STAT 482 PROJECT

Modeling California Wildfires

(Poisson Process)

Submitted To: Dr. Korosteleva

Prepared By: Joe Soria

29 November 2022

Contents

Introduction

Besides working on statistics during the semester, I have hobbies in botany and hiking. I take advantage of hikes to test my knowledge on plants and the ecosystems they inhabit. I have been wanting to work on a data set related to my hobbies in previous classes for Dr. Olga but had no success, until now. The dataset was surprisingly easy to find and piqued my interest immediately, it was titled "California Wildfires (2013-2020)". The dataset included what I needed to create a model using the Poisson process. I was keen on visualizing wildfire occurrences and utilizing the Poisson process to do this and more.

Background

California has frequently seen numerous record-breaking wildfires this century. According to the California Department of Forestry and Fire Protection (Cal Fire), nine of the ten biggest wildfires in the state's history have occurred in the past decade. And eight of those nine have occurred in the past five years. From 2010 to 2020, 11% of California's land mass burned according to NASA.gov. These fires have burned millions of acres each year and taken numerous lives as well according to Cal Fire. Furthermore, the average cost of suppression has averaged over $100 million since 2010, with the biggest spending year coming in 2018-2019 at $890 million (Cal Fire). It is known that there is a season for wildfires. But is there a way we can statistically model wildfires and gain a better understanding of them in LA and nearby counties?

Data Description

The data was obtained from Kaggle.com. The original data set contains information on wildfires in California from 2013 to 2020. It is a csv file containing 40 variables and 1,634 observations. For the analysis I used the two variables "Counties" and "Started." For the variable "Counties" I filtered for Los Angeles, San Bernardino, and Riverside. For the variable "Started" I extracted the date in each cell using an Excel function and I only selected observations from 2019.

Results

I created the code in R Studio according to the Poisson process. I created a histogram and conducted a chi-squared goodness-of-fit test to see if the times followed an exponential distribution. The resulting histogram showed characteristics of an exponential distribution. I further conducted the goodness-of-fit test, where I found the *p-value* to be $p = 0.5175706$. This value is greater than 0.05, indicating that the wildfires in the given time frame occurred according to a Poisson process. The mean was calculated from interarrival times, it came out

to $\mu = 5.56$ days. The value of lambda was calculated by taking the inverse of the mean, so $\lambda = \frac{1}{\mu} = \frac{1}{5.56} = 0.1796$ fires per day. The next predicted fire was obtained by taking the last observed fire in the data set and adding the mean, which is on November 6, 2019.

Conclusion

In summary, wildfires in the three counties occur at a rate of 0.1796 fires per day and a mean of 5.56 days. The next fire is projected to occur on November 6, 2019. The 2019 fire season actual trajectory was able to be plotted and so was a simulated trajectory using the Poisson process. This project provided me the opportunity to apply concepts learned in this Stat 482 class.

Appendix

A. Poisson process

a. R Studio Code

```
incidents <- read.csv(file = "C:/Users/t14nu/Desktop/STAT482/PROJECT/Code&Data/CAWildfires.csv",
                      header = TRUE, sep = ",")

# POISSON

# creating date-time variable
datetime<- as.POSIXct(as.Date(incidents$Date, "%m/%d/%Y"))
datetime

# computing lag
datetime.lag<- c(0,head(datetime, -1))

# computing inter arrival times (in days) and removing 1st val
int.time<- (as.numeric(datetime)-as.numeric(datetime.lag))/(3600*24)
int<- int.time[-1]

# plotting histogram
hist(int, main="", col="green", xlab="Interarrival Time")

#binning inter arrival times
binned.int<- as.factor(ifelse(int<5,"1",
                              ifelse(int>=5 & int<10,"2",ifelse(int>=10 & int<15,"3","4"))))

#computing observed frequencies
obs<- table(binned.int)
obs

#estimating mean for exponential distribution
mean.est<- mean(int)
mean.est

#computing expected frequencies
exp<- c(1:4)
exp[1]<- length(int)*(1-exp(-5/mean.est))
exp[2]<- length(int)*(exp(-5/mean.est)-exp(-10/mean.est))
exp[3]<- length(int)*(exp(-10/mean.est)-exp(-15/mean.est))
exp[4]<- length(int)*exp(-15/mean.est)
round(exp,1)

#computing chi-squared statistic
print(chi.sq<- sum((obs-exp)^2/exp))

#computing p-value
print(p.value<- 1-pchisq(chi.sq, df=2))
```

```
# mean
datetime<- as.POSIXct(as.Date(incidents$Date, "%m/%d/%Y"))

datetime.lag<- c(0,head(datetime, -1))

int.time<- (as.numeric(datetime)-as.numeric(datetime.lag))/(3600*24)
int<- int.time[-1]

mean.est<- mean(int)
mean.est

# lambda estimate
lambda.est <- 1/mean.est

# waiting time until next fire
nextFire <- 39/lambda.est
nextFire
```

```
#---PLOTTING ACTUAL TRAJECTORY---
fires <- c(1:38)
date<- as.POSIXct(as.Date(incidents$Date, "%m/%d/%Y"))

#plotting stock price against date
plot(date, fires, type="n",
     xlab="Time", ylab="Number of Fires", first.panel=grid())


segments(date[-length(date)], fires[-length(date)], date[-1]-0.07, fires[-length(date)],
         lwd=2, col="purple")
points(date, fires, pch=20, col="blue")
points(date[-1], fires[-length(date)], pch=1, col="purple")
```

```
#---SIMULATING TRAJECTORIES---
lambda.est <- 1 / mean.est
t<- 10
nfires <- 20
lambda.est


#defining states
N<- 0:nfires

#setting time as vector
time<- c()

#setting initial value for time
time[1]<- 0

#specifying seed
set.seed(483650)

#simulating trajectory
for (i in 2:(nfires+1))
  time[i]<- time[i-1]+round((-1/lambda.est)*log(runif(1)),2)

#plotting trajectory
plot(time, N, type="n", xlab="Days", ylab="Number of Fires", panel.first = grid())
segments(time[-length(time)], N[-length(time)], time[-1]-0.07, N[-length(time)],
         lwd=2, col="brown")
points(time, N, pch=20, col="red")
points(time[-1], N[-length(time)], pch=1, col="red")
```

b. Results

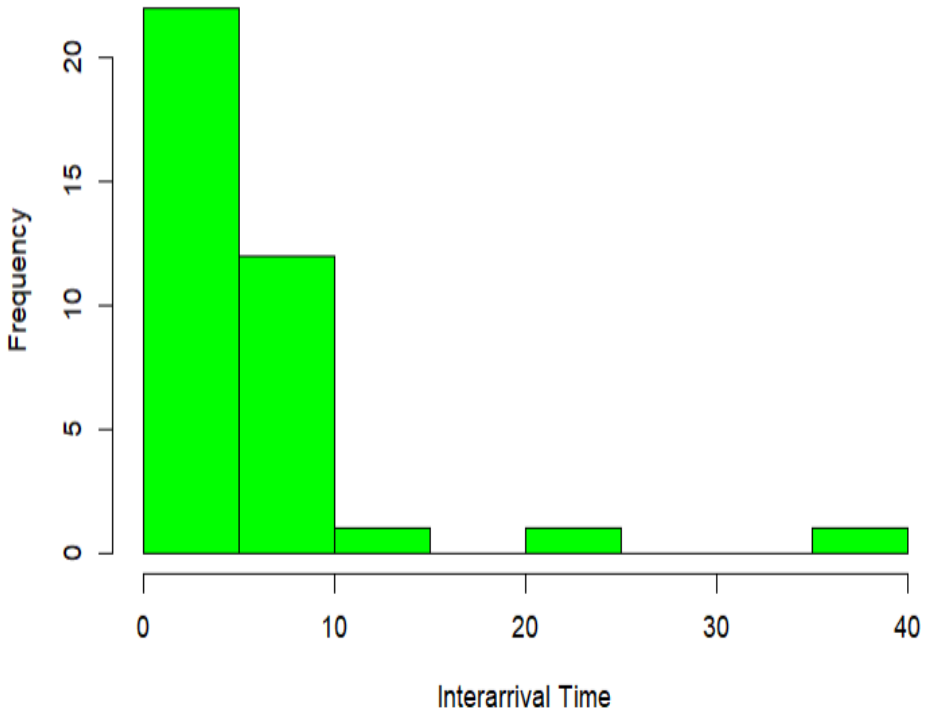

Figure 1: Histogram showing interarrival times

```
> obs
binned.int
 1  2  3  4
22 11  2  2
> round(exp,1)
[1] 21.9  8.9  3.6  2.5
> #computing chi-squared statistic
> print(chi.sq<- sum((obs-exp)^2/exp))
[1] 1.317218
> #computing p-value
> print(p.value<- 1-pchisq(chi.sq, df=2))
[1] 0.5175706
```

Figure 2: Results showing binned intervals, chi-squared, and goodness-of-fit test

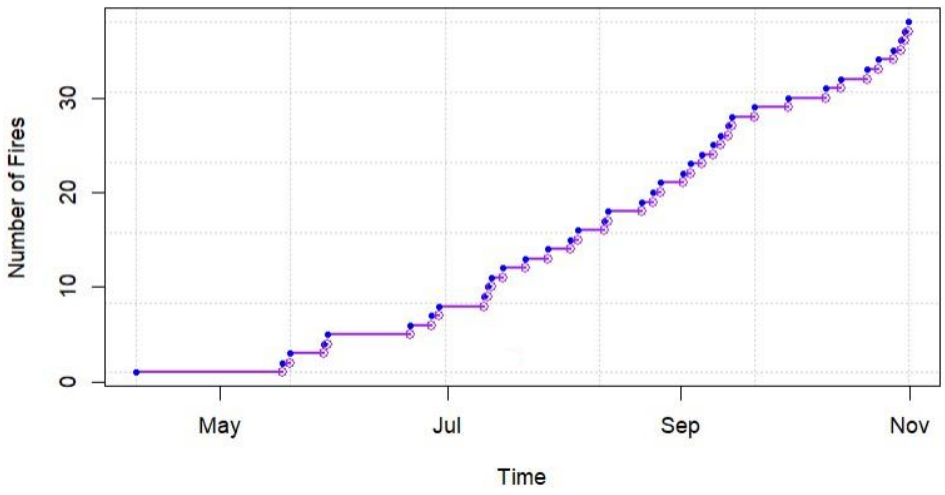

Figure 3: Actual trajectory plotted

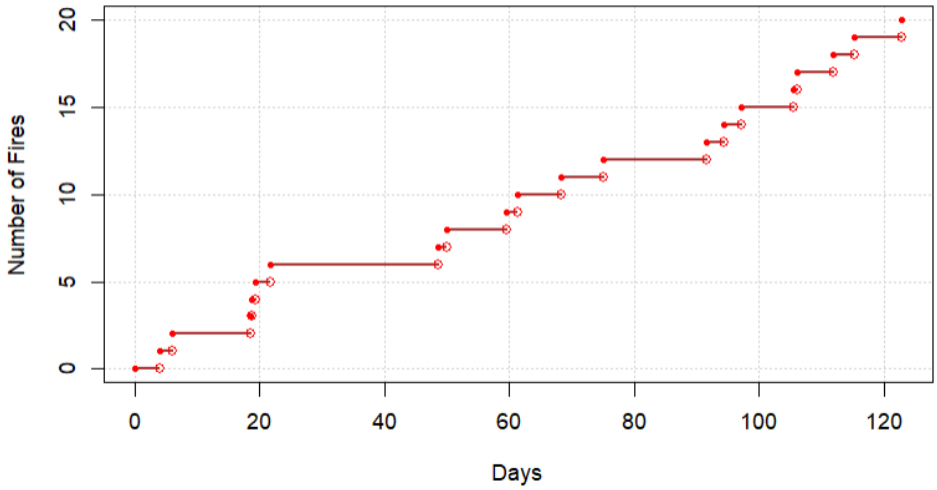

Figure 4: Simulated trajectory plotted

```
> mean.est
[1] 5.567568
```

Figure 5: Mean of interarrivals, in days

```
> lambda.est
[1] 0.1796117
```

Figure 6: Estimated Lambda calculated from inverse of mean, in fires per day

Date	Mean	Next Fire
11/1/2019	5.56	11/6/2019

Figure 7: Adding mean to last observed date in data set to obtain date of next fire

References

California Wildfires (2013-2020). (2020, Feb 9). Kaggle. Retrieved November 4, 2022. https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020

Cal Fire. (2022, Oct 2). *Top 20 Largest California Wildfires*. fire.ca.gov. https://www.fire.ca.gov/media/4jandlhh/top20_acres.pdf

NASA. (2021, Oct 4). *What's Behind California's Surge of Large Fires?* NASA Earth Observatory. https://earthobservatory.nasa.gov/images/148908/whats-behind-californias-surge-of-large-fires