

Cleaned up the names of countries in countrytemp.

Clean up dates of olymweat

```
# Create new variables for starting/ending months and days
olymweat2 <- olymweat %>%
  rename("StartDate" = `Start Date`,
         "EndDate" = `End Date`) %>%
  mutate(StartDate = as.character(StartDate),
         EndDate = as.character(EndDate),
         StartMonth = strsplit(StartDate, split = "-")[[1]][2],
         EndMonth = strsplit(EndDate, split = "-")[[1]][2],
         StartDay = strsplit(StartDate, split = "-")[[1]][3],
         EndDay = strsplit(EndDate, split = "-")[[1]][3]) %>%

# Get rid of original StartDate and EndDate columns
  select(-StartDate, -EndDate)
```

Next, we made changes to the dates contained in the olymweather's dataset. We created new variables for the starting month, ending month, starting day, ending day of the each olympics game. We called this "olymweather2".

Combine athletes and region

```r
athletes2 <- athletes %>%
  left_join(regions, by = c('NOC' = 'NOC')) %>%
  select(-notes, -NOC, -Games)
```

We then joined the athletes dataset with regions dataset by NOC so that each athlete had a corresponding regions that they were from. These regions were country. We called this dataset "athletes2".

Combine olymweat2 and athletes and filter out indoor sports

```r
# Join athletes2 and olymweat2 dataset
olympics1 <- left_join(athletes2, olymweat2, by = c("Year" = "Year", "Season" = "Season")) %>%
  select(-City.y) %>%
  rename(City = City.x) %>%

# Filter out Olympics that do not follow modern format
  filter(!(Year %in% c(1896,1900,1904,1906,1908,1912,1920,1924,1928) & Season == "Summer")) %>%
  mutate(CityTemp1 = TempMon1,
         CityTemp2 = TempMon2) %>%

# Join inout dataset that contains information about indoor/outdoor sports
  left_join(inout, by = c("Sport" = "Sport")) %>%

# Filter out indoor sports
  filter(isindoor == 0) %>%

# Select out columns without further use
  select(-ID, -Sex, -Age, -Height, -Weight, -Team, -Event, -TempMon1, -TempMon2)
```

We then joined countrytemp and athletes2 by Year and Season, then joined it further with inout by Sport. The created dataset had each Year, Season, Host City, Sport, Medal, region (Country), Altitude of Host City, Starting Month, Ending Month, Starting Day, Ending Day, Number of Days, Host City Temperature of Month1 and Host City Temperature of Month2 and isIndoor for every single athlete that competed in the olympics in the past.We then filtered out athletes from older Olympics (1896,1900,1904,1906,1908,1912,1920,1924,1928) because they were hosted over longer periods (50+ days), which is different from more modern Olympics which are hosted over around 20 days. We also filtered out athletes competing in indoor sports since we thought indoor sports aren't affected by temperature or elevation difference. We called this dataset "olympics1".

Combine olympics1 with countrytemp

```r
olympics2 <- left_join(olympics1, countrytemp, by = c("region" = "name"))

# create vector of the starting and ending months of each olympic
startmon <- as.character(as.integer(olympics1$StartMonth))
endmon <- as.character(as.integer(olympics1$EndMonth))
```

```
# Loop through olympics2 to gather the temperature of months that corresponds to the olympic months for
# Store this data in histtemp1 and histtemp2
histtemp1 <- c()
histtemp2 <- c()
for (i in 1:nrow(olympics2)){
  histtemp1[i] <- olympics2[[i,startmon[i]]]
  histtemp2[i] <- olympics2[[i,endmon[i]]]
}

# Create new column corresponding to historic temperature of home countries in each Olympics
olympics2$histtemp1 <- histtemp1
olympics2$histtemp2 <- histtemp2

# Select out columns without further use
olympics2 <- olympics2 %>%
  select(-`1`,-`2`,-`3`,-`4`,-`5`,-`6`,-`7`,-`8`,-`9`,-`10`,-`11`,-`12`, -Annual_temp)
```

We next joined olympics1 with countrytemp. We then filtered out all the historic average monthly temperature of the athlete's home country if they did not correspond to the Olympic months. So for example, if an athlete competed in an Olympic held from August 23rd to September 7th, then the dataset would contain historic average temperature for August and September, in addition to all the information contained in olympics1. We called this dataset "olympics2".

Calculate difference between host city temperature and home country temperature

```
# Calculate the ratio of days in the first and second month
olympics3 <- olympics2 %>%
  mutate(NumDayMon1 = ifelse(as.integer(EndDay) > `Number of Days`,
                             `Number of Days`,
                             `Number of Days` - as.integer(EndDay)),
         NumDayMon2 = `Number of Days` - NumDayMon1,
         RatioMon1 = NumDayMon1/`Number of Days`,
         RatioMon2 = 1 - RatioMon1,
         CityTemp2 = ifelse(is.na(CityTemp2), 0, CityTemp2)) %>%

# Calculate the total difference
  mutate(tempdiff = (CityTemp1 - histtemp1)*RatioMon1 + (CityTemp2 - histtemp2)*RatioMon2) %>%

# Select out columns without further use
  select(-CityTemp1, -CityTemp2, -histtemp1, -histtemp2, -NumDayMon1,
         -NumDayMon2, -RatioMon1, -RatioMon2, -`Number of Days`, -StartMonth,
         -EndMonth, -StartDay, -EndDay, -isindoor)
```

We then used olympics2 to calculate the temperature difference between the host city and home country. If the Olympics were held in just one month, the calculation was simply subtracting the home country temperature from host city temperature. If the Olympics were held across two months, the calculation was a bit more complex. We first calculated the temperature difference for both months and added the weighted sum. The weights were calculated by the ratio of days in each months. We called this dataset "olympics3".

Combine elevation data with olympics3 and compute elevation difference

```
# Join olympics3 and elevation dataset
olympics4 <- olympics3 %>%
  left_join(elevation, by = c("region" = "Country")) %>%

# Calculate average elevation difference
  mutate(elevdiff = Altitude - Elevation) %>%

# Select out columns without further use
  select(-Elevation, -Altitude)
```

We next joined olympics3 and the elevation dataset and computed the elevation difference for each athletes. We called this dataset "olympics4".

Clean up medal data

```
# Quantify each medals as follows: Gold = 3, Silver = 2, Bronze = 1, None = 0
olympics5 <- olympics4 %>%
  mutate(Medal = ifelse(is.na(Medal), "0", Medal),
         Medal = ifelse(Medal == "Bronze", "1", Medal),
         Medal = ifelse(Medal == "Silver", "2", Medal),
         Medal = ifelse(Medal == "Gold", "3", Medal),
         Medal = as.integer(Medal))
```

We then quantified each medals using the following point system: NA = 0, Bronze = 1, Silver = 2 and Gold = 3. We used this point system because looking at sports literature, this was the most commonly used. For example "https://www.topendsports.com/events/summer/medal-tally/rankings.htm". We called the dataset after the quantification "olympics5".

Create dataset that contains total medals for each olympics

```
# Group by Year and Season and take the sum of Medal
totalmedals <- olympics5 %>%
  group_by(Year,Season) %>%
  summarise(TotMed = sum(Medal))
```

Using olympics5, we calculated the total medal points awarded in each Olympics. To do this, we first grouped by Year and Season, then took the sum of all medal points. An important point to note is that the medal points only include medals awarded in outdoor sports. We called this dataset "totalmedals".

Calculate percentage of medals won by each country

```
# Calculate the sum of all medals won by each country in each olympics
olympics6 <- olympics5 %>%
  group_by(Year, Season, region) %>%
  summarise(medalswon = sum(Medal)) %>%

# Calculate the proportion of total medals that each country won in each olympics
  left_join(totalmedals, by = c("Year" = "Year", "Season" = "Season")) %>%
  mutate(medratio = medalswon / TotMed)
```

Next, we grouped by Year, Season and Region in olympics5 and calculated the sum of medals. We leftjoined the totalmedals dataset onto the result and calculated the proportion of medals (medal ratio) won by each country in each Olympics. We called this dataset "olympics6". In essence, olympics6 has medal ratios for every country for every olympics.

Get rid of distinctions between each athletes and focus on country level performance

```r
# Get rid of all individual distinctions between athletes and focus on country
olympics7 <- olympics5 %>%
  select(Year, Season, region, tempdiff, elevdiff)

# There are multiple athletes from the same countries in every olympics, creating duplicates
# Find which rows has duplicates in olympics7
duplicated <- olympics7 %>%
  duplicated()

# Get rid of duplicates
olympics8 <- olympics7[!duplicated,]
```

Here, we return to olympics5. We created olympics7 by selecting only Region, Year, Season, Temperature Difference and Elevation Difference. We also created a dataset called "duplicate" since olympics7 has a lot of duplicate rows that needs to be taken out. We called the dataset created after taking out duplicate values olympics8. In essence, olympics8 had the temperature difference and elevation difference for all countries for every Olympics.

Combined join medalratio data onto temperature/elevation data

```r
# Join dataset that contains medalratio info onto olympics8
olympics9 <- olympics6 %>%
  left_join(olympics8, by = c("Year" = "Year", "Season" = "Season", "region" = "region")) %>%

# Select out columns that does not have further use
  select(-medalswon, -TotMed)
```

After leftjoining olympics8 onto olympics6, we were able to create the final dataset, olympics9. Using this data, we performed out analysis.

Analysis

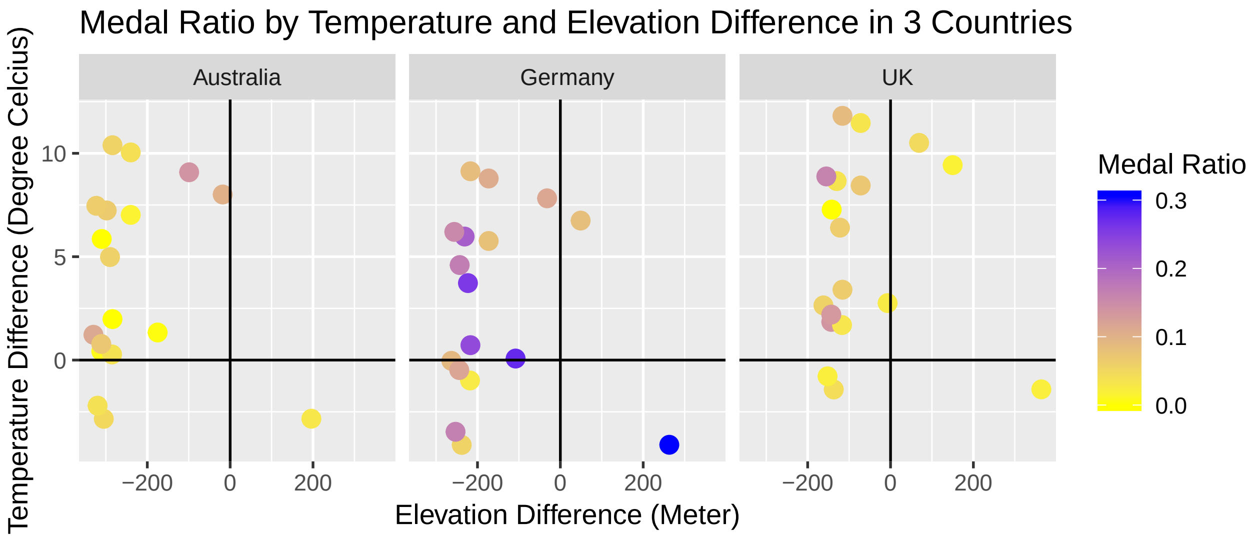
Plots exploring effects of elevation/temperature difference on medal ratios

```r
# Create dataset for graphical exploration
olympicsds1 <- olympics9 %>%

# Studied Australia, Germany, UK as they are consistent performers
  filter(region %in% c("Australia", "Germany", "UK")) %>%
# Got rid of winter olympics and rows with missing data
  filter(!is.na(elevdiff), !is.na(tempdiff), Season == "Summer")


olympicsds1 %>%
#Filtered out extreme outliers
  filter(!(elevdiff > 1500)) %>%

# Generate plot
  ggplot(aes(x = elevdiff, y = tempdiff)) + geom_point(size = 3.0) + aes(color = medratio)  +
    facet_wrap(~region, ncol = 4) + theme(legend.position = "right") + labs(title = "") +
    scale_color_gradient(low = "yellow", high = "blue") +
    geom_vline(xintercept = 0) + geom_hline(yintercept = 0) +
    labs(x = "Elevation Difference (Meter)",
         y = "Temperature Difference (Degree Celcius)",
         title = "Medal Ratio by Temperature and Elevation Difference in 3 Countries",
         color = "Medal Ratio")
```



We generated this plot to see whether there were any patterns that could be detected visually. This graphic plots Elevation Difference and Temperature Difference onto an x-y plane and the color of the points represents medal ratio. We picked Australia, Germany and UK because they are consisting high performers in the Olympics and because they have smaller landmass than other high performers(US,Russia,China). Looking at the graph, we do not see any noticable relationships between Elevation Difference, Temperature Difference and Medal Ratio. At this point, we started to suspect that there are no strong connections between a country's olympic performance and the temperature/elevation difference.

Results

Create a linear mixed model

```
# Only keep rows that are summer and is not missing temperature and elevation difference
formodel1 <- olympics9 %>%
  filter(Season == "Summer", !is.na(tempdiff), !is.na(elevdiff))

# Create linear mixed model with clustering as country/region
```

```r
model <- lmer(medratio ~ tempdiff + elevdiff + (1|region), data = formodel1)
```

In our linear mixed model, we found there to be no explanatory power from neither temperature difference nor elevation difference. We came to this conclusion because of temperature difference's and elevation difference's t-value. The R package we use (lme4) lacks P-value computations and uses t-value and the t distribution to gauge significance. We know that the further the t-value is to zero, the smaller the p-value. Luckily, for our model summary, both temperature difference and elevation difference had t-values very close to zero, allowing us to deduce that these two factors had no explanatory power, like we hypothesized.

```r
# Output model summary
summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: medratio ~ tempdiff + elevdiff + (1 | region)
##    Data: formodel1
##
## REML criterion at convergence: -5548
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8440 -0.2434 -0.0723 -0.0128  8.0059
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  region   (Intercept) 0.0004102 0.02025
##  Residual             0.0005308 0.02304
## Number of obs: 1243, groups:  region, 132
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 1.011e-02  2.054e-03   4.922
## tempdiff    1.135e-04  1.448e-04   0.784
## elevdiff    6.862e-07  1.282e-06   0.535
##
## Correlation of Fixed Effects:
##          (Intr) tmpdff
## tempdiff 0.122
## elevdiff 0.278  0.224
```

LMM mainly uses t-values instead of p-values. Because of this caveat, our assessment of the model is not wholly accurate. Fortunately, the model with both temperature difference and elevation difference had respective t-values of 0.784 and 0.535 which are both very close to zero. Through this, we deduce that neither elevation difference nor temperature difference has no significant effect on q country's medal ratio.

While keeping in mind that neither predictor is significant, we observe the coefficients' signs are interesting because both are positive, indicating that with an increase in elevation difference or temperature difference, there is a boost in medal ratio.

Temperature difference's coefficient is 1.135e-4, which can be interpreted as: with an one degree celsius increase, a country's medal ratio will increase by .000135. For elevation difference, its coefficient of 6.862e-7 can be interpreted as: with a one meter increase in elevation, a country's medal ratio increases by 6.862e-7. However, it is very important to note that these interpretations are severely limited because neither factor are significant in the model. These interpretations are viable only if the predictors were significant.

Conclusion

Ultimately, from our current model, we do not detect temperature difference or elevation difference to have an effect on a country's medal performance, disproving our initial hypothesis and answering our research question. It is important to note that this study does not imply temperature difference and elevation difference have no effect on a country's performance. It merely states we did not DETECT any significant difference.

We started this investigation to study whether temperature and elevation difference between an Olympic host city and a country has any effect on the performance of each country. Our linear mixed model suggests that temperature difference and elevation difference do not have significant relationships with a country's performance in the olympics. However, there are several limitations to this model.

The main limitation to our model is the lack of P-values. Without p-values, it is harder to interpret the model and determine the validity of a model's meaning. However, instead of relying solely on p-values, we used other statistical measures to evaluate the model such as t-values and coefficients' signs. While these are far from "best practice", the lack of p-values reminded us of the industry's overreliance on p-values and the flaws p-values have as well (https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.WBCslMmJQ20). A way to navigate around this is to contact different authors of R packages related to LMMs and see their thoughts on testing the validity of LMMs. In improving our model, close work with one of the R package authors would be a key future step.

Also, the fact that random effects are on a zero-inflated distribution may have created bias in our model. A zero-inflated distribution means that a poisson or exponential distribution would have been better than the usually-assumed normal distribution for the random effect.

Furthermore, there are multiple packages that support LMM. We decided to user lme4 because it was used most frequently and there were more documentation related to the package. For the future, we could do further research on different R packages that uses LMM and find the pros/cons of each one to see which fits our topic the best.

For our data, one limitation was that we could not collect many of the host city temperature data for the winter olympics, mostly due to the fact that winter olympics are held in remote areas. This limited the range of analysis to the summer olympics. This was unfortunate since we were interested in comparing the two olympics.

Another limitation was that we were not able to collect data related to the range of temperature and elevation. In geographically large countries like the United States and China, there is a large range in elevation and temperature. To accurately depict larger countries, using values such as min/max or IQR would have been ideal, but there were no sources that collected such data many years into the past. Finding the correct metric to capture the variations in elevation and temperature would be a promising effort for the future. Additionally, instead of just temperature and elevation, adding other weather factors such as humidity, air pressure, and precipitation could be a useful endeavor.