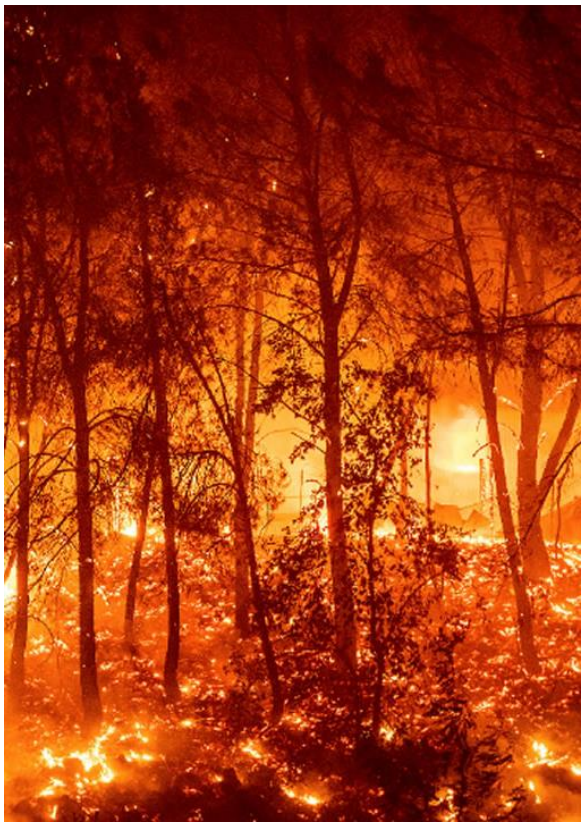
Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación



# PROYECTO FINAL

Análisis de la actividad de los incendios forestales en el oeste de Estados Unidos

MINERÍA DE DATOS

Ing. En Ciencias de la Computación
Catedrático: Dr. Abraham Sánchez López
Presenta: Alondra Sánchez Molina
Secc. 002

# Índice

## Introducción

Los incendios forestales han consumido áreas cada vez mayores de los bosques del oeste de EE. UU. Cientos de hogares son quemados anualmente por incendios forestales y los daños a los recursos naturales son a veces extremos e irreversibles. Se cree que la actividad de los incendios forestales en el oeste de los Estados Unidos ha aumentado en las últimas décadas, sin embargo, ni el alcance de los cambios recientes ni el grado en que el clima puede estar impulsando cambios regionales en los incendios forestales se ha documentado sistemáticamente. Gran parte del debate público y científico sobre los cambios en los incendios forestales del oeste de Estados Unidos se ha centrado en cambio en los efectos de la historia del uso de la tierra en los siglos XIX y XX.

Los estudios sugieren que, durante las últimas décadas, la cantidad y el tamaño de los incendios forestales han aumentado en todo el oeste de los Estados Unidos. La duración promedio de la temporada de incendios forestales también ha aumentado significativamente en algunas áreas. Según la Unión de Científicos Preocupados (UCS), todos los estados del oeste de los EE. UU. Han experimentado un aumento en el número promedio anual de grandes incendios forestales (más de 1,000 acres) durante las últimas décadas. El noroeste del Pacífico, incluidos Washington, Oregón, Idaho y la mitad occidental de Montana, han tenido temporadas de incendios forestales particularmente desafiantes en los últimos años.

La temporada de incendios forestales de 2017 rompió récords y le costó al Servicio Forestal de los EE. UU. $2 mil millones sin precedentes. Desde los incendios forestales de Oregón hasta los incendios de finales de temporada en Montana, y el momento muy inusual de los incendios de California en diciembre, fue un año ajetreado en el oeste de los Estados Unidos.

Numerosos estudios han encontrado que los grandes incendios forestales en el oeste de los EE. UU. Se han producido casi cinco veces más a menudo desde los años setenta y ochenta. Estos incendios están quemando más de seis veces la superficie terrestre que antes y duran casi cinco veces más.

La mayor parte del oeste de los EE. UU. es árido con una preponderancia de precipitación anual que llega en invierno. Gran parte del área forestal se concentra en cadenas montañosas donde los efectos orográficos aumentan la cantidad de precipitación y la elevación aumenta la probabilidad de que las precipitaciones invernales caigan en forma de nieve que puede acumularse y llevar la humedad de la precipitación de la estación fría al verano más árido.

Se cree que el cambio climático es la causa principal del aumento de los grandes incendios forestales con temperaturas en aumento que conducen a un volumen más temprano y menor de nieve que se derrite, una disminución de las precipitaciones y condiciones del bosque que son más secas durante períodos de tiempo más largos.

Un aumento de las enfermedades de los árboles forestales por perturbación de insectos también se ha asociado con el cambio climático y puede conducir a grandes áreas de bosques muertos o moribundos altamente inflamables. Otras causas potenciales del aumento de la actividad de los incendios forestales incluyen las prácticas de manejo forestal y un aumento de los incendios forestales causados por humanos debido a accidentes o incendios.

Mediante el siguiente análisis, se plantea corroborar si la cantidad y el tamaño de los incendios a aumentado en las últimas décadas, mas concretamente, entre los años de 1980 y 2016. Así como si el cambio climático es la principal causa de los incendios forestales en el oeste de los Estados Unidos.

Otra de las interrogantes a las cuales se planea dar solución es si, no solo ha aumentado el tamaño y la cantidad de incendios, si no que también, la temporada en la que se presentan ha crecido. Además, si la superficie quemada difiere según la organización federal que ha registrado los incendios. Todo ello, centrándose en la oleada de los grandes incendios forestales. Dando respuesta a la pregunta de que, si ha aumentado la actividad y el tamaño de los incendios forestales, o simplemente parece así porque estamos más atentos a las malas noticias y las redes sociales.

## Desarrollo

En este caso de estudio, se utilizarán no solo las habilidades técnicas, si no que también, las habilidades analíticas que se han ido puliendo a lo largo del curso de Minería de datos.

### Datos

Los datos por tratar acerca de los incendios forestales fueron recuperados de la Base de datos federal de ocurrencia de incendios forestales, proporcionada por el Servicio Geológico de EE. UU. (USGS) para visualizar el cambio en la actividad de los incendios forestales de 1980 a 2016. El análisis se limitará al oeste de los Estados Unidos, incluidos California, Arizona, Nuevo México, Colorado, Utah, Nevada, Utah, Oregón, Washington, Idaho, Montana. y Wyoming.

El dataset StudyArrea.csv, cuenta con los siguientes atributos:

- FID. Número de id.
- ORGANIZATI. Agencia que reportó el incendio.
- UNIT. Unidad que reportó el incendio.
- SUBUNIT. Subunidad que reportó el incendio.
- SUBUNIT2. Unidad identificadora de reporte (USFS District Number).
- FIRENAME. El nombre del evento de incendio forestal.
- CAUSE. Causa categorizada por el USFS.
- YEAR_. Año del incendio.
- STARTDATED. Fecha de descubrimiento del incendio forestal.
- CONTRDATED. Fecha en que se controló el incendio forestal.
- OUTDATED. Fecha en que se declaró extinguido el incendio forestal.
- STATE. Estado en el cual sucedió el incendio.
- STATE_FIPS. Código FIPS.
- TOTALACRES. Acres en el momento de control.

### Herramienta utilizada

El estudio acerca de los incendios forestales fue realizado haciendo uso de R pues es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Es interpretado, es decir, ejecuta las instrucciones directamente, sin una previa compilación del programa a instrucciones en lenguaje máquina. Estas razones, entre otras, lo hace una herramienta ideal para realizar un análisis como el planteado.

## Exploración inicial

Primeramente, se realizó una exploración inicial al dataset con la finalidad de tener un primer acercamiento con ellos, se observan los tipos de datos de los atributos, así como, mediante un summary(), valores estadísticos iniciales.

```
> summary(wildfire)
      FID          ORGANIZATI           UNIT             SUBUNIT           SUBUNIT2          FIRENAME           CAUSE
 Min.   :     0   Length:439362      Length:439362      Length:439362      Length:439362      Length:439362      Length:439362
 1st Qu.:171157   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :289967   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :313041
 3rd Qu.:469159
 Max.   :589645
     YEAR_       STARTDATED         CONTRDATED         OUTDATED           STATE              STATE_FIPS     TOTALACRES
 Min.   :1980   Length:439362      Length:439362      Length:439362      Length:439362      Min.   : 4.0   Min.   :     0.0
 1st Qu.:1990   Class :character   Class :character   Class :character   Class :character   1st Qu.: 6.0   1st Qu.:     0.1
 Median :1998   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :16.0   Median :     0.1
 Mean   :1998                                                                             Mean   :22.2   Mean   :   191.1
 3rd Qu.:2006                                                                             3rd Qu.:41.0   3rd Qu.:     1.0
 Max.   :2016                                                                             Max.   :56.0   Max.   :590620.0
```

En esta primera etapa, además de obtener la clase de cada atributo, corroboramos que efectivamente, los años de los datos a tratar van de 1980 a 2016; así como se plantea la idea de que en algunos incendios forestales no se quemaron acres, mientras que existe al menos un incendio, donde la cantidad de acres quemados fue de 590,620.

## Preprocesamiento y preparación de los datos

En base a la información recabada, se optan por reducir la dimensión del dataset con la finalidad de obtener respuesta a las interrogantes planteadas anteriormente.

### Selección de atributos

Los atributos seleccionados se listan a continuación:

- ORGANIZATI. Agencia que reportó el incendio.
- CAUSE. Causa categorizada por el USFS.
- YEAR_. Año del incendio.
- STARTDATED. Fecha de descubrimiento del incendio forestal.
- OUTDATED. Fecha en que se declaró extinguido el incendio forestal.
- STATE. Estado en el cual sucedió el incendio.
- TOTALACRES. Acres en el momento de control.

### Filtrado de datos

Todo el análisis, se centrará en los incendios forestales con mayor tamaño, catalogándolos como los que queman mas de 1000 acres. Por esta razón, se realiza un filtrado a las filas las cuales cumplan la siguiente condición.

*wildfire <- filter(wildfire, TOTALACRES >= 1000)*

Un dataset, puede contar con campos vacíos, por ello, se realiza una búsqueda de celdas vacías, y se eliminarán para realizar un tratamiento sin datos incompletos.

```
> sum(is.na(wildfire))
[1] 268
> wildfire <- wildfire[complete.cases(wildfire),]
> sum(is.na(wildfire))
[1] 0
```

Anteriormente, se observó que las fechas en el dataset se encontraban en formato carácter, por esta razón, el siguiente paso es convertirlas en formato date, con la finalidad demás adelante poder operarlas sin ningún problema.

```
> wildfire$STARTDATED <- as.Date(wildfire$STARTDATED, "%m/%d/%y %H:%M")
> wildfire$OUTDATED <- as.Date(wildfire$OUTDATED, "%m/%d/%y %H:%M")
```

### Creación de variables

Se crearon dos nuevas variables: DECADE y DAYS_ON_FIRE. La primera, nace de la necesidad de agrupar los incendios por década, y la segunda, de tener el conocimiento del total de los días que duró el incendio.

La variable DECADE, se calcula mediante un algoritmo de decisión que va etiquetando cada incendio de acuerdo con el año en el que sucedió colocándole una etiqueta por década.

La variable DAYS_ON_FIRE, se calcula realizando una resta entre la fecha de finalización y la fecha de inicio. Es por ello que anteriormente fueron convertidas en date, para utilizar las matemáticas permitidas por la librería lubridate, anteriormente cargada.

## Regresión Lineal como método de análisis

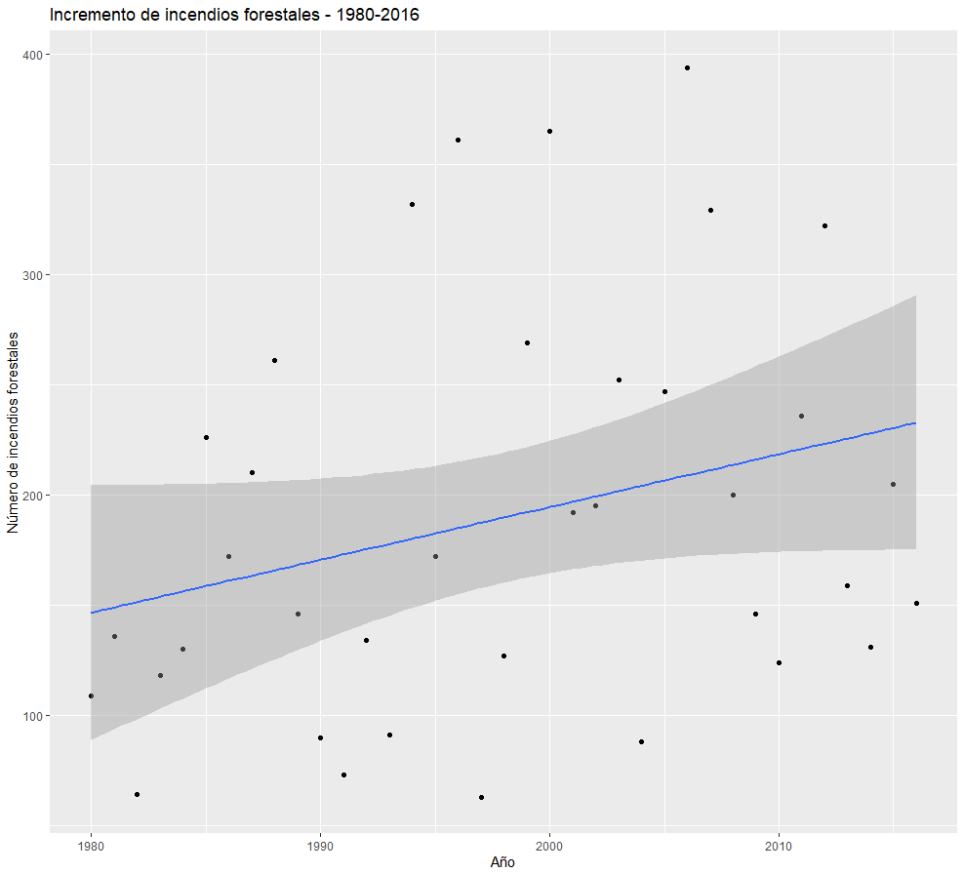
El método utilizado es la regresión lineal puesto que esta, identifica la recta óptima que atraviesa una nube de puntos y es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, y las variables independientes $X_i$ siendo un método del aprendizaje supervisado.

Las pruebas y resultados mostrados a continuación fueron ploteamientos de diagramas de dispersión, con los cuales se plantea observar la información, estos, contendrán una línea, la cual mostrará la predicción realizada a partir del uso de la regresión lineal. En algunos casos, incluirán un intervalo de confianza, y en otros, con fines visuales; no.

## Pruebas

Al inicio, se planteo la finalidad del estudio de los datos contenidos en el dataset, ahora, se dará respuesta a las preguntas realizadas.

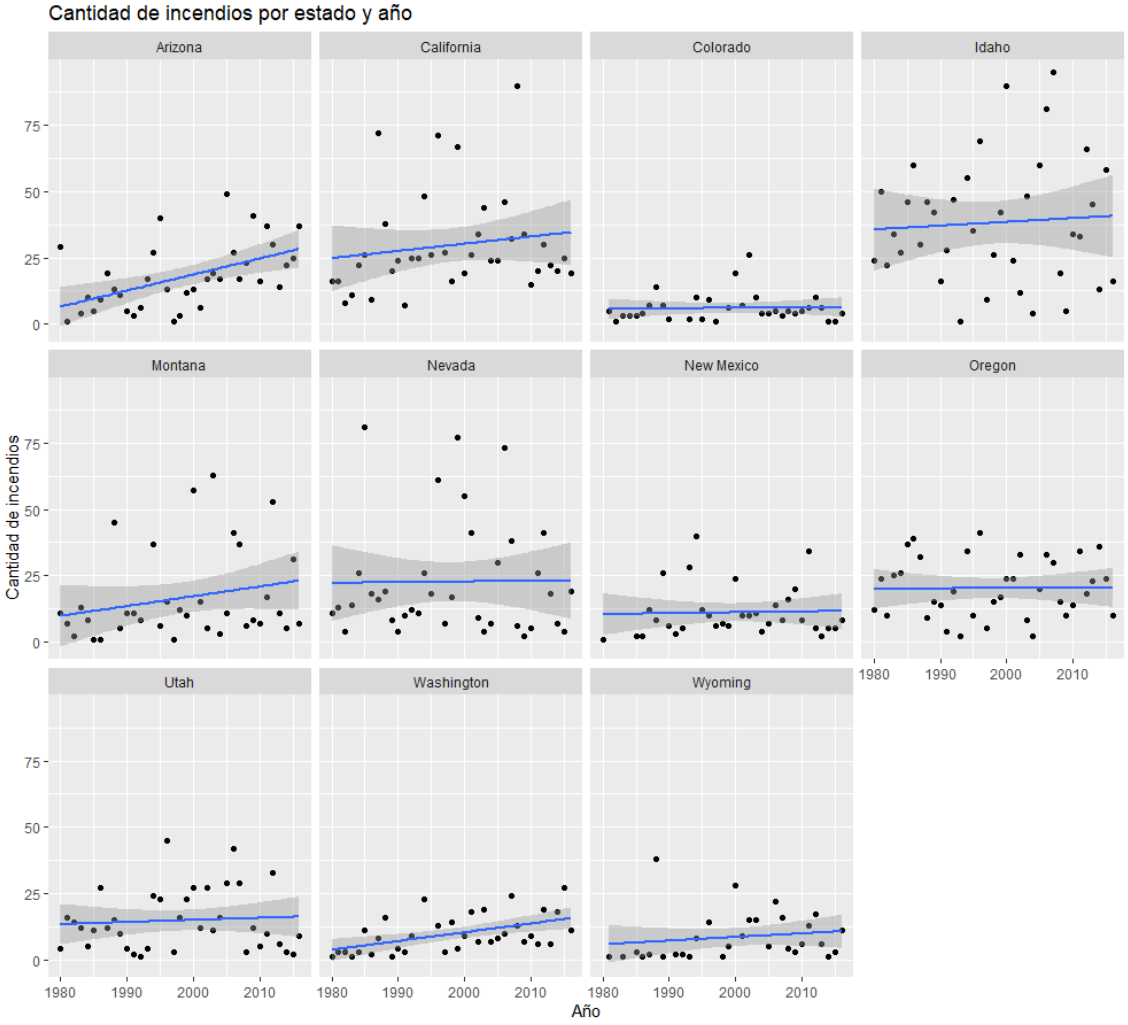
### Número de incendios forestales en las últimas décadas




En la gráfica obtenida se observa que existe un aumento en el numero de incendios forestales en las ultimas décadas, podemos observar, que conforme han pasado los años, el numero de incendios en el oeste de los Estados Unidos ha ido en aumento.

Siendo este análisis, llevado acabo con los datos donde los incendios quemaban igual o mas de 1000 acres en los once estados contenidos en el dataset. Por lo que podemos inferir, que, aunque el numero en general a aumentado, existe la posibilidad que no todos los estados hallan contribuido a esta tendencia.

## Exploración por estado

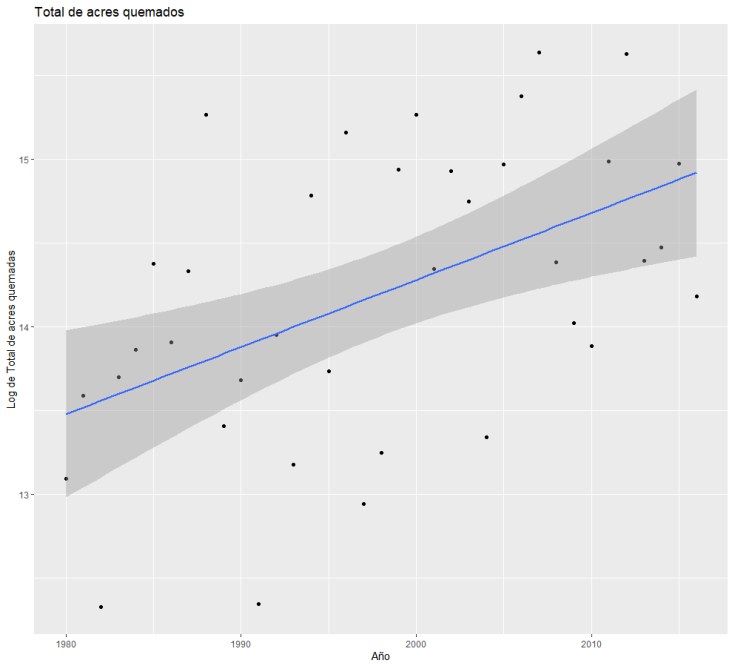



Al realizar el mismo procedimiento que en el grafico anterior, pero ahora por estado, se puede notar que, en algunos estados, el incremento ha sido notable, estados como Arizona, California, Montana y Washington; han tenido un aumento significativo al numero de incendios presentados.
Por otra parte, estados como Colorado, New México Oregón, Nevada y Utah; muestran tendencias al casi nulo aumento en su numero de incendios a lo largo del tiempo.

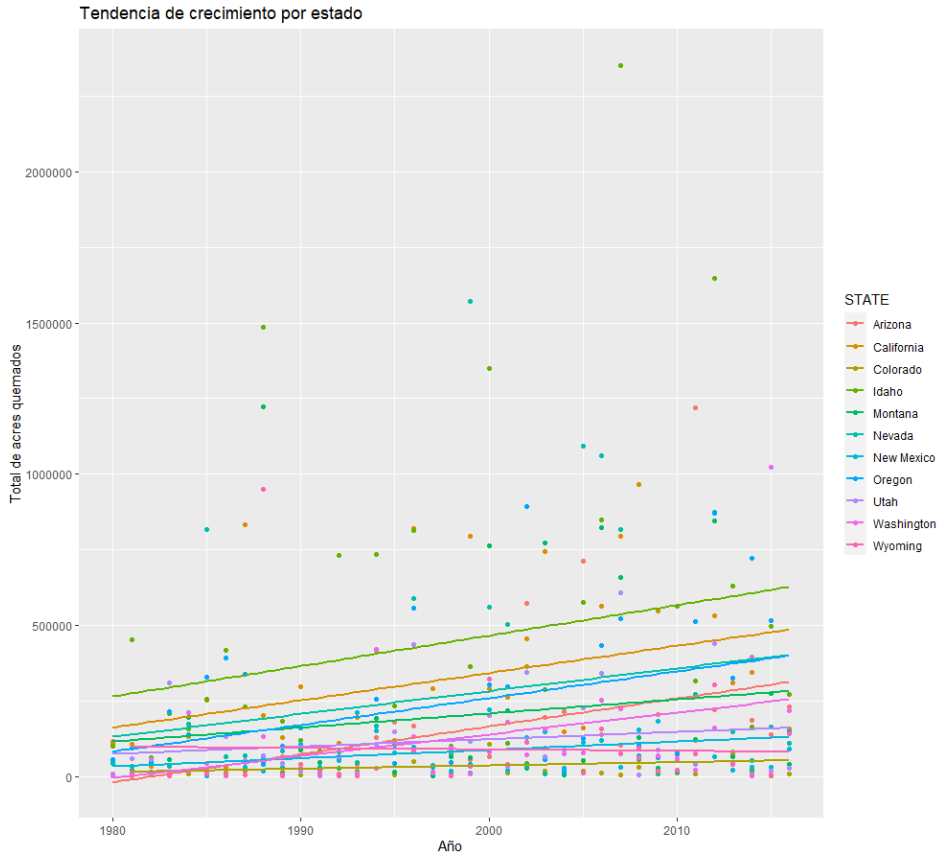
## Superficie quemada con el tiempo

El análisis realizado para el ploteamiento de este diagrama de dispersión fue el siguiente, se opto por agrupar el numero total de acres quemados por año, y convertirlo a una escala logarítmica para una mejor visualización.

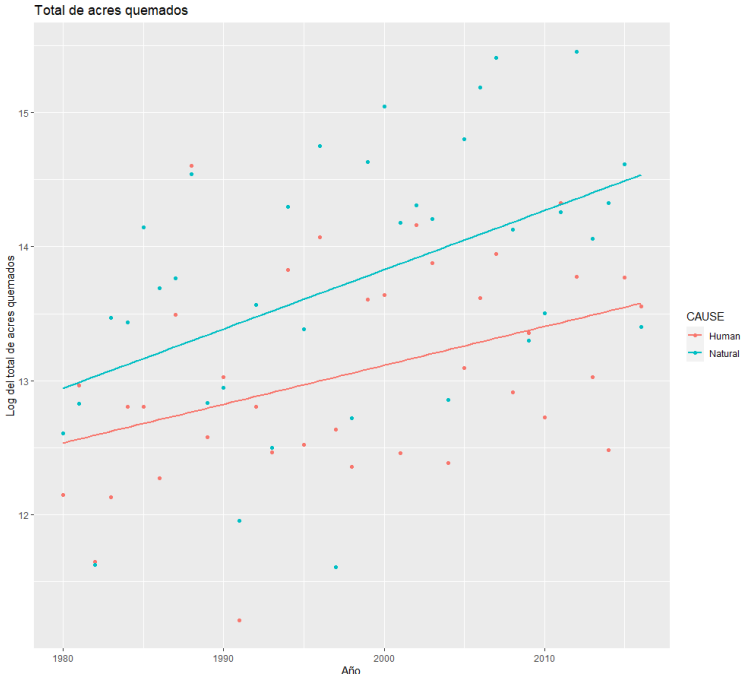
Como se observa a continuación, el numero de acres quemados al año, a aumentado significativamente para el oeste de Estados Unidos.




Si optamos por revisar las tendencias de acres quemados por estado, conocemos que, aunque la cantidad de incendios en Idaho no a tenido un gran aumento, el numero de acres quemados, sí. California, no solo ha crecido en numero de incendios, si no, además, en numero de acres perdidos por los incendios.



Estadísticamente hablando, el 33% de los incendios forestales analizados, ocurren por causas humanas, esto, es una cifra significativa. Al realizar un análisis por causa, se puede observar cómo tanto el nivel de acres perdidos, independientemente de la causa ha ido en aumento.




## Predicción

Mediante regresión lineal, se predice el siguiente modelo, con la finalidad de predecir el número de acres quemados, en un incendio metafórico.

```
Call:
lm(formula = TOTALACRES ~ ORGANIZATI + CAUSE + STATE + DECADE +
    DAYS_ON_FIRE, data = wildfirePred)

Residuals:
    Min      1Q  Median      3Q     Max
-150681   -8552   -5475   -1063  572290

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3419.48    1686.32   2.028   0.0426 *
ORGANIZATIBLM      2543.11    1337.68   1.901   0.0573 .
ORGANIZATIBOR      1389.28   12586.83   0.110   0.9121
ORGANIZATIFS       -294.54    1393.10  -0.211   0.8326
ORGANIZATIFWS      4865.22    5521.12   0.881   0.3782
ORGANIZATINPS       -50.30    2093.11  -0.024   0.9808
CAUSENatural       -280.28     758.66  -0.369   0.7118
CAUSEUndetermined  4286.90   10342.35   0.414   0.6785
STATECalifornia    1114.94    1446.37   0.771   0.4408
STATEColorado     -3929.33    2288.77  -1.717   0.0861 .
STATEIdaho         2648.14    1371.87   1.930   0.0536 .
STATEMontana       1529.21    1615.56   0.947   0.3439
STATENevada        3694.49    1552.60   2.380   0.0174 *
STATENew Mexico    -992.52    1847.92  -0.537   0.5912
STATEOregon        1843.55    1553.67   1.187   0.2354
STATEUtah          -862.17    1687.47  -0.511   0.6094
STATEWashington    2778.97    1907.95   1.457   0.1453
STATEWyoming       -275.76    2109.45  -0.131   0.8960
DECADE1990-1999      59.98     983.08   0.061   0.9514
DECADE2000-2009    2170.20     943.59   2.300   0.0215 *
DECADE2010-2016    4303.45    1072.86   4.011  6.1e-05 ***
DAYS_ON_FIRE         81.56       7.12  11.454  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27910 on 6997 degrees of freedom
Multiple R-squared:  0.03089,   Adjusted R-squared:  0.02798
F-statistic: 10.62 on 21 and 6997 DF,  p-value: < 2.2e-16
```

Con él, se llevó acabo la siguiente predicción:

Predecir el total de acres quemados cuando las variables dependientes tengan los siguientes valores:

ORGANIZATI = "NPS",
CAUSE = "Human",
STATE = "California",
DECADE = "2010-2016",
DAYS_ON_FIRE = 10

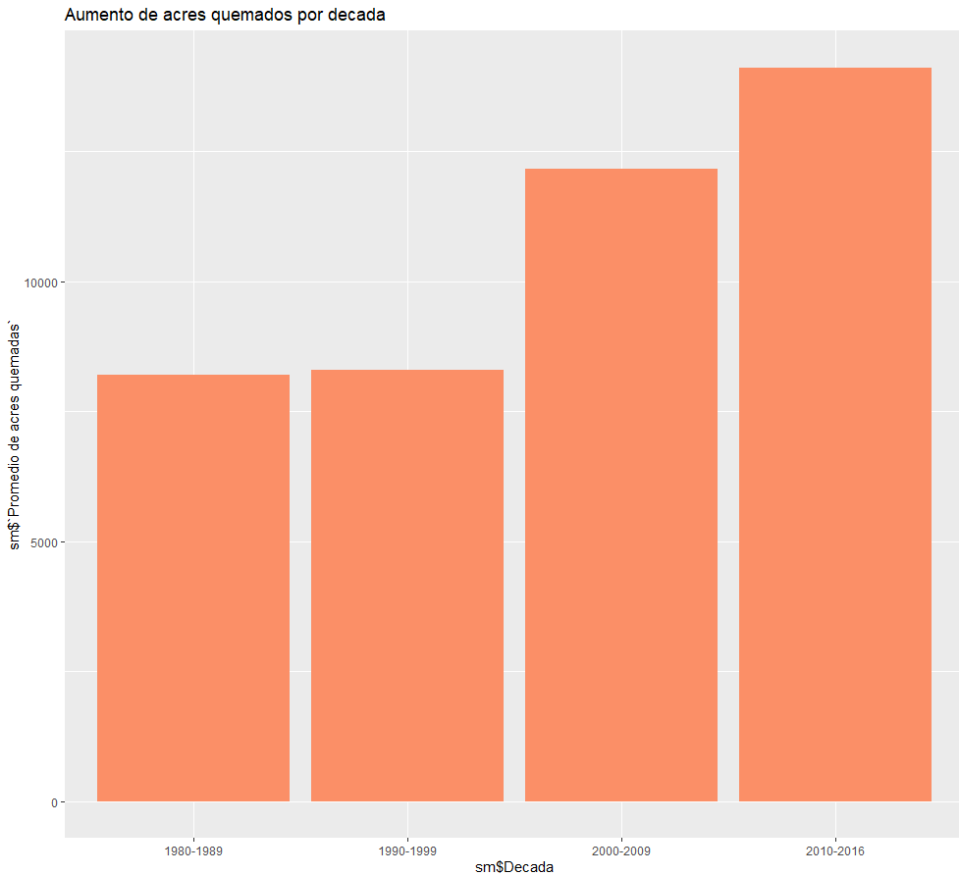
Resultado obtenido:

```
> acres_pred
       1
9603.124
```

Es decir, en un incendio que presente las anteriores características descritas, es probable que la cantidad de acres quemados sean 9603.

## Tamaño de los incendios forestales individuales con el tiempo

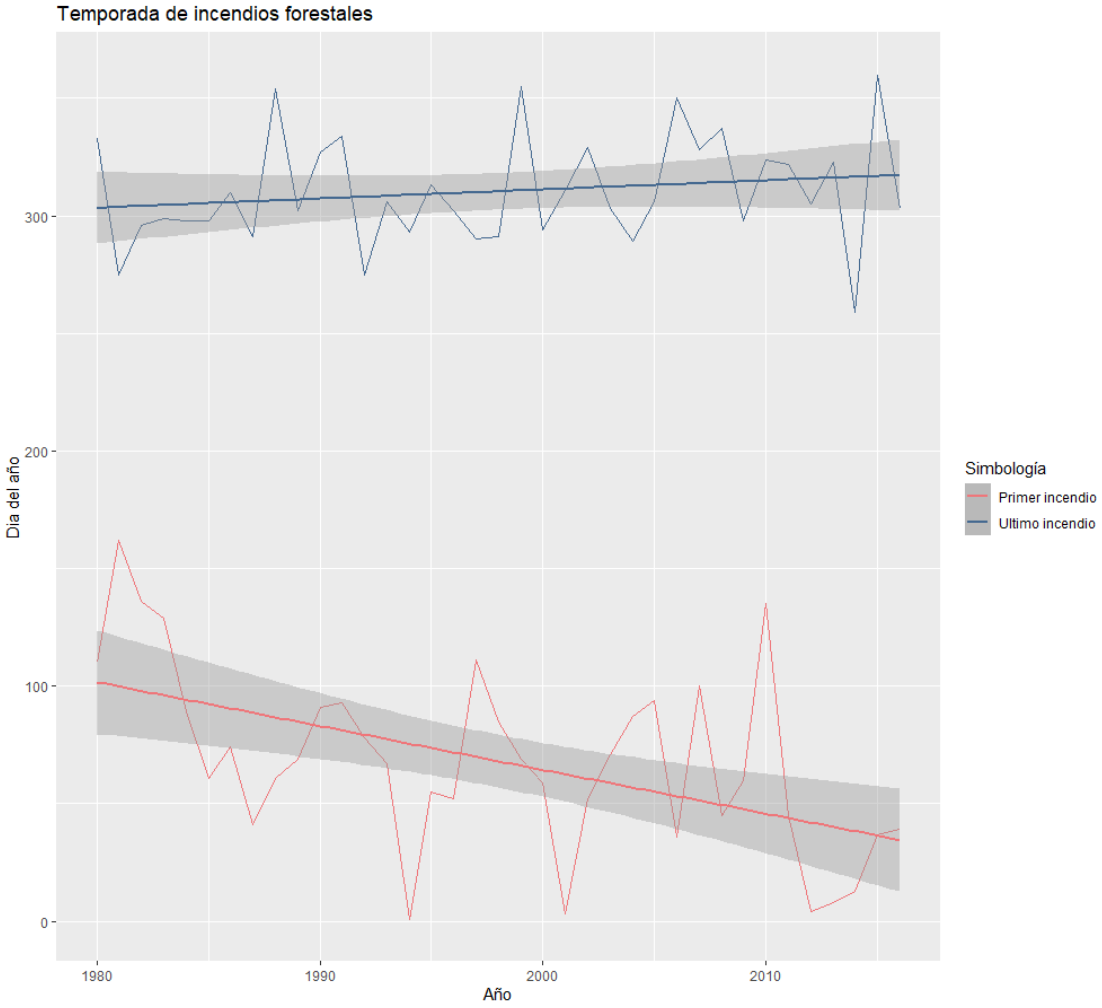
Para responder a la pregunta de que, si ha habido un aumento con el tiempo de los tamaños de los incendios forestales, se utiliza la variable década y se grafica con respecto al promedio de acres quemados. Como se puede notar a continuación, las primeras dos décadas que incluye el dataset, de 1980 a 1999, mantuvo un valor parecido, pero, para las ultimas dos, hubo un aumento notorio en e incremento de acres quemados.




## Duración de la temporada de incendios con el tiempo

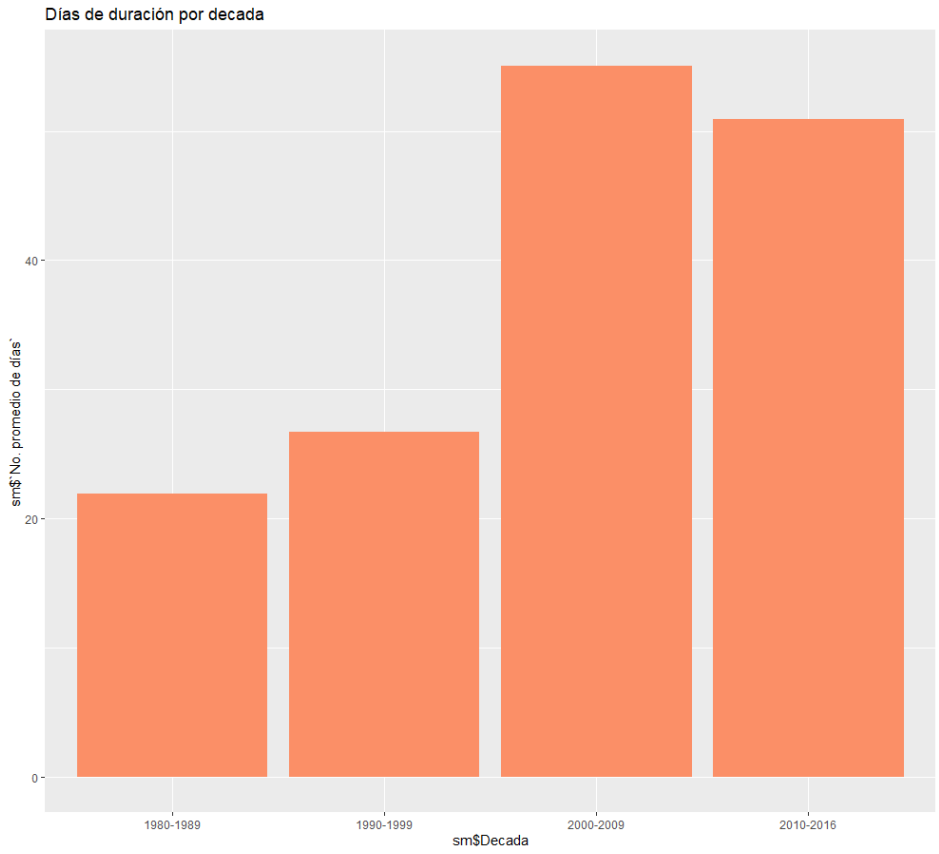
La temporada de incendios, se define como el periodo de tiempo en un año en el cual el número de incendios forestales es constante; es por ello, que para definir dicha temporada se utiliza la fecha de inicio del primero incendio del año, y la fecha del ultimo incendio del año.

Con esta información, podemos plotear la información para observar que ha habido un aumento en la temporada de incendios, ya que los incendios tienden a empezar mucho antes que en las décadas anteriores; el incremento en el final de temporada también creció, pero el inicio de temporada creció de una manera significativa. Anteriormente, la temporada de incendios comenzaba en el cuarto mes del año; en el ultimo año registrado, esto cambio a iniciar en el segundo.




Además de la temporada se decidió, visualizar el promedio de los incendios en días por década, la gráfica que a continuación se presenta, cumple la función de indicarnos dicha información.

Se hace notar, que los incendios en la década de 1980 – 1989 duraban un promedio de 22 días, pero en la última década disponible, duraban un periodo de 51 días, mas del doble que hace tres décadas.




## Predicción

Mediante regresión lineal, se predice el siguiente modelo, con la finalidad de predecir el número de días de duración en un incendio metafórico.

```
Call:
lm(formula = DAYS_ON_FIRE ~ ORGANIZATI + CAUSE + STATE + TOTALACRES +
    DECADE, data = wildfirePred)

Residuals:
    Min      1Q  Median      3Q     Max
-131.46  -18.87   -4.61   11.16 1879.86

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.043e+01  2.803e+00  -3.722 0.000199 ***
ORGANIZATIBLM     -6.603e+00  2.224e+00  -2.969 0.003001 **
ORGANIZATIBOR     -1.787e+01  2.094e+01  -0.853 0.393442
ORGANIZATIFS       4.405e+01  2.257e+00  19.520  < 2e-16 ***
ORGANIZATIFWS     -1.228e+01  9.184e+00  -1.338 0.181076
ORGANIZATINPS      4.446e+01  3.441e+00  12.919  < 2e-16 ***
CAUSENatural       1.069e+01  1.256e+00   8.518  < 2e-16 ***
CAUSEUndetermined  9.152e+00  1.720e+01   0.532 0.594772
STATECalifornia    2.453e+01  2.388e+00  10.272  < 2e-16 ***
STATEColorado      1.748e+01  3.802e+00   4.596 4.37e-06 ***
STATEIdaho         1.322e+01  2.277e+00   5.807 6.63e-09 ***
STATEMontana       2.689e+01  2.668e+00  10.076  < 2e-16 ***
STATENevada        8.874e+00  2.582e+00   3.437 0.000591 ***
STATENew Mexico    8.659e+00  3.072e+00   2.818 0.004840 **
STATEOregon        2.864e+01  2.562e+00  11.179  < 2e-16 ***
STATEUtah          1.604e+01  2.801e+00   5.727 1.07e-08 ***
STATEWashington    2.764e+01  3.157e+00   8.755  < 2e-16 ***
STATEWyoming       2.012e+01  3.501e+00   5.747 9.45e-09 ***
TOTALACRES         2.257e-04  1.970e-05  11.454  < 2e-16 ***
DECADE1990-1999    1.052e+00  1.635e+00   0.643 0.520126
DECADE2000-2009    2.010e+01  1.552e+00  12.956  < 2e-16 ***
DECADE2010-2016    1.728e+01  1.775e+00   9.737  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.44 on 6997 degrees of freedom
Multiple R-squared:  0.3088,    Adjusted R-squared:  0.3067
F-statistic: 148.8 on 21 and 6997 DF,  p-value: < 2.2e-16
```

Con él, se llevó acabo la siguiente predicción:

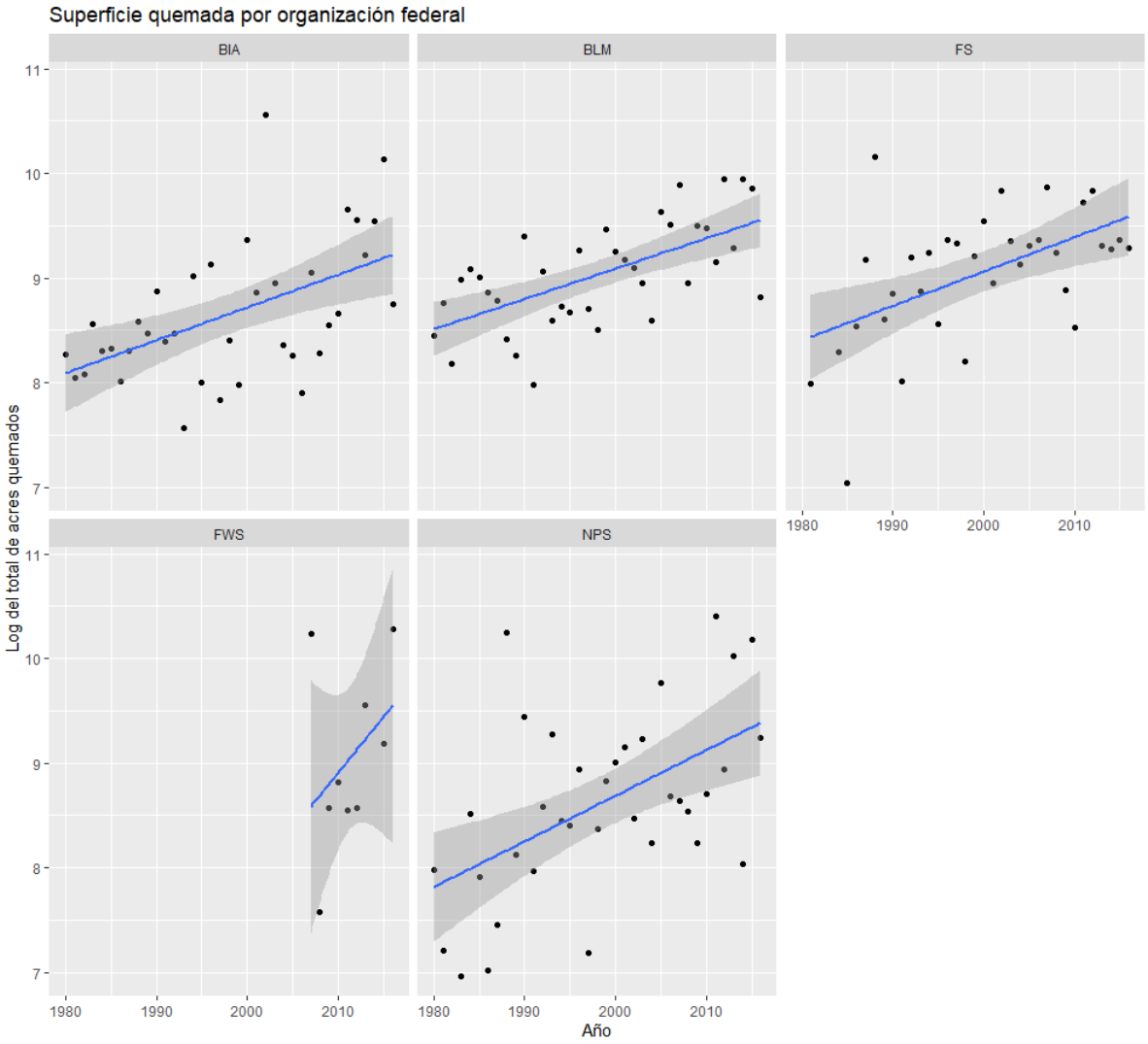
Predecir el número de días de duración del incendio cuando las variables dependientes tengan los siguientes valores:

ORGANIZATI = "FS",
CAUSE = "Human",
STATE = "Arizona",
TOTALACRES = 1000,
DECADE = "2010-2016"

Resultado obtenido:

```
> days_pred
       1
51.12391
```

## Acres quemados según la organización federal




Para obtener un resultado acerca de este tema, se ploteo el log del total de acres quemados por año según la organización que registro cada incendio. Se observa, que la superficie quemada, independientemente de la organización de la cual vengan los datos, muestra un incremento serio sobre los acres perdidos por los incendios forestales de su zona.

## Conclusiones

El presente estudio, ha tenido como finalidad el explorar los incendios forestales que han ocurrido en la zona del oeste de Estados Unidos, haciendo énfasis en once países específicos y las organizaciones que registran estos. Las variables de estudio analizadas fueron los acres quemados, la duración de los incendios forestales, la temporada en la cual ocurre, la cantidad de ellos y tendencias por estado, así como por organizaciones.

Todo ello se llevó acabo con el fin de descubrir que tan veras es la información que se difunden a través de las redes sociales sobre el gran aumento que ha habido en la actividad y el tamaño de los incendios forestales. A través del trabajo aquí realizado, dicha información se puede corroborar como veras, puesto que el análisis de los datos de 1980 a 2016 así lo indican.

Estudios previos han sugerido que la actividad de incendios forestales en el oeste de los EE. UU. está aumentando debido a un clima más cálido y al deshielo primaveral más temprano, aquí, vemos que la frecuencia de los incendios forestales y el área quemada en los bosques del noroeste del Pacífico han aumentado más rápidamente, y que esto no ha sido simplemente por pausas naturales, el 33% de los incendios causados en el periodo analizado son causados por los humanos, y estos, también han ido en aumento.

Estados como Arizona y Washington, han tenido un aumento en la cantidad de incendios a lo largo de los años, pero estados como Idaho y Montana son los que presentan un mayor incremento en la cantidad de acres quemados.

Cada vez son más la cantidad de acres que se pierden, en general, la temporada de incendios forestales ha aumentado significativamente, dejando no solo perdidas de bienes materiales o naturales, sino además un gran derrame económico, e inclusive a veces cobrando vidas humanas.

## Fuentes consultadas


- Department of the Interior by the United States Geological Survey. (2019, 30 octubre). Federal Wildland Fire Occurrence Data. Federal Fire Occurrence Website. https://wildfire.cr.usgs.gov/firehistory/data.html
- Spracklen, D. V., Mickley, L. J., Logan, J. A., Hudman, R. C., Yevich, R., Flannigan, M. D., & Westerling, A. L. (2009). Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. Journal of Geophysical Research: Atmospheres, 114(D20).
- Westerling, A. L. (2016). Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1696), 20150178.
- Westerling, A. L., Hidalgo, H. G., Cayan, D. R., & Swetnam, T. W. (2006). Warming and earlier spring increase western US forest wildfire activity. science, 313(5789), 940-943.
- NASA. 2012. "Climate Models Project Increase in U.S. Wildfire Risk." Accessed September 15, 2015. http://www.nasa.gov/topics/earth/features/climate-fire.html.
- Canavos, George C.; Probabilidad y Estadística. Aplicaciones y Métodos. McGraw-Hill. México. ISBN 9684518560.