CUNY MSDA 606 Spring 2016 Project Proposal

James Topor

Research question

The earth's climate is affected by a variety of phenomena, many of which are external to the earth's atmosphere. These include the wobble of the earth's axis, gamma rays that continuously bombard the planet's atmosphere from outer space, the rate at which the earth's oceans absorb and release heat, natural variations in the earth's orbit around the sun, and variations in the sun's irradiance resulting from the sun's own internal stellar dynamics.

For example, scientists have long speculated on the possibility of a relationship between sunspot activity and variations in the earth's atmospheric temperatures and precipitation. Fortunately, regular periodic observations of sunspot counts, atmospheric temperatures, and precipitation for the past 120 years or so are readily available to us. As such, we can conduct our own analysis into the relationship between sunspots and some of the earth's climatic variations.

Specifically, we will examine the following:

1) Whether the average number of sunspots observed within a given calendar year might be predictive of the average atmospheric temperature observed within the continental United States for that same calendar year;
2) Whether the average number of sunspots observed within a given calendar year might be predictive of the average precipitation amount observed within the continental United States for that same calendar year.

To facilitate this analysis we will be drawing on data from two separate sources: one of which records sunspot observations on a regular basis; and another which tracks atmospheric temperature and precipitation observations for the continental United States. Both are described in greater detail in the sections labeled "**Data Collection**" and "**Data Source**".

Cases

Each case is comprised of three distinct observations recorded for each calendar year for the years 1900 - 2015. The three observations are:

1) The average number of sunspots observed daily as recorded by The Royal Observatory of Belgium.
2) The average of recorded atmospheric temperatures throughout the continental United States.
3) The average of recorded precipitation amount throughout the continental United States.

We have a total of 116 cases: one for each calendar year for the years 1900 - 2015.

Data collection

Sunspot Data

Scientists at the Royal Observatory of Belgium have recorded daily sunspot observations since January, 1818. Access to the data repository containing those observations is available via the internet at the following web page:

http://www.sidc.be/silso/datafiles

The specific sunspot data set used for this analysis can be obtained from that web page by scrolling down to the item labeled **"Yearly mean total sunspot number [1700 - now]"** and selecting the **CSV** option, which generates a .csv file containing the mean number of daily sunspot observations for each calendar year for the period 1700 - 2015.

The .csv file contains 5 columns of data described as follows (source: http://www.sidc.be/silso/infosnytot):

- **Column 1**: Gregorian calendar year (mid-year date)
- **Column 2**: Yearly mean total sunspot number.
- **Column 3**: Yearly mean standard deviation of the input sunspot numbers from individual stations.
- **Column 4**: Number of observations used to compute the yearly mean total sunspot number.
- **Column 5**: Definitive/provisional marker. '1' indicates that the value is definitive. '0' indicates that the value is still provisional.

From this .csv file we will utilize data for the years 1900 - 2015 as part of our analysis.

Atmospheric Temperature and Precipitation Data

The National Oceanic and Atmospheric Administration maintains a repository of climate-related observations that is accessible via the internet. Individual temperature and precipitation measurements are recorded at 1,218 different locations throughout the continental United States on a daily basis. The NOAA aggregates those measurements and computes average temperature and precipitation amounts for the continental United States and updates its online repository of that data on a regular basis.

Average temperature observations for the continental United States for the period 1900 - 2015 can be obtained via the following web page:

http://www.ncdc.noaa.gov/cag/time-series/us/110/0/tavg/12/12/1900-2015

The specific data set to be used in this analysis can be obtained as follows from that web page:

- For **"Parameter"** select "Average Temperature" from the drop down list;
- For **"Time Scale"** select "12-Month";
- For **"Month"** select "December";
- For **"Start Year"** select "1900";

- For **“End Year”** select “2015”
- For **“State(Region)”** select “Contiguous US”
- For **“Climate Division/City”** select “All 48 States”

Then, click the “**Plot**” button. Once plotting has been completed, the data set is displayed on that page and can be downloaded as a .csv by scrolling to the “**Download**” label and clicking the MS Excel icon. The .csv file will have a header that explains the two columns of data therein.

Precipitation observations for the continental United States for the period 1900 - 2015 can be obtained via the following web page:

http://www.ncdc.noaa.gov/cag/time-series/us/110/0/pcp/12/12/1900-2015

The specific data set to be used in this analysis can be obtained as follows from that web page:

- For **“Parameter”** select “Precipitation” from the drop down list;
- For **“Time Scale”** select “12-Month”;
- For **“Month”** select “December”;
- For **“Start Year”** select “1900”;
- For **“End Year”** select “2015”
- For **“State(Region)”** select “Contiguous US”
- For **“Climate Division/City”** select “All 48 States”

Then, click the “**Plot**” button. Once plotting has been completed, the data set is displayed on that page and can be downloaded as a .csv by scrolling to the “**Download**” label and clicking the MS Excel icon. The .csv file will have a header that explains the two columns of data therein.

Type of study

Since we are relying upon data recorded each year for the period 1900 - 2015, this is strictly a retrospective observational study.

Data Sources

Sunspot Data
As noted above, the sunspot data is provided by The Royal Observatory of Belgium. Their website is:

http://www.sidc.be/silso/home

The proper citation for the Royal Observatory of Belgium and the data contained therein is:

Source: WDC-SILSO, Royal Observatory of Belgium, Brussels.

Atmospheric Temperature and Precipitation Data

As noted above, atmospheric temperature and precipitation data have been obtained via the National Oceanic and Atmospheric Adminstration (NOAA), which is a branch of the United States Department of Commerce. Their website is:

http://www.ncdc.noaa.gov/

The temperature and precipitation data used in this analysis are accessible via this specific web page:

http://www.ncdc.noaa.gov/cag/

Response Variables

For purposes of this analysis we will be examing two separate response variables:

1) **Average Temperature**: A continuous numeric variable whose value represents an average of recorded atmospheric temperatures in degrees Farenheit throughout the continental United States for a given calendar year for the years 1900 - 2015.
2) **Average Precipitation**: A continuous numeric variable whose value represents an average of recorded precipitation measured in inches throughout the continental United States for a given calendar year for the years 1900 - 2015. All forms of atmospheric precipitation, including rain, snow, sleet, and hail, are encompassed by this variable. (*NOTE: The NOAA converts all non-rain precipitation measures to a rain-equivalent metric, thereby ensuring a standard unit of measurement for all types of precipitation*).

Explanatory Variable

For purposes of this analysis we will utilize a single explanatory variable:

- **Average Sunspot Count**: A continuous numeric variable representing the mean daily number of sunspots observed within a given calendar year for the years 1900 - 2015 as recorded by the Royal Observatory of Belgium.

Relevant summary statistics

Sunspot Data

Prior to computing summary statistics we need to refine the raw sunspot data so that it conforms with our intended analytical approach.

Step 1: Read the .csv file containing sunspot data into a dataframe. (*NOTE: The .csv file has no header info*):

```
# Read raw sunspot data from semicolon delimited CSV file
sscsv <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-606/master/Project/ss_obs.csv",
                  sep=";", header = FALSE, stringsAsFactors = FALSE)

str(sscsv)
```

```
## 'data.frame':    316 obs. of  5 variables:
##  $ V1: num  1700 1702 1702 1704 1704 ...
##  $ V2: num  8.3 18.3 26.7 38.3 60 96.7 48.3 33.3 16.7 13.3 ...
##  $ V3: num  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ V4: int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ V5: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(sscsv)
```

```
##       V1   V2 V3 V4 V5
## 1 1700.5  8.3 -1 -1  1
## 2 1701.5 18.3 -1 -1  1
## 3 1702.5 26.7 -1 -1  1
## 4 1703.5 38.3 -1 -1  1
## 5 1704.5 60.0 -1 -1  1
## 6 1705.5 96.7 -1 -1  1
```

Step 2: Create a data frame using the 2 specific columns containing the calendar year and the correspondig sunspot observation metric, trimming whitespaces where needed:

```
ssdf <- data.frame(sscsv$V1, sscsv$V2)

head(ssdf)
```

```
##   sscsv.V1 sscsv.V2
## 1   1700.5      8.3
## 2   1701.5     18.3
## 3   1702.5     26.7
## 4   1703.5     38.3
## 5   1704.5     60.0
## 6   1705.5     96.7
```

Step 3: Rename columns to match full attribute names

```
names(ssdf)
```

```
## [1] "sscsv.V1" "sscsv.V2"
```

```
names(ssdf)<-c("Year","AvgSpots")

row.names(ssdf)<-NULL

names(ssdf)
```

```
## [1] "Year"     "AvgSpots"
```

Step 4: Substract 0.5 from each year to get a precise calendar year value

```
ssdf$Year <- ssdf$Year - 0.5
head(ssdf$Year)
```

```
## [1] 1700 1701 1702 1703 1704 1705
```

Step 5: Discard data for years prior to 1900

```
ssdf <- subset(ssdf, Year >= 1900)
row.names(ssdf)<-NULL
head(ssdf)
```

```
##   Year AvgSpots
## 1 1900     15.7
## 2 1901      4.6
## 3 1902      8.5
## 4 1903     40.8
## 5 1904     70.1
## 6 1905    105.5
```

Now that the sunspot data has been refined to include data for only the years 1900 - 2015, we can compute some summary metrics:

```
nrow(ssdf)
```

```
## [1] 116
```

As we can see from the output or R's **nrow** function, the data frame we've obtained via our sunspot data refinement efforts has the 116 cases we described earlier.

Let's take a look at the output of R's **summary** and **sd** functions for our sunspot data:

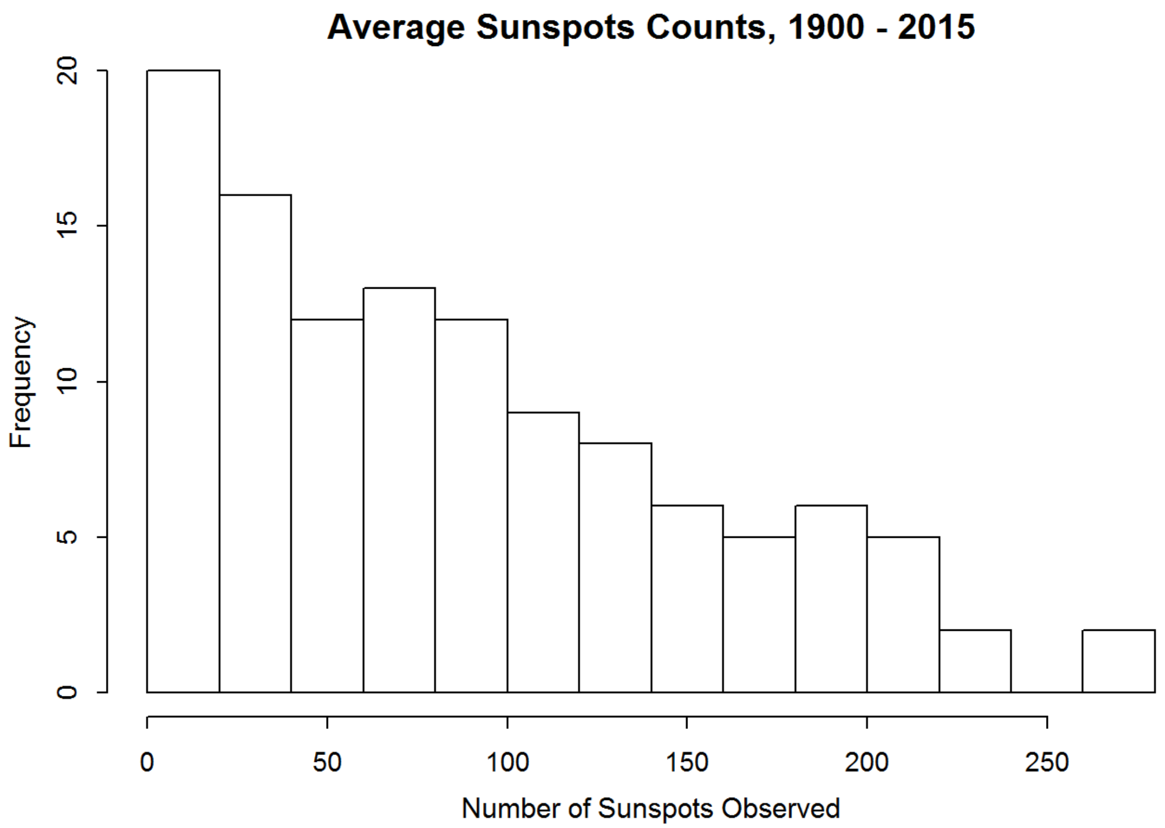
```
summary(ssdf$AvgSpots)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.40   27.70   76.25   88.24  132.90  269.30
```

```
sd(ssdf$AvgSpots)
```

```
## [1] 66.94521
```

The mean number of sunspots observed each year for the period 1900 - 2015 is 88.24, the median is 76.25, and the standard deviation is 66.945. Since the mean is so much larger than the median, it appears that we have a right-skewed distribution. A plot of a histogram of the individual values confirms this:



Temperature Data

Our refinement of the pertinent temperature data is as follows:

Step 1: Read the .csv file containing temperature data into a dataframe. (*NOTE: The .csv file does have header info on its third line*):

```
# Read temperature data from CSV file. Skip the first 2 lines since header is on third line
tempdf <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-606/master/Project/tempUS.txt",
                   header = TRUE, skip=2, stringsAsFactors = FALSE)

str(tempdf)
```

```
## 'data.frame':    116 obs. of  2 variables:
##  $ Date : int  190012 190112 190212 190312 190412 190512 190612 190712 190812 190912 ...
##  $ Value: num  52.8 51.9 51.6 50.6 51.2 ...
```

```
head(tempdf)
```

```
##     Date Value
## 1 190012 52.77
## 2 190112 51.87
## 3 190212 51.59
## 4 190312 50.62
## 5 190412 51.16
## 6 190512 51.00
```

Step 2: Truncate last 2 digits of all 'Date' values since they aren't needed for our analysis

```
# convert 'year' values to string
truncYear <- as.character(tempdf$Date)
head(truncYear)
```

```
## [1] "190012" "190112" "190212" "190312" "190412" "190512"
```

```
# truncate last 2 characters
truncYear <- substr(truncYear, 1, nchar(truncYear)-2)
head(truncYear)
```

```
## [1] "1900" "1901" "1902" "1903" "1904" "1905"
```

```
# convert the 'year' strings back to numeric format
tempdf$Date <- as.numeric(truncYear)

head(tempdf)
```

```
##   Date Value
## 1 1900 52.77
## 2 1901 51.87
## 3 1902 51.59
## 4 1903 50.62
## 5 1904 51.16
## 6 1905 51.00
```

Now that the temperature data has been refined we can compute some summary metrics:

```
nrow(tempdf)
```

```
## [1] 116
```

As we can see from the output or R's **nrow** function, the data frame we've obtained via our temperature data refinement efforts has the 116 cases we described earlier.

Let's take a look at the output of R's **summary** and **sd** functions for our sunspot data:

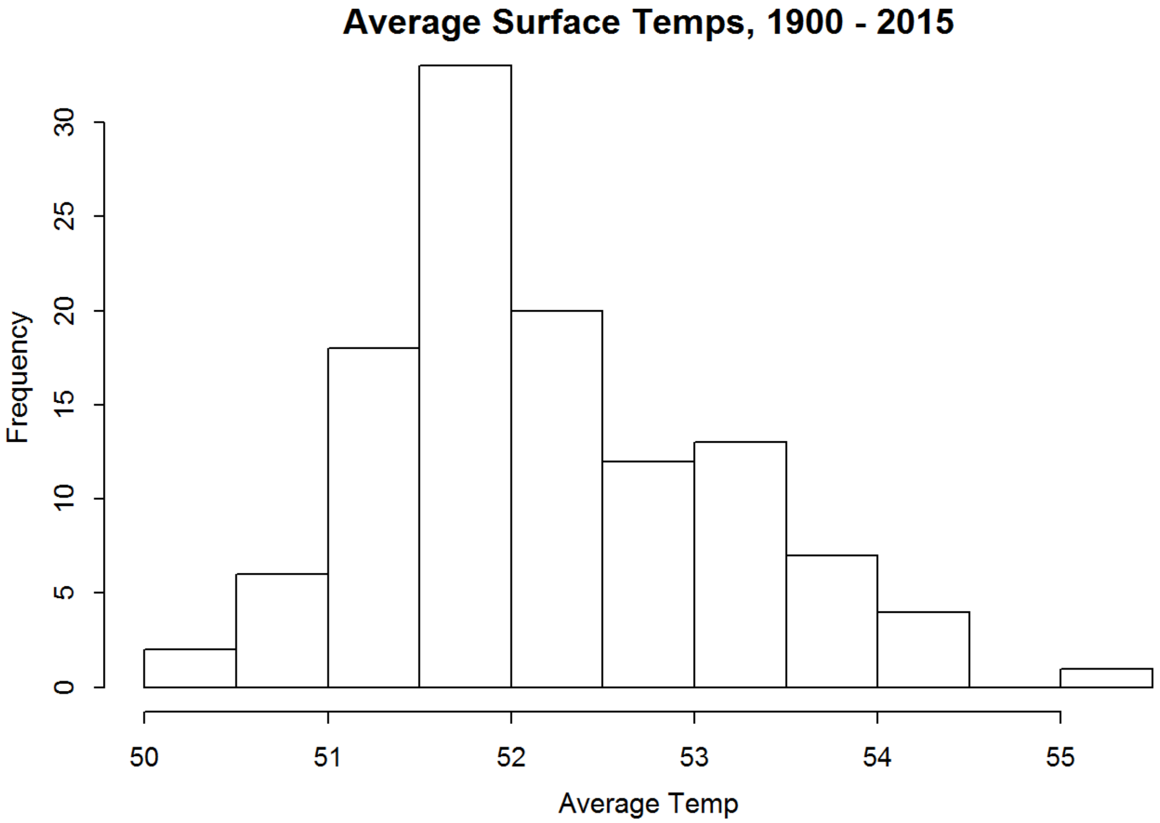
```
summary(tempdf$Value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.06   51.58   51.98   52.20   52.80   55.28
```

```
sd(tempdf$Value)
```

```
## [1] 0.9327599
```

The temperature readings for the continental United States for the years 1900-2015 have a mean of 52.2, a median of 51.98, and a standard deviation of 0.9327. The mean being greater than the median suggests that we once again may be dealing with a skewed distribution. A plot of a histogram for the temperature data reveals a right-skew characteristic:



Precipitation Data

Our refinement of the pertinent precipitation data is as follows:

Step 1: Read the .csv file containing precipitation data into a dataframe. (*NOTE: The .csv file does have header info on its third line*):

```
# Read precipitation data from CSV file. Skip the first 2 lines since header is on third line
precipdf <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-606/master/Project/prcpUS.txt"
                     header = TRUE, skip=2, stringsAsFactors = FALSE)

str(precipdf)
```

```
## 'data.frame':    116 obs. of  2 variables:
##  $ Date : int  190012 190112 190212 190312 190412 190512 190612 190712 190812 190912 ...
##  $ Value: num  30.6 27.6 30.6 29.4 27.9 ...
```

```
head(precipdf)
```

```
##     Date Value
## 1 190012 30.63
## 2 190112 27.63
## 3 190212 30.63
## 4 190312 29.36
## 5 190412 27.95
## 6 190512 32.60
```

Step 2: Truncate last 2 digits of all 'Date' values since they aren't needed for our analysis

```
# convert 'year' values to string
truncYear <- as.character(precipdf$Date)
head(truncYear)
```

```
## [1] "190012" "190112" "190212" "190312" "190412" "190512"
```

```
# truncate last 2 characters
truncYear <- substr(truncYear, 1, nchar(truncYear)-2)
head(truncYear)
```

```
## [1] "1900" "1901" "1902" "1903" "1904" "1905"
```

```
# convert 'year' strings back to numeric format
precipdf$Date <- as.numeric(truncYear)

head(precipdf)
```

```
##   Date Value
## 1 1900 30.63
## 2 1901 27.63
## 3 1902 30.63
## 4 1903 29.36
## 5 1904 27.95
## 6 1905 32.60
```

Now that the precipitation data has been refined we can compute some summary metrics:

```
nrow(precipdf)
```

```
## [1] 116
```

As we can see from the output or R's **nrow** function, the data frame we've obtained via our precipitation data refinement efforts has the 116 cases we described earlier.

Let's take a look at the output of R's **summary** and **sd** functions for our sunspot data:

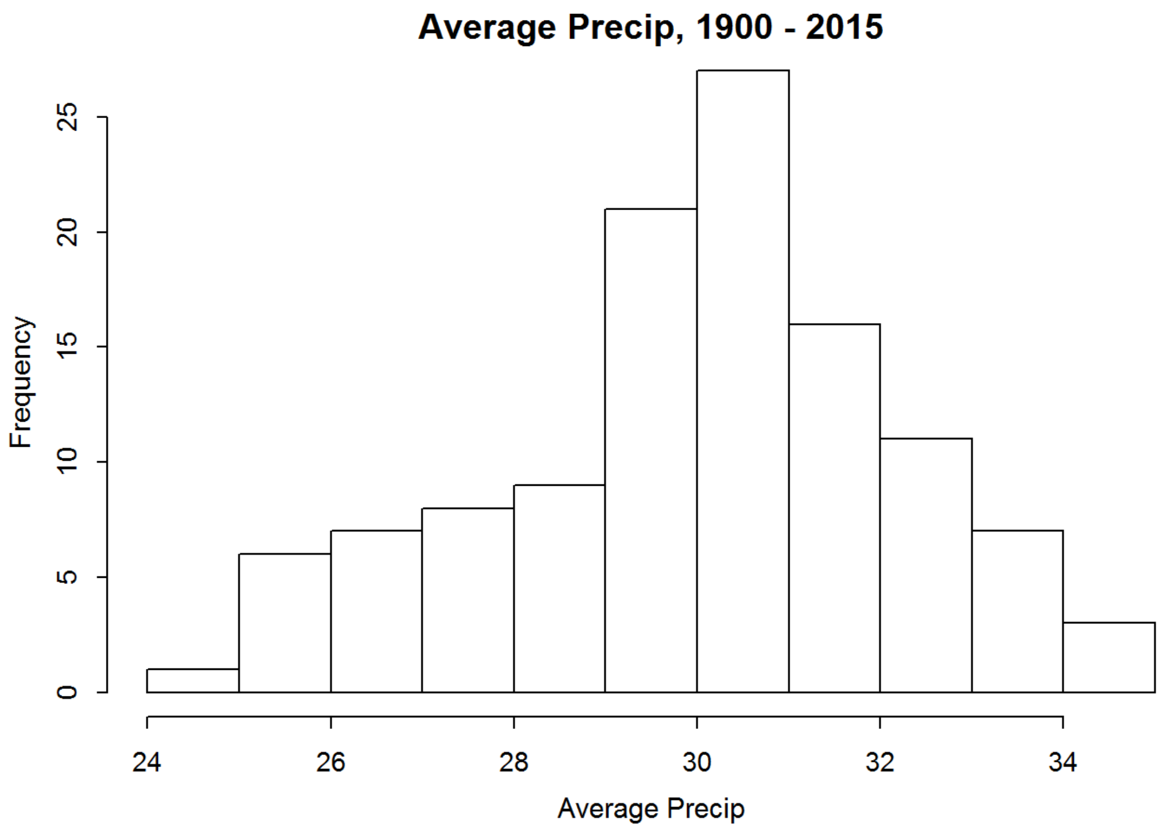
```
summary(precipdf$Value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.91   28.78   30.28   30.04   31.31   34.96
```

```
sd(precipdf$Value)
```

```
## [1] 2.22302
```

The precipitation data for the continental United States for the years 1900-2015 have a mean of 30.04, a median of 30.28, and a standard deviation of 2.22. The mean being nearly identical to the median suggests that we may have a nearly normal distribution of average precipitation amounts. A plot of a histogram for the precipitation data reveals a slight amount of left skew:


Average Precip, 1900 - 2015
Frequency
Average Precip