# Drought in Rhode Island

## Environmental Data Science

Zhaosen Guo

11/20/2020

## Introduction

Rhode Island is a state in the United States situated on the north-east coast. It is in New England and known for sandy shores and seaside Colonial towns. It would probably come to most people around the nation as a surprise, that despite of its coastal location, Rhode Island has been experiencing a rather harsh drought in the past few months. This project aims to investigate the shortage of water in this state, which is not commonly covered by national news media, and explore factors such as time, place, and residents affected.

---

## Problem Statement

The ultimate question that I seek to answer is: Looking at relevant data in recent years, has the drought in Rhode Island been "expected" in terms of timing and magnitude?

## Data and Methodology

### Original Data

The data used in this project is retrieved from *The U.S. Drought Monitor* hosted by The National Drought Mitigation Center at University of Nebraska-Lincoln. The data are weekly collections of categorical scales of drought values, based on percentage of areas and populations affected. A sample of the selected columns of the first 5 rows from the raw data can be seen here:

| StateAbbreviation | None | D0 | D1 | D2 | D3 | D4 | ValidStart | ValidEnd |
|---|---|---|---|---|---|---|---|---|
| RI | 0 | 0 | 55.86 | 44.14 | 0.00 | 0 | 2020-11-17 | 2020-11-23 |
| RI | 0 | 0 | 55.86 | 44.14 | 0.00 | 0 | 2020-11-10 | 2020-11-16 |
| RI | 0 | 0 | 55.86 | 44.14 | 0.00 | 0 | 2020-11-03 | 2020-11-09 |
| RI | 0 | 0 | 0.00 | 55.61 | 44.39 | 0 | 2020-10-27 | 2020-11-02 |
| RI | 0 | 0 | 0.00 | 55.61 | 44.39 | 0 | 2020-10-20 | 2020-10-26 |
| RI | 0 | 0 | 0.00 | 0.98 | 99.02 | 0 | 2020-10-13 | 2020-10-19 |

It shows that the percent of areas under different drought categories -
**None** for normal, **D0** for abnormally dry, **D1** for moderate drought, **D2** for severe drought, **D3** for extreme drought, and **D4** for exceptional drought.
For more detailed classification parameters, please refer to *this webpage*.
One note: each level of drought automatically covers the previous one, for example, for the week of 2020-11-17 in RI, although the numerical values is 0 under **D0**, it is assumed that 100% of the area there has to be "abnormally dry" before being considered "moderate drought" or "severe drought" in **D1** and **D2**. Similarly, although only 55.86% is under **D1**, in fact the rest of 44.14% under **D2** has to be qualified in **D1** before getting a more severe rating.

In addition to the comprehensive data shown, some data will be introduced to visualize the drought on a map.

### Data Cleaning and Merging

The first step to make a data set usable is to clean it, and here I will be removing irrelevant columns in both the population and area data sets, such as $StateAbbreviation. After that, for the easy of accessing the data, I will merge the two data set based on respective weeks. The categories are simply labeled as "Dx" in both sets, so to avoid confusion, I will rename columns prior to the merging.

### Research Methods

First, I will investigate the duration of drought in each year and their patterns, focusing on the overall week counts. Then, the severity of the drought throughout the years will be checked under statistically analysis on proportion of the areas under different level of drought in each season. Moreover, I will introduce an index that weights all the factors can combine 5 categories into one single representative number. Lastly, the population and areas will be compared and contrasted to see if the drought has changed its patterns, and it will be aided with some map images.

---

## Analysis

### Yearly Data

The yearly average of percentage under each drought category can be derived using group_by & summarise_at from the dplyr package, and the result is:

| Year_Group | Area_None | Area_D0 | Area_D1 | Area_D2 | Area_D3 | Area_D4 |
|---|---|---|---|---|---|---|
| 1 | 50.18404 | 30.398077 | 19.418077 | 0.000000 | 0.0000000 | 0 |
| 2 | 36.44904 | 29.380577 | 21.117308 | 13.051538 | 0.0011538 | 0 |
| 3 | 48.49154 | 15.160577 | 36.347885 | 0.000000 | 0.0000000 | 0 |
| 4 | 88.30904 | 11.690962 | 0.000000 | 0.000000 | 0.0000000 | 0 |
| 5 | 93.24654 | 6.753462 | 0.000000 | 0.000000 | 0.0000000 | 0 |
| 6 | 61.64462 | 11.742885 | 6.917115 | 8.929423 | 10.7659615 | 0 |

In terms of areas under each categories, it is clear that there are some discrepancies among the average area proportions. It is also interesting to see that for the past 6 years Rhode Island has never gotten to a category D4 drought on its land. Also, in terms of the $Year_Group variable, "1" means the year cycle closest to the

date of data retrieval (Nov. 18 2020-2019).

Now lets look at the populations and check on the average percent affected by drought:

| Year_Group | Pop_None | Pop_D0 | Pop_D1 | Pop_D2 | Pop_D3 | Pop_D4 |
|---|---|---|---|---|---|---|
| 1 | 52.39635 | 31.506923 | 16.095577 | 0.00000 | 0.0000000 | 0 |
| 2 | 34.35385 | 29.531538 | 22.819039 | 13.29346 | 0.0011538 | 0 |
| 3 | 49.99038 | 13.046731 | 36.962885 | 0.00000 | 0.0000000 | 0 |
| 4 | 89.52154 | 10.478461 | 0.000000 | 0.00000 | 0.0000000 | 0 |
| 5 | 94.38731 | 5.612692 | 0.000000 | 0.00000 | 0.0000000 | 0 |
| 6 | 61.16135 | 12.360000 | 6.239039 | 10.21596 | 10.0238462 | 0 |

And now, to answer the question of if the proportions of each drought categories have been consistent throughout the years, a MANOVA test is used. Much like ANOVA, MANOVA simply adds a "multivariate" to the front and the null hypothesis is that the means on multiple dependent variables (in this case the percent land/population under each categories) are equal across groups; the alternative hypothesis is that the variables are not the same, hence the patterns in the recent years are different.

```
## [1] "Manova on Area"

##             Df  Pillai approx F num Df den Df    Pr(>F)
## Year_Group   1 0.17772   13.227      5    306 1.125e-11 ***
## Residuals  310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "------------------------------------------------------------"

## [1] "Manova on population"

##             Df Pillai approx F num Df den Df    Pr(>F)
## Year_Group   1 0.1815   13.571      5    306 5.722e-12 ***
## Residuals  310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

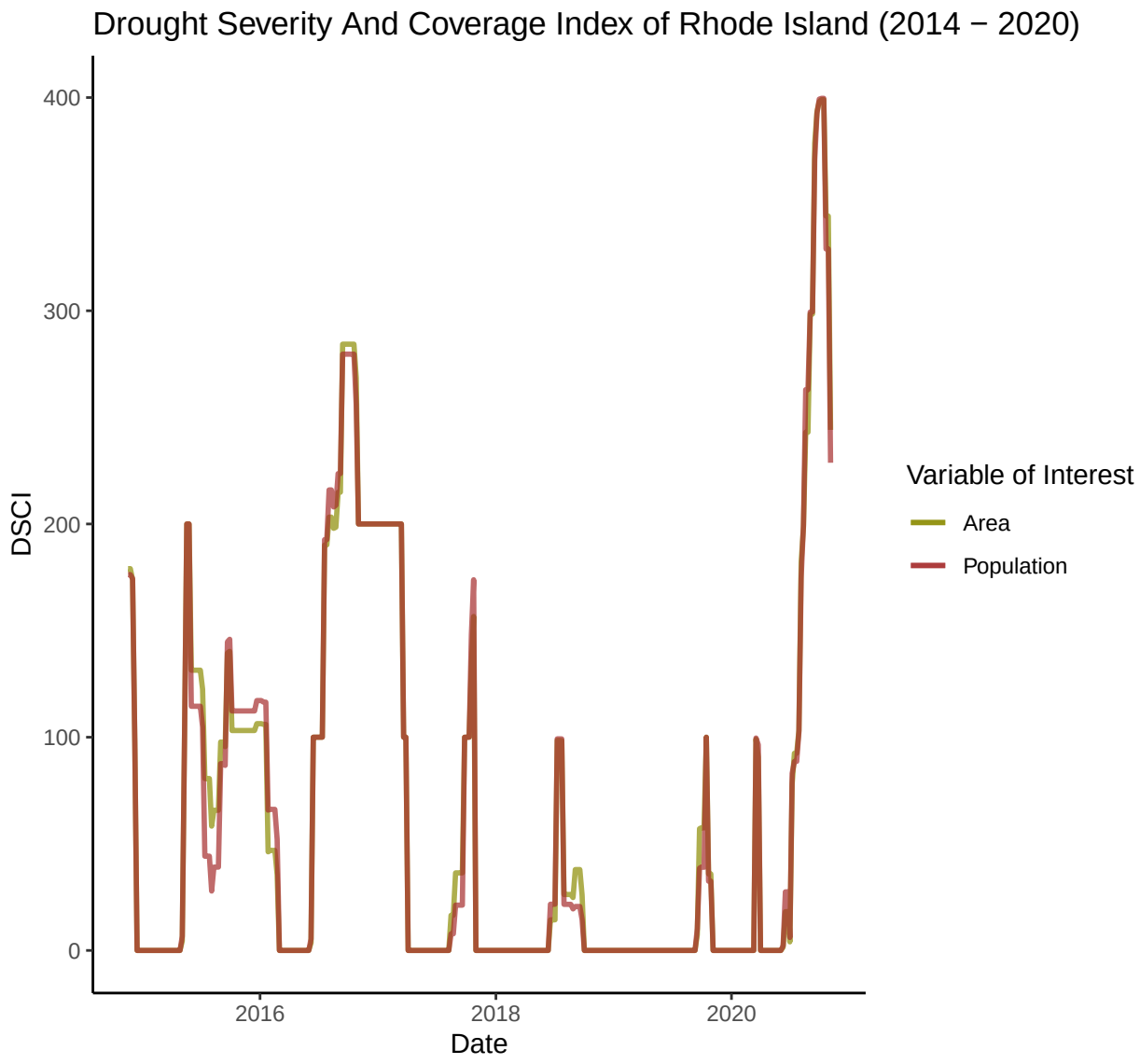
From the results of two MANOVA tests above, it is clear to observe that both tests yielded an extremely low P-value, meaning the null hypothesis should be rejected and that the proportions of drought categories change from year to year. Moreover, the **Pillai** in the summary output is short for "Pillai-Bartlett Trace." It is used as a test statistic in MANOVA, a positive valued statistic ranging from 0 to 1. Increasing values means that effects are contributing more to the model, and from the result both Pillai's trace seems to have pretty high values.

## Drought Severity And Coverage Index

The DSCI is a proposed way to convert the drought levels to a single value that provides a more comprehensive representation of the situation in a specific location. This equation here explains how the weighted sum is calculated:

$$\text{DSCI} = 0 \cdot \text{None} + 1 \cdot D_0 + 2 \cdot D_1 + 3 \cdot D_2 + 4 \cdot D_3 + 5 \cdot D_4$$

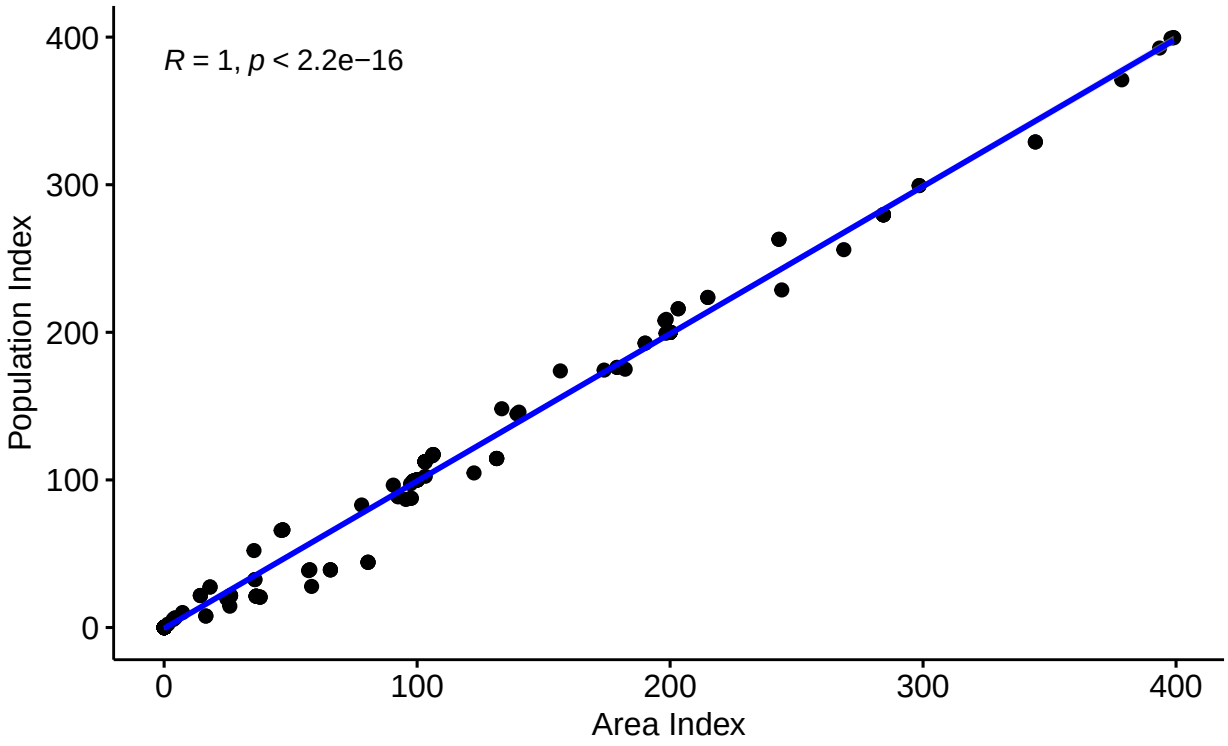
So after calculating the appropriate DSCI, we can visualize it:




And just like the results we obtained from the previous tests, the overall DSCI index has not shown any visible patterns; in additions, the recent spike of drought seems to be easing off in some degree, but it is still higher than normal.

## Population vs. Area

The graph above provides a good representation of the drought severity, however, it also highlighted the differences between two metrics that were used in the data set. The percent area effected and the percent population effected some times have discrepancies. With a state of 1,212 $mi^2$ and 1.059 million residents, that different could be huge in situations like the trends shown from 2015 to 2016, as well as in 2018.
Here is an overview of DSCI(Area) plotted against DSCI(Population):



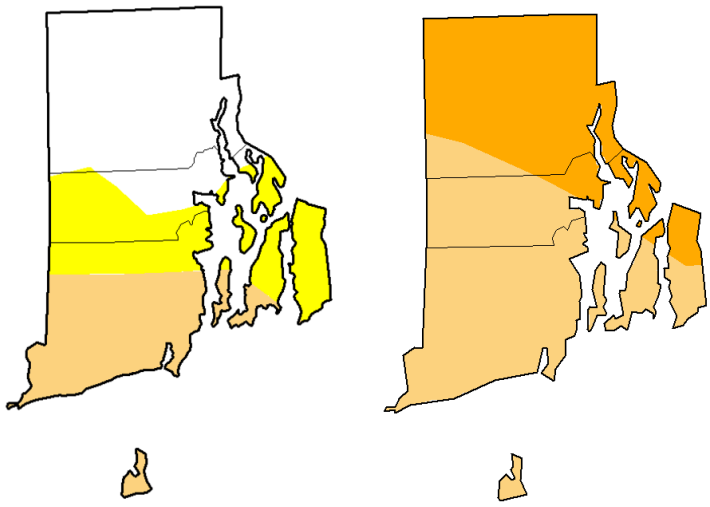

The plot actually shows a pretty solid linear correlations, however, the outliers are still worthy of inspection.

To do so, I first calculated the absolute proportional differences between **DSCI_Area** and **DSCI_Pop**

$$\frac{|\text{DSCI}_{Area} - \text{DSCI}_{population}|}{\text{DSCI}_{Area} + \text{DSCI}_{population}}$$

, and look at dates that are interesting.

Here are two examples of weeks where Index values differ a lot between population and area.
Left: 2015-07-28, when DSCI_Area = 2 * DSCI_Pop,
Right: 2020-08-25, when DSCI_Pop = 20 + DSCI_Area



When the drought is more severe in the south of RI, because there's not much population, the DSCI Area index would be higher; on the other hand, when the drought is more concentrated in the North, including Providence, more population is affected than the relative percentage of the land.

---

## Conclusion

The drought in Rhode Island has been proven historical, and even though in common conception usually fall is a season that lacks natural rainfall and other means of water, the magnitude of this year's drought, in comparison to previous 5 years, is not comparable and predictable. In addition, further investigations on drought levels should be reminded to use a combination of residents affected as well as land area affected. As shown previously, only considering one factor may misrepresent the situation.

# Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
#setup the environment
library(tidyverse)
library(ggplot2)
library(ggpubr)
library(png)
data.area <- read_csv(
  "https://raw.githubusercontent.com/guozhaosengzs/ENVDS/master/final_project/area_p_6yrs.csv")
data.pop <- read_csv(
  "https://raw.githubusercontent.com/guozhaosengzs/ENVDS/master/final_project/population_p_6yrs.csv")
kable(head(data.area[,2:10]))
data.area.lite <- data.area %>% select(c(MapDate, None, D0, D1, D2, D3, D4))
colnames(data.area.lite) <-
  c("MapDate","Area_None","Area_D0", "Area_D1", "Area_D2", "Area_D3", "Area_D4")

data.pop.lite <- data.pop %>% select(c(MapDate, None, D0, D1,
                                       D2, D3, D4, ValidStart, ValidEnd))

colnames(data.pop.lite) <-
  c("MapDate", "Pop_None", "Pop_D0", "Pop_D1",
    "Pop_D2", "Pop_D3", "Pop_D4", "ValidStart", "ValidEnd")

data.complete <- full_join(
  data.area.lite, data.pop.lite, by = c("MapDate" = "MapDate"))

# Label groups by year (e.g. 2014-11-18 ~ 2015-11-16 as group 1)
# Each year starting from 2014-11-18 is counted as 51 weeks
data.complete$Year_Group <- NA
data.complete$Year_Group[263:314] <- 1
data.complete$Year_Group[211:262] <- 2
data.complete$Year_Group[159:210] <- 3
data.complete$Year_Group[107:158] <- 4
data.complete$Year_Group[55:106] <- 5
data.complete$Year_Group[3:54] <- 6

#two weeks are left out, so we exclude them from the data set
data.complete <- data.complete[3:314,]
data.year.mean.area <-
  group_by(data.complete, Year_Group) %>%
  summarise_at(vars(Area_None, Area_D0, Area_D1, Area_D2, Area_D3, Area_D4),
               mean, na.rm = TRUE)

kable(data.year.mean.area)
data.year.mean.pop <-
  group_by(data.complete, Year_Group) %>%
  summarise_at(vars(Pop_None, Pop_D0, Pop_D1, Pop_D2, Pop_D3, Pop_D4),
               mean, na.rm = TRUE)

kable(data.year.mean.pop)
# Excluding D4 in the process because all entries are 0
```

```r
print("Manova on Area")
manova.area <- manova(cbind(Area_None, Area_D0, Area_D1, Area_D2, Area_D3) ~
                        Year_Group, data = data.complete)

summary(manova.area, tol=0)

print("-------------------------------------------------------------")
print("Manova on population")
manova.pop <- manova(cbind(Pop_None, Pop_D0, Pop_D1, Pop_D2, Pop_D3) ~
                       Year_Group, data = data.complete)

summary(manova.pop, tol=0)
data.complete$DSCI_Area = 1 * data.complete$Area_D0 +
  2 * data.complete$Area_D1 + 3 * data.complete$Area_D2 +
  4 * data.complete$Area_D3 + 5 * data.complete$Area_D4

data.complete$DSCI_Pop = 1 * data.complete$Pop_D0 +
  2 * data.complete$Pop_D1 + 3 * data.complete$Pop_D2 +
  4 * data.complete$Pop_D3 + 5 * data.complete$Pop_D4

DSCI <- ggplot(data = data.complete) +
  geom_line(aes(x = ValidStart, y = DSCI_Area, colour = "Area"),
            size = 1, alpha = 0.7) +
  geom_line(aes(x = ValidStart, y = DSCI_Pop, colour = "Population"),
            size = 1, alpha = 0.7) +
  labs(title="Drought Severity And Coverage Index of Rhode Island (2014 - 2020)",
       x ="Date", y = "DSCI") +
  scale_colour_manual(name = "Variable of Interest",
                      values=c(Area ="yellow4", Population="brown")) +
  theme_classic()
DSCI
corr <- ggscatter(data.complete, x = "DSCI_Area", y = "DSCI_Pop",
          add = "reg.line",
          add.params = list(color = "blue", fill = "lightgray"),
          conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Area Index", ylab = "Population Index")
corr
data.complete$DSCI_Diff = abs(
  data.complete$DSCI_Area -
    data.complete$DSCI_Pop) / (data.complete$DSCI_Area + data.complete$DSCI_Pop)

data.complete$DSCI_Diff = data.complete$DSCI_Area - data.complete$DSCI_Pop
knitr::include_graphics("20150728_20200825.png")
```

For complete project repository please go to https://github.com/guozhaosengzs/ENVDS/tree/master/final_project.