UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS

FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS

" *La ciencia es sólo curiosidad organizada* "
-ERIC LANDER

MM-530

TEORÍA DE MUESTREO

INFORME PROYECTO FINAL:
"Análisis de base de datos de fulguraciones solares"

Primer periodo, 2019
Fecha: Jueves 2 de Mayo del 2019

Entregado por:

Luis Felipe Flores Machado

Número de cuenta:

20141030258

Contents

Parte I

SECCIÓN PREAMBULAR

"Ahora soy un genio.
Antes de eso, fui un esclavo del trabajo."

-Niccolo Paganini

1.1 Introducción

En el presente informe se realiza un estudio breve de la base de datos de con 113,942 observaciones y 18 variables, bajo el nombre de `hessi.solar.flare.2002to2016`. Esta servirá como población la cual se examinará mediante técnicas de muestreo aprendidas en clase.

Esta base de datos comprende las observaciones obtenidos por el experimento "Reuven Ramaty High Energy Solar Spectroscopic Imager" lanzado por la Administración Nacional de la Aeronáutica y del Espacio (NASA) en el 2002 con el objetivo de explorar la física de aceleración de partículas en altas energías y su radiación electromagnética percibible desde el sol.

Como ejemplo principal en esta base de datos tenemos las **fulguraciones solares** (conocido quizá mejor por su nombre en inglés: solar flares), los cuales siempre están emparejados con otro fenómeno que es de naturaleza espacial/geométrico: Las manchas solares. Estas consisten en regiones del sol cuya iluminación percibida se ve mucho más ténue de lo usual debido a actividades magneticas subyacentes. Se realiza un análisis tanto conceptual-teórico del fenómeno subyacente a los datos extraídos, como un análisis espacial,cuantitativo y categórico de los datos propiamente por medio de un muestreo y programas computacionales.

1.2 Objetivos

El objetivo principal y general es utilizar técnicas de muestreo, ordenamiento e interpretación de datos para explorar eficientemente la base de datos `hessi.solar.flare.2002to2016` cuyo tamaño no hace viable la rápida exploración de toda la población.

Además se busca extraer información suficiente para realiar conclusiones particulares sobre el fenómeno de fulguraciones solares que esta base representa. Esto se logrará mediante una exploración espacio-temporal de las energías, sus duraciones, categorías en las que podemos clasificarlas, etc.

Enumeramos a continuación la lista consisa de objetivos particulares

1. Dada una jerarquía de rangos de energía (que clasifica la naturaleza de la fulguración), encontrar las proporciones que hay de cada una de ellas.

2. Encontrar la duración media de las fulguraciones solares y encontrar si hay una correlación con la energía que producen. Es decir, responder a la pregunta:

   ¿Las fulguraciones de altas energías duran más (o menos) que las de más bajas energías?

3. Comparar la efectividad entre las técnicas de muestreo por estratificación de la muestra versus la sistematización de la muestra en base al orden natural de duraciones.

Parte II

MARCO TEÓRICO

"Si desea describir la verdad,
deje la elegancia al sastre."

-Ludwig Boltzmann

2.1 El experimento

2.1.1 Descripción de la base de datos

La base de datos seleccionada tiene el nombre `hessi.solar.flare.2002to2016.csv` en alegoría a ser muestras carcterísticas de fulguraciones solares (solar flares) tomadas en observaciones que duraron desde el año 2002 hasta el 2016 por el proyecto "Reuven Ramaty High Energy Solar Spectroscopic Imager"

Este es llamado por sus siglas en inglés, RHESSI, originalmente llamado sólo "High Energy Solar Spectroscopic Imager" o HESSI. Fue renombrado en honor a Reuven Ramaty, pionero de la física solar, y la astronomía de rayos-gamma.), un satélite observatorio de fulguraciones solares de la NASA.

Su lanzamiento fue parte de la sexta misión del programa "Small Explorer", seleccionado en Octubre de 1997 y lanzado en el 5 de Febrero del 2002. Su misión principal fue el explorar la física de la aceleración de partículas y energías liberadas durante las fulguraciones solares.

la base de datos tiene las siguientes entradas:

1. `flare:` Es un código numérico identificador de la fulguración (flare) observada. Sencillamente representa una etiqueta y no es de mayor relevancia para parámetros estadísticos. Se obviara o retirará a conveniencia.

2. `start.date:` Etiqueta temporal en formato año-mes-día (aa-mm-dd) que especifica cuándo comenzó la fulguración.

3. `start.time:` Etiqueta temporal en formato hora-minuto-segundo (hh-mm-ss) que especifica cuándo comenzó la fulguración en el día particular etiquetada por `start.date`.

4. `peak:` Hace referencia al tiempo (igualmente en formato hora-minuto-segundo (hh-mm-ss) como la etiqueta anterior) en que ocurre el máximo valor de flujo de rayos-x irradiado por la actividad solar en dicho punto observado.

   Este valor es conocido como el **pico** (peak) y es encontrado al observar el punto en cuestión por determinado lapso (usualmente el lapso completo de la fulguración) y obteniendo el máximo de todos dichos valores.

5. `end:` Hace referencia al tiempo (igualmente en formato hora-minuto-segundo (hh-mm-ss) como la etiqueta anterior) en que la fulguración acaba.[1]

6. `durtion.s:` Una etiqueta que describe el tiempo total de duración de la fulguración. Es obtenible desde las dos etiquetas previas `start.time` y end, sin embargo, es de mucho mayor utilidad para el análisis estadístico el tenerlo como un parámetro numérico a parte.

7. `peak.c.s:` Conteo de picos por segundo detectados en el rango de 6-12 KeV, promediado a través de todos los colimadores activos, incluyendo la actividad de fondo.

8. `total.counts:` Conteo de energías en el rango de 6-12 KeV integrado a traves de todas las fulguraciones solares, sumados a través de todas los colimadores, incluyendo la actividad de fondo.

[1] Los límites de cuándo comienza y termina una fulguración solar se determina por los niveles de radiación emitidos sobre la usual radiación promedio del sol. Este incremento en radiación es causado por la repentina aceleración de partículas cargadas desde el interior del sol, por ende, el lapso de la fulguración es equivalentemente descrito en términos del lapso en que dichas partículas son aceleradas.

9. `energy.kev`: La más alta banda (rango) de energía en que se observa la radiación emitida durante la fulguración. Medida en kilo electron volteo[2].

10. `x.pos.asec`: Valor de la coordenada x medida desde un marco de referencia con centro en un punto acordado de ser llamado "centro del sol" y con ejes orientados en direcciones constantes a través de todas las mediciones. Las unidades están en arcsec (arcosegundo, o segundo sexagesimal).

    Nota para las unidades: Hay 360° en un círculo, 60 arcominutos en 1°, y 60 arcosegundos en 1 arcominuto. Esto quiere decir que 1 arcosegundo es una 1/3600 parte de un grado. Desde la distancia promedio en la que yace la tierra respecto al sol (y en la que se mantienen las mediciones), 1 arcosegundo equivale a 725 kilómetros. Como ilustración, el diámetro de la tierra mediría 17.5 arcosegundos; usualmente las fulguraciones solares y manchas solares son mucho más grandes que eso.

11. `y.pos.asec`: Valor de la coordenada y medida desde un marco de referencia con centro en un punto acordado de ser llamado "centro del sol" y con ejes orientados en direcciones constantes a través de todas las mediciones. Las unidades están en arcsec (arcosegundo, o segundo sexagesimal).

12. `radial`: Valor del radio respecto al corigen medida desde un marco de referencia con centro en un punto acordado de ser llamado "centro del sol" y con ejes orientados en direcciones constantes a través de todas las mediciones. Las unidades están en arcsec (arcosegundo, o segundo sexagesimal).

13. `active.region.ar`: Un identificador para la región activa más cercana a la particular localidad donde el dato fue recogido. Se le define como **región activa** a cierta región espacial en nuestro marco de referencia (y posteriormente etiquetada y rastreada independiente a dicho marco de referencia coordenado) una vez que se han observado históricamente una actividad frecuente de fulguraciones solares durante un lapso considerable.

    Cabe destacar además que estas regiones activas han sido observadas desde varios estudios previos o colaterales, además de observaciones indirectas por otros medios.

14. `flags`: Códigos para la descripción de calidad del dato. Estos pueden manifestar que el dato es ruido probablemente proveniente de otra fuente distinta al sol, o bien no tener una alta certeza de ser de naturaleza solar. Además, algunos tienen una medida de calidad enumerada en una escala del 1 al 11.

Notas:

- Solamente eventos con posición no-cero y rangos de energía diferentes al de 3-6 keV son confirmados como de fuente solar.

- Eventos que no tienen una posición y se muestran principalmente sólo en los detectores más en frente, pero no pudieron ser procesados como imagen, son etiquetados con la advertencia (flag) "PS" (Possibly solar).

[2]El electronvoltio (símbolo eV) es una unidad de energía que representa la variación de energía cinética que experimenta un electrón al moverse desde un punto de potencial V_a hasta un punto de potencial V_b cuando la diferencia $V_{ab} = V_b - V_a = 1$ V, es decir, cuando la diferencia de potencial del campo eléctrico es de 1 voltio.

Su valor se determina de forma experimental. El valor dado por CODATA 2014 es 1.602 176 620 8 $\times 10^{-19}$ J (incertidumbre relativa: 6.1 $\times 10^{-9}$ J; incertidumbre porcentual 6.1 $\times 10^{-7}$ %) obteniéndose este valor de multiplicar la carga fundamental por la unidad de potencial eléctrico (V).

- Eventos qe no tienen una posición válida y son confirmados de no venir de una fuente solar son etiquetados con la advertencia "NS" (non-solar).

Podemos de hecho hacer una lista completa de las posibles advertencias (flags) que pueden estar en las últimas columnas de nuestra base de datos:

- a0 - En un atenuador en el estado 0 (Ninguno) en algún momento durante la fulguración.
- a1 - En un atenuador en el estado 1 (Delgado) en algún momento durante la fulguración.
- a2 - En un atenuador en el estado 2 (Grueso) en algún momento durante la fulguración.
- a3 - En un atenuador en el estado 3 (Ambos) en algún momento durante la fulguración.
- An - Estado del atenuador (0, 1, 2, 3) durante la medición del pico.
- DF - Cuentas de los segmentos frontales fueron perdidas en algúm momento durante.
- DR - Cuentas de los segmentos laterales fueron perdidas en algúm momento durante.
- ED - Hubo un eclipse (noche) en algún momento durante la fulguración.
- EE - La fulguración comenzó durante un eclipse (noche).
- ES - La fulguración terminó durante un eclipse (noche).
- FE - La fulguración continuaba cuando terminó el archivo
- FR - En modo "fast rate" (frecuencia de medición aumentada).
- FS - La fulguración estaba ya comenzada al inico del archivo.
- GD - Brecha en los datos durante la fulguración (razones variadas),
- GE - La fulguración terminó durante la brecha.
- GS - La fulguración inició durante la brecha.
- MR - La nave yacía en altas altitudes durante la fulguración.
- NS - Evento no solar
- PE - Evento de partículas (hubieron partículas detectables presente).
- PS - Posible fulguración solar. Detectado por los sensores frontales pero sin posición..
- Pn - Calidad de posición: P0 = Posición NO válida, P1 = Posición válida.
- Qn - Calidad del dato: Q0 = Mejor calidad, Q11 = Peor calidad.
- SD - La nave estuvo en la SSA (south atlantic anomaly) durante la fulguración.
- SE - La fulguración acabó cuando la nave estaba en la SSA (south atlantic anomaly).
- SS - La fulguración comenzó cuando la nave estaba en la SSA (south atlantic anomaly)

2.1.2 Resumen Fenomenológico

El modelo principal que se tuvo en mente para este muestreo fue la teoría **magnetohidrodinámica**, particularmente el fenómeno de resistividad simple. Las fulguraciones solares se explican ocurrir cuando ocurre reconección magnética entre las fronteras de plasmas altamente conductores que interactuan.

La reconección magnética consiste en un reordenamiento de las lineas de flujo magnético en que su topología cambia (surgen más componentes conexas y/o diferente número de canales en que la energía puede propagarse) y resulta en una conversión de la energía magnética almacenada por el campo hacia energía cinética, térmica y aceleración de las partículas mismas del plasma.

1. Sunspot/mancha solar: Regiones de la fotósfera del sol con temperaturas más bajas que la media, y por ende se muestran más opacas. Son debidas a campos magnétidos fuertes y densos generados por plasma circulante que por veces surge por la fotósfera debido a su entrelazamiento.

   Los bajos de temperatura causados por ello son alrededor de 1000 K en la región del sunspot. A tal región se le llama umbra y es rodeada por un borde delgado llamado penumbra. Su extensión tiene un rango desde "pequeños" (del tamaño más o menos de la tierra" hasta los grandes que pueden ser la mitad de la superficie.

2. Solar flare/fulguración solar: Erupciones violentas en la cromósfera del sol, ocasionadas por intensa actividad magnética. Durante tal erupción, los brillos suben alrededor de 1000 km sobre la cromósfera y la temperatura del plasma rápidamente incrementa a 20 millones de grados. Estos además liberan alrededor de 10^{25} Joules de energía (comparrable a unos cuantos millones de erupciones en la tierra).

3. La frecuencia de los sunspot y solar flares están fuertemente correlacionadas. Además son importantes de predecir/entender sus patrones de ocurrencia debido a que afectan la atmósfera terrestre electricamente y por ende interfiere con señales de radio. Fenómenos como las auroras borealis y australis son también causados por partículas energéticas que son inyectadas al planeta tras un solar flare.

NOTAS:

- **Reconección magnética**: Es un proceso en los plasmas conductores en el cual la topología del campo magnético es reordenada y la energía magnética convertida en cinética, térmica y aceleración de las partículas mismas. Ocurre (según magnedohidrodinámica) por resistividad eléctrica cerca de la capa fronteriza entre plasmas que se opone a las corrientes magnéticas **J** que describen las ecuaciones de Maxwell:

$$\nabla \times \mathbf{B} = \mu \mathbf{J} + \mu\epsilon \frac{\partial \mathbf{E}}{\partial t}.$$

- **Fotósfera**: Cascarón exterior de una estrella en que la luz es irradiada. En el caso del sol, sualmente entre los 4500 y 6000 K de temperatura (con temperatura efectiva de 5777 K. Es 100 kilómetros de ancho y con una densidad alrededor de 10^{-3} a 10^{-6} kg/m^3.

- **Cromósfera**: Está por encima de la fotósfera y tiene una densidad sólo de 10^{-4} de ella. Esto la hace usualmente invisible (debido al brillo de la fotósfera detrás) y solo visible durante eclipses totales con un color rojizo o rosas.

  La temperatura, a pesar de yacer 2000 km por encima de la fotósfera, es mucho mayor, llegando hasta los 25000 K. Esto es posiblemente explicado por reconección magnética.

2.1.3 Resumen de Clasificaciones de fulguraciones solares

Es importante notar que existen sistemas de clasificaciónes para las fulguraciones solares, manchas solares y sus acumulaciones. Esto permite la mejor determinación de conceptos como "regiones activas" y "regiones históricamente complejas".

Además, podemos con mayor facilidad (a costo de menor detalle cuantificable) caracterizar la evolución temporal de dichas regiones y verifcar así las hipótesis y modelos que tenemos para física de altas energías y físicaa de plasmas.

Estas clasificaciones son:

1. Valores-Z: (Clasificación modificada de Zurich para manchas solares).
    - A - Mancha solar pequeña, unipolar y aislada. Representa ya sea la etpa final o formativa de la evolución estelar.
    - B - Grupo de manchas solares bipolares sin ninguna penumbra.
    - C - Grupo de manchas solares bipolares con penumba en uno de los lados.
    - D - Grupo de manchas solares bipolares. Ambos extremos deben tener penumbra. Extensión longitudinal que no excede los 10 grados.
    - E - Grupo de manchas solares bipolares. Ambos extremos deben tener penumbra. Extensión longitudinal que sí excede los 10 grados pero no los 15.
    - F - Grupo enlongado de manchas solares bipolares. Ambos extremos deben tener penumbra. Extensión longitudinal que sí excede los 15 grados.
    - H - Grupo de manchas solares unipolar con penumbra.
2. Valores-p:
    - x - **Sin** penumbra (Clasificación de clase es A o B).
    - r - Penumba **rudimentaria** parcialmente rodeada de una mancha más grande. Suele ser una penumbra incompleta y granular en lugar de filamental. Es más brillante que una penumbra madura y se extiende tan poco como 3 arcsec de la umbra puntual. Puede que esté, ya sea en etapa de formación o de disolución
    - s - Penumbra pequeña **simétrica**. La mancha más grande tiene una penumbra oscura, filamental y madura en forma de círculo o elipse con poca irregularidad. El diámetro de norte-sur de la penumbra es menor o igual a 2.5 grados.
    - a - Penumbra pequeña **ásimetrica**. La penumbra de la mancha más grande tiene irregulares fuertes y las múltiples umbras del interior están separadas. El diámetro norte-sur es menor o igual a 2.5 grados
    - h - Grande y simétrica (similar a la clasificación H en Zurich). Tiene la misma estructura que el tipo "s" pero el diámetro de su penumbra puede ser mayor a 2.5 grados.
    - k - Grande y simétrica (similar a la clasificación H en Zurich). Tiene la misma estructura que el tipo "a" pero el diámetro de su penumbra puede ser mayor a 2.5 grados.

Parte III

TRATAMIENTO DE LOS DATOS

"El nitrógeno de nuestro ADN, el calcio de nuestros dientes, el hierro de nuestra sangre, el carbono de nuestras tartas de manzana se hicieron en los interiores de las estrellas en proceso de colapso.
¡Estamos hechos de sustancia estelar!."

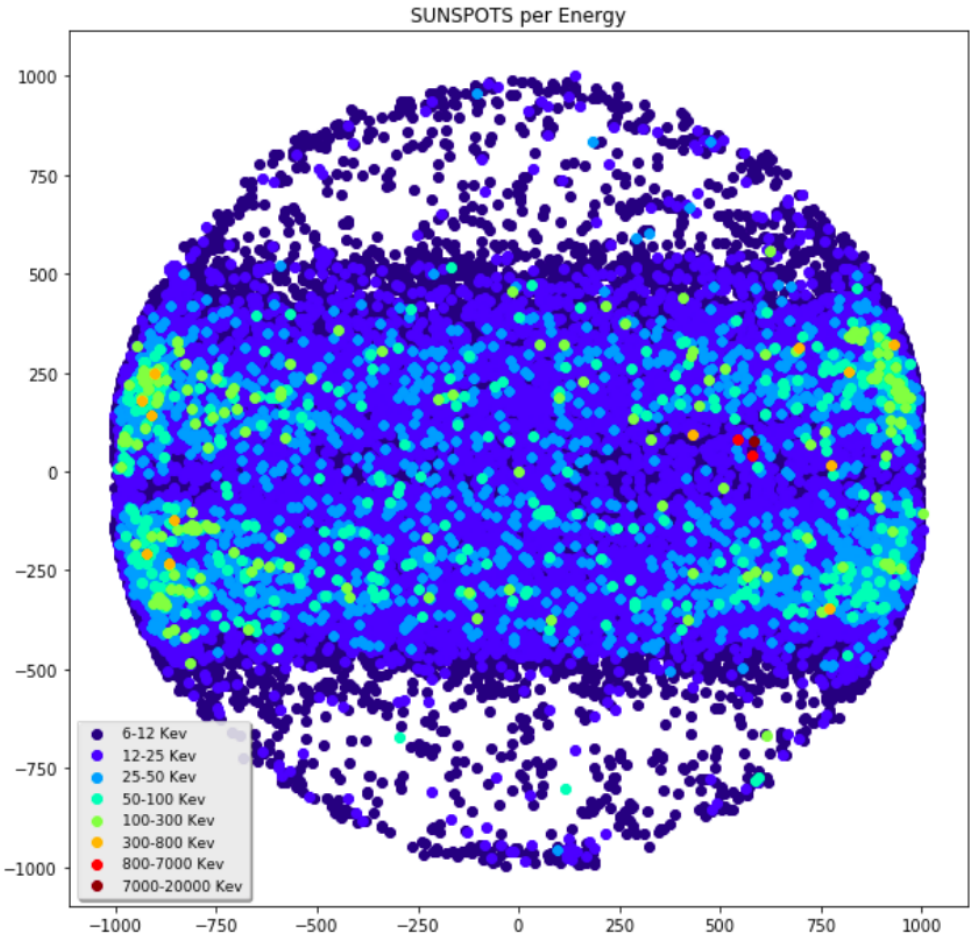
-Carl Sagan

3.1 Metodología de muestreo

3.1.1 Justificación de la metodología

Para comenzar realizaremos un análisis previo a los datos auxiliándonos del lenguaje de programación Python:

```
## Graficaremos las regiones solares segun energias
import matplotlib.pyplot as plt

# Codificamos los colores
colors = plt.cm.jet(np.linspace(0,1,len(CENERGY['energy.kev.i'].values)))

# Construimos los elementos de la figura
fig, ax = plt.subplots(figsize=(10,10))
# Recorremos todas las posiciones y sus rangos de energia
for i,irange in enumerate(CENERGY['energy.kev'].values):
    # Recolectamos los puntos de datos
    AUX = DATA1[DATA1['energy.kev']==irange][['x.pos.asec','y.pos.asec']]
    # Agregamos el punto al grafico
    plt.scatter(AUX['x.pos.asec'].values,AUX['y.pos.asec'].values,color=
                                        colors[i],label='%s Kev'%irange)
    ax.legend(loc='best',fontsize=9,shadow=True)
    # limpiamos memoria
    del(AUX)
# Titulo
plt.title('SUNSPOTS per Energy')
# Grafica
plt.show()
```



En la figura anterior se ilustra la distribución de energías a lo largo de la geometría real del sol. Para ello se ha utilizado las características de posición `x.pos.asec` y `y.pos.asec` para ubicar cada punto en relación a los demás, y utilizado la característica de energía `energy.kev` para determinar un código de colores según categoría de energías y así ilustrar la actividad de fulguraciones.

Esto nos provee con información importante para evitar concluir equivocadamente sobre ciertas medias que posteriormente calcularemos Por ejemplo, si pensásemos en buscar la posición vertical `y.pos.asec` media de la actividad de fulguraciones (con esperanzas de que eso nos indique el valor esperado de *dónde observar* la mayoría de fulguraciones):

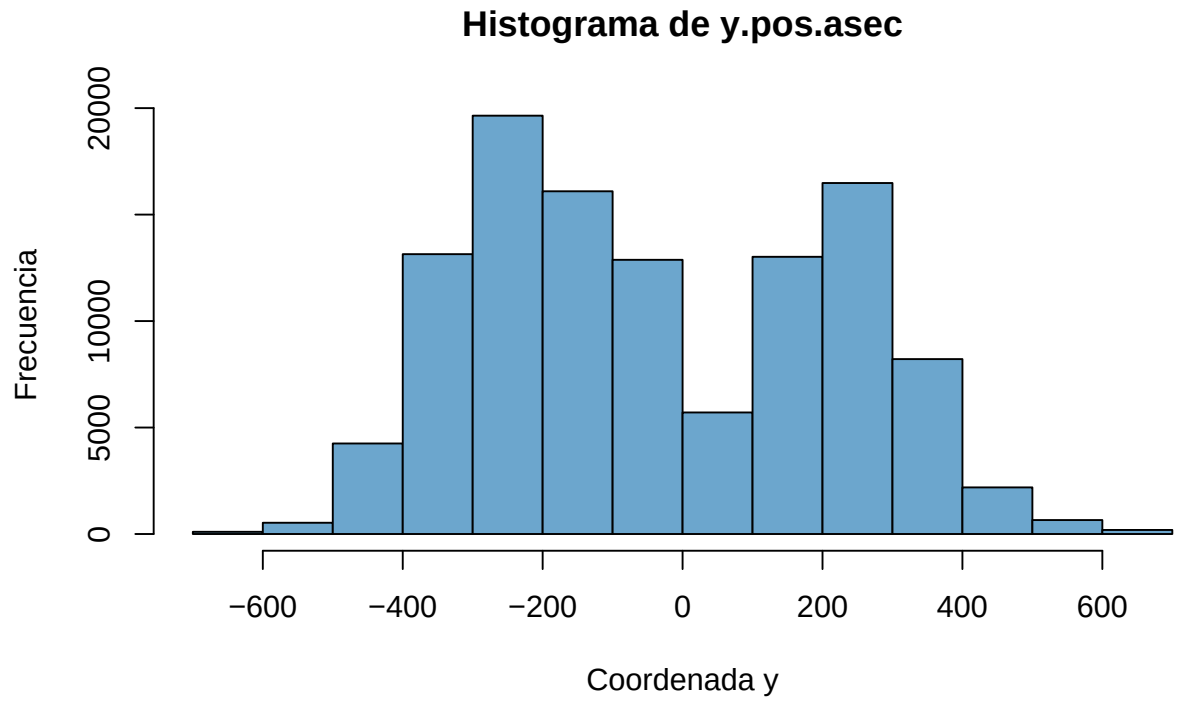
```
data <- read.csv("hessi.solar.flare.2002to2016.csv")
y_pos_bar <- mean(data$y.pos.asec)
y_pos_bar

## [1] -43.12981
```

Lo cual nos ubica muy cercanos al origen de nuestra figura anterior (note la escala utilizada en los ejes coordenados). Esto no nos provee una zona de *mayor actividad esperada* debido a que las reales zonas de mayor actividad son dos casi equidistantes del origen.

En otras palabras, la distribución de actividad es una distribución sesgada por sus momentos de orden superior a 2, lo que causa que la media no coincida con el valor máximo. Esto nos indica que debemos ser cuidadosos en asumir propiedades de nuestra distribución de datos

Para mayor ilustración, podemos graficar algunos de los datos *no tan extremos*[1] de la siguiente manera:

```
hist(data$y.pos.asec[abs(data$y.pos.asec)<700], xlab = "Coordenada y",
     ylab = "Frecuencia", main = "Histograma de y.pos.asec",
     col = "skyblue3")
```



[1] Con coordenada-y a no más de 700 de arcosegundos del origen.

Nuestros datos están ordenados cronológicamente en base a las características `start.date` y `start.time`. Evidentemente nuestra base, más que como una **muestra aleatoria**, está mejor descrita como una **serie de tiempo**, o al menos esa sería la inicial sospecha en base a la estructura de la toma de datos.

Para que pudiésemos considerarla más como una **muestra aleatoria** se deben cumplir las siguientes dos propiedades de nuestro conjunto $\{x_1, x_2, \ldots, x_N\}$:

- x_i y x_j deberían de ser independientes para todo $i \neq j$.
- Cada uno de los elementos de $\{x_1, x_2, \ldots, x_N\}$ deberían provenir de *la misma distribución.*

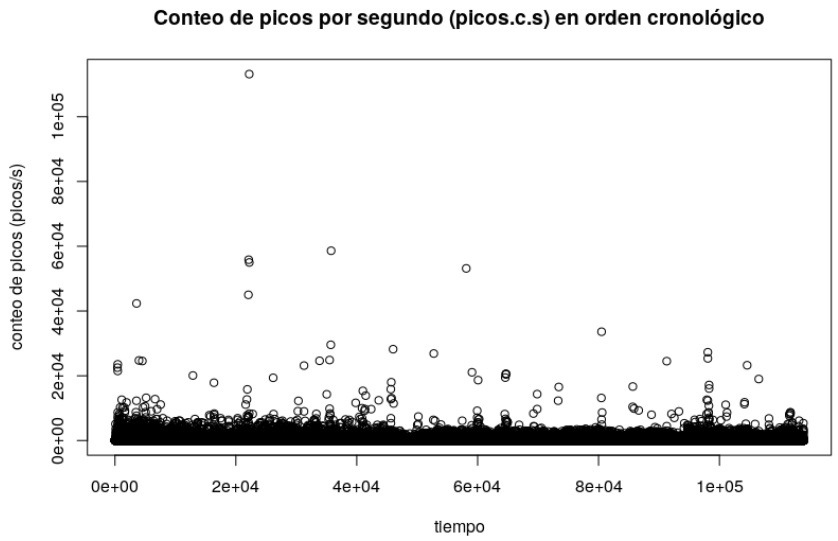
Dichas propiedades, trasladadas a las propiedades que nuestro fenómeno físico debería cumplir equivalentemente serían:

- Cada fulguración solar observada debería de tener todas sus propiedades enlistadas en la base de datos (niveles de energía, duración, conteo de picos, etc.) independiente de las propiedades de cualquier otra observación distinta.
- Cada fulguración solar debería de tener la misma (o suficientemente similar) causa fundamental, pues esto resultaría en que la *forma* de las fluctuaciones entre observaciones depende sólo de **desconocer pequeños cambios en variables incontrolables del fenómeno** pero no provienen de fenómenos fundamentalmente distintos en cuanto al sistema dinámico determinista que lo puede describir (utilizando en este caso leyes y restricciones físicas).

  Por ejemplo, el mero hecho que la colisión de plasmas que subyace la fulguración nunca se dará de la misma manera pues su volumen, forma y distribución interna de energías cambia, mas no significativamente para cambiar lo que nuestras predicciones teóricas dirían del fenómeno.

La primera de estas puede ser verificada buscando patrones periódicos en nuestros datos ordenados cronológicamente. Es decir, si fuese el caso que existiese una dependencia entre datos, encontraríamos que **siempre que cierto tipo de dato ocurriese, tendríamos una respuesta específica de los datos subsiguientes** y por ende crear cierta periodicidad local que debería ser posible observar en un gráfico.

Proseguimos entonces a observar nuestros datos graficados:



Dada la ausencia de patrones periódicos, podemos concluir con suficiente certeza inicial que no habrá una correlación apreciable entre nuestros datos y por ende considerar correcta nuestra primera asunción.

Para argumentar sobre el segundo punto, lo haremos más físico y fenomenológico. Consideremos las escalas de tiempo en que se dan las etapas de una estrella y, dado que nuestro sol actualmente es una estrella de tipo-G de la secuencia principal, aún quedan billones de años para que su composición y comportamiento energético cambie significativamente.

Esto quiere decir que, para las escalas de tiempo en que la base de datos fue tomada (del 2002 al 2016, equivaliendo a 14 años aproximadamente), el comportamiento de actividad energética del sol ha sido la misma. Esto garantiza, asumiendo además el correcto mantenimiento del equipo de medición y consideraciones de eventos astronómicos extraordinarios, que nuestros datos recibidos provienen de una distribución que se ha mantenido estacionaria a lo largo de todo el experimento.

3.1.2 Muestreo de proporciones de energías

Proseguimos entonces a nuestro primer objetivo: Categorizar las fulguraciones solares en niveles de energía en base a proporciones.

Proseguimos entonces a tomar una muestra aleatoria de tamaño 5000 de nuestra base de datos. Esto lo haremos mediante **muestro sistemático** en base un orden creciente en la duración de las fulguraciones. Esto garantizará que obtendremos fulguraciones de todas las duraciones en nuestra muestra (en proporción a cuántas hay de cada duración en toda la población) y así capturar cualquier correlación que pueda existir entre cuánto dura el evento y las energías emitidas.

Elegimos entonces nuestro k (tamaño de paso sistemático), número aleatorio R y extraemos la muestra:

```
N <- length(data$flare)
n <- 5000
k <- ceiling(N/n)
R <- trunc(runif(1,1,k))+1
indices <- seq(R, N, k)
muestra <- data[order(data$duration.s),][indices,]
```

Con ello veremos la proporción de fulguraciones en las siguientes categorías:

- 3-6 KeV
- 6-12 KeV
- 12-25 KeV
- 25-50 KeV
- 50-100 KeV
- 100-300 KeV
- 300-800 KeV
- 800-7000 KeV
- 7000-20000 KeV

Note que a medida las categorías van incluyendo energías más altas, también se vuelven más amplias. Esto es debido a que, a medida un evento de fulguración incrementa en energía, es menos probable que lo haga en "pequeños pasos"[2] y por ende, para tener categorías que sigan siendo útiles, se amplían acorde a cuánto se espera teóricamente que sigan variando la distribución es posibles fulguraciones a dichos niveles de energía.

[2]Tener diferentes niveles de energía en el evento significa usualmente una naturaleza distinta en número y aceleración de partículas en el plasma que causó el evento.

Creamos una función para calcular las proporciones reales, las estimadas y sus intervalos de confianza para comparar:

```
proporciones <- function(rango){
  y <- ifelse(muestra$energy.kev == rango, 1, 0)
  p_hat <- mean(y)
  var_p <- (1-n/N)*(p_hat*(1-p_hat))/(n-1)
  LI <- p_hat - qnorm(0.975)*sqrt(var_p)
  LS <- p_hat + qnorm(0.975)*sqrt(var_p)

  y <- ifelse(data$energy.kev == rango, 1, 0)
  p <- mean(y)
  x <- c(p, p_hat, max(LI, 0), LS)
  return(x)
}
```

Además de una función que nos ayude a imprimir nuestro resultado como tabla:

```
Imprimir_prop <- function(){
  rangos <- c("3-6", "6-12", "12-25", "25-50", "50-100",
              "100-300", "300-800", "800-7000", "7000-20000")
  tabla <- matrix(data = 0, nrow = length(rangos), ncol = 4)
  i <- 1
  for(k in rangos){
    tabla[i,] <- as.numeric(proporciones(k))
    i <- i + 1
  }
  rownames(tabla) <- rangos
  library(xtable)
  A <- xtable(tabla, align = "ccccc", digits = c(0,5,5,5,5))
  names(A) <- c("P real", "P estimado", "LI", "LS")
  print(A, sanitize.text.function = function(x){x})
}
```

E imprimimos la tabla generada:

```
Imprimir_prop()
```

	P real	P estimado	LI	LS
3-6	0.05561	0.05390	0.04778	0.06002
6-12	0.75209	0.74041	0.72853	0.75230
12-25	0.16995	0.18187	0.17142	0.19233
25-50	0.01706	0.01897	0.01528	0.02267
50-100	0.00357	0.00363	0.00200	0.00526
100-300	0.00159	0.00081	0.00004	0.00158
300-800	0.00011	0.00040	0.00000	0.00095
800-7000	0.00002	0.00000	0.00000	0.00000
7000-20000	0.00001	0.00000	0.00000	0.00000

Algo que cabe resaltar, es que algunas de nuestras estimaciones son idénticamente cero (tanto el valor medio como los intervalos de confianza).

Esto no es erróneo, pues el valor real de la proporción P es sumamente cercano a cero[3] y al existir tan pocos datos, nuestra varianza $\widehat{\text{Var}}(\hat{P})$ es también muy poca y eso vuelve angosto el intervalo de confianza; en este caso, idénticamente cero para efectos prácticos.

Para formalmente reportar los intervalos de confianza, tenemos[4]:

- **3-6 KeV**
  IC del 95% de $\hat{P}$: $[4.78\%, 6\%]$
- **6-12 KeV**
  IC del 95% de $\hat{P}$: $[72.85\%, 75.23\%]$
- **12-25 KeV**
  IC del 95% de $\hat{P}$: $[17.14\%, 19.23\%]$
- **25-50 KeV**
  IC del 95% de $\hat{P}$: $[1.53\%, 2.27\%]$
- **50-100 KeV**
  IC del 95% de $\hat{P}$: $[0.2\%, 0.53\%]$
- **100-300 KeV**
  IC del 95% de $\hat{P}$: $[0\%, 0.16\%]$
- **300-800 KeV**
  IC del 95% de $\hat{P}$: $[0\%, 0.09\%]$
- **800-7000 KeV**
  IC del 95% de $\hat{P}$: $[0\%, 0\%]$
- **7000-20000 KeV**
  IC del 95% de $\hat{P}$: $[0\%, 0\%]$

Los intervalos de confianza se han denotado, con ligero abuso de notación, con cantidades porcentuales. Esto es para mejor ilustrar lo que veremos a continuación con un gráfico de tipo circular.

Además, las cantidades han sido redondeadas a solamente 2 cifras luego del punto decimal[5], en visión que nuestra muestra puede no ser lo suficientemente grande para aseverar mayor precisión.

Notamos que tenemos, incluso con la actual precisión, intervalos de confianza de longitud 0, esto sólo indica que su proporción es despreciable respecto a nuestra escala de precisión y serán agrupados entre ellos en una categoría llamada "$<$ 100 KeV".

Con el siguiente código realizamos el gráfico circular para ilustrar las proporciones de cada categoría de energías:

```
chart <- c(A$`P estimado`[1:4], 1-sum(A$`P estimado`[1:4]))
etiq <- paste(round(chart*100, 1), "%")
leyenda <- c(paste(rangos[1:4], "KeV"), "<50 Kev")
library("RColorBrewer")
colores <- brewer.pal(5, "BuPu")
pie(chart, main = "Proporción de fulguraciones por energías",
    labels = etiq, radius = 1, col = colores)
legend("bottomleft", leyenda, fill = colores)
```

[3]Incluso a pesar de que las categorías de altas energías son sumamente amplias; esto pues re-enfatiza lo anteriormente discutido.

[4]Los intervalo de confianza con longitud cero indican proporciones demasiado pequeñas para ser adecuadamente estimadas con el tamaño de muestra actual, sin embargo, no es una estimaciónimprecisa, pues sus valores reales sí yacen sumamente cercanos a 0.

[5]Lo correcto es hablar de cifras significativas, no obstante, en nuestro caso, las cifras decimales en porcentajes definen las cifras significativas.

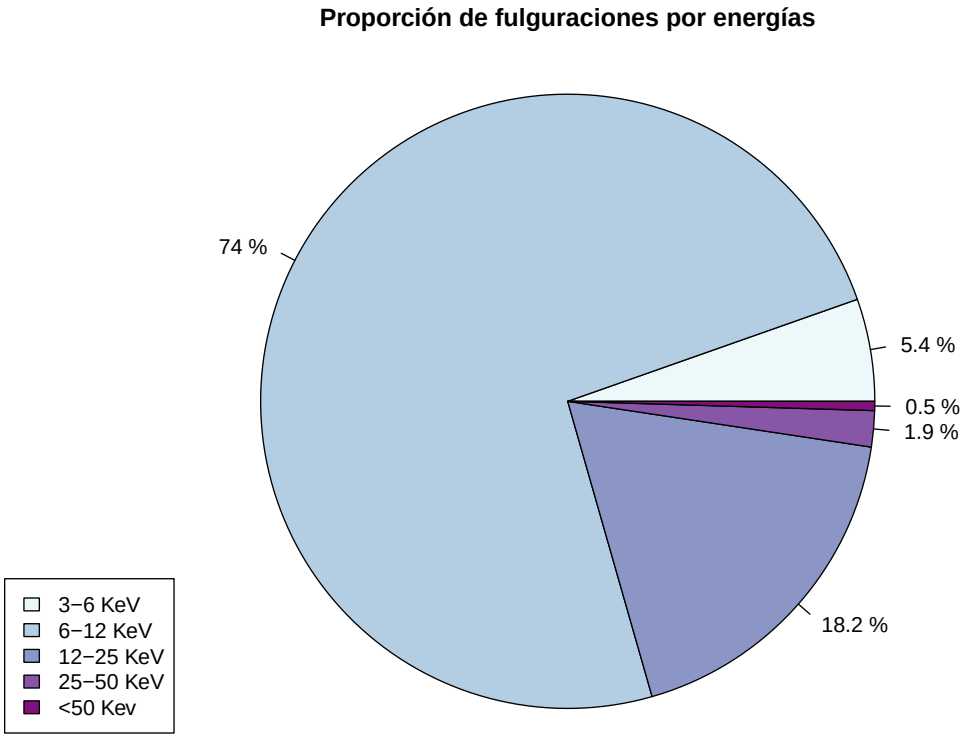


Figura 1: Gráfico circular de las proporciones porcentuales medias por energías.

3.1.3 Muestreo de duración media de una fulguración

Aprovechando la misma muestra anteriormente extraida podemos realizar una estimación de la duración media de fulguraciones. Tenemos la garantía que en nuestra muestra existe la diversidad adecuada de duraciones debido a que la toma sistemática de ella fue realizada acorde el orden de duraciones.

Prosguimos a calcular nuestro valor central:

```
y <- muestra$duration.s
y_bar <- mean(y)
y_bar

## [1] 493.1126
```

Contruimo el intervalo de confianza del 95%:

```
var_y_bar <- (1-n/N)*var(y)/n
LI <- y_bar - qnorm(0.975)*sqrt(var_y_bar)
LS <- y_bar + qnorm(0.975)*sqrt(var_y_bar)
```

Reportamos entonces el intérvalo de confianza como:

$$\bar{y} \in [481.4, 504.9] \text{ con un } 95\% \text{ de confianza.}$$

Podemos encontrar la media real de nuestra población (la base de datos entera) para comparar. Lo hacemos de la siguiente manera:

```
mu <- mean(data$duration.s)
mu

## [1] 493.0517
```

El error entre la media estimada y la real (error absoluto) y su error relativo son respectivamente:

$$\text{ERA}_{\bar{y}} = |493.0517 - 493.1126| = 0.0609$$

$$\text{ERR}_{\bar{y}} = \frac{|493.0517 - 493.1126|}{493.0517} \times 100\% = 0.01235\%$$

Ambos errores delatan una satisfactoria estimación de las proporciones. Sin embargo, puede que deseemos disminuir el tamaño muestral (como propondremos pronto) y aún tener la misma o mejor precisión.

Para mejorar nuestras estimaciones, busquemos utilizar la estructura de los datos para obtener mayor información. Ya hemos utilizado en esta ocasión el orden de las duraciones, sin embargo, podemos sospechar que hay una manera de clasificar (ya sea por conglomerados o por estratos) para capturar clases de agrupaciones de datos que nos provean una optimización en la predicción.

En sospecha de que la duración de cambia drásticamente entre escalas de energía, verificamos su distribución de la siguiente manera:

```
distr_eng <- tapply(data$duration.s, data$energy.kev, mean)
distr_eng

##    100-300      12-25      25-50        3-6    300-800     50-100
##  1278.2762   779.4290  1023.5885   482.5625  2078.0000  1224.1572
##       6-12 7000-20000   800-7000
##   411.7342   124.0000   422.0000
```

En efecto, notamos una gran diferencia entre las duraciones promedio de las fulguraciones en algunas categorías.

Por ejemplo, comparemos la categoría de de “12-25” con la de “300-800”. Además, esta relación no parece ser lineal, ni tan si quiera monótona, como se observa en la siguiente gráfica:

```
plot(distr_eng,
     xlab = "Energías (KeV)",
     ylab = "Tiempo (seg)",
     xaxt = 'n',
     main = "Distribución de duraciones según energías",
     col = "skyblue",
     pch = 19,
     cex = 2)
lines(distr_eng)
text(1:9, par("usr")[1] - 0.25, srt = 45,
     adj = 1,
     labels = paste(rangos, " "),
     xpd = TRUE)
```

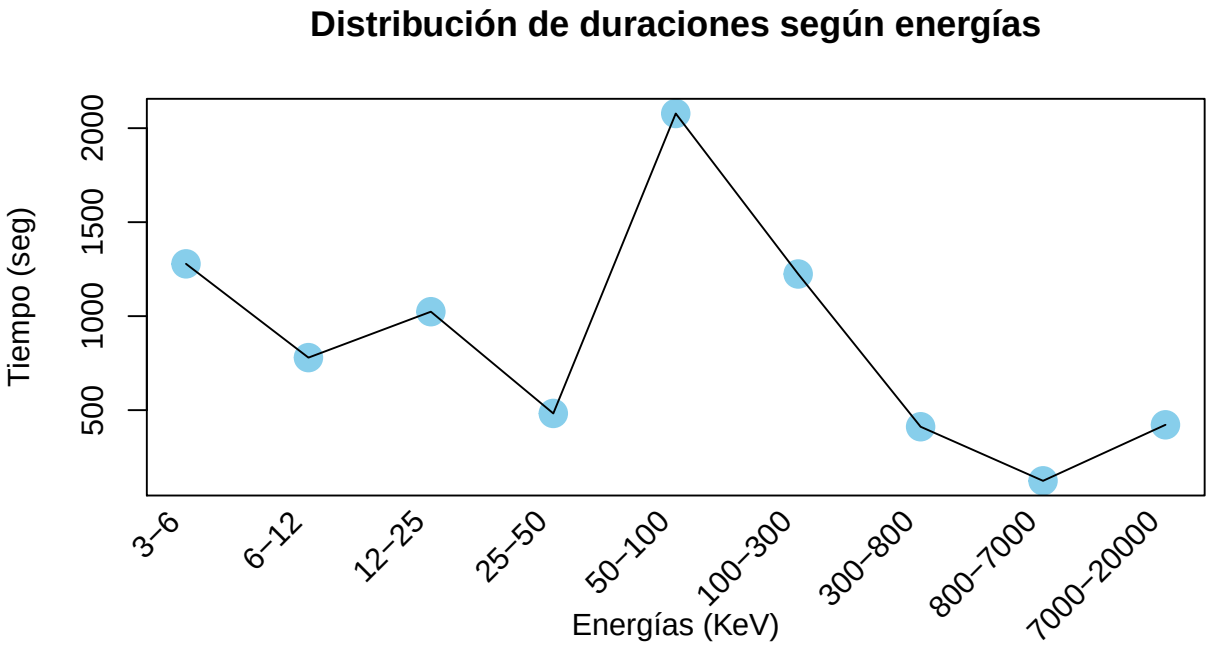


Figura 2: Gráfico de duraciones (en segundos) promedio que se da en los grupos de diferentes rangos de energía. Se puede apreciar la relación altamente no lineal ni fácil de predecir. Esto sugiere que utilizar estratos es una buena idea.

Sin embargo puede aún ser el caso que el muestreo estratificado no sea el correcto para nuestros datos si sucede que hay demasiada variación interna en los estratos propuestos (en este caso las categorías de energía). Eso se verificiará posterior a analizar los errores de este procesos de estratificación versus a los obtenidos anteriormente.

Se prosigue a realizar el muestreo estratificado y calcular la media:

```
excluidos <- c("300-800", "800-7000", "7000-20000")
data <- subset(data, !(data$energy.kev %in% excluidos))
n <- 5000
Nh <- table(data$energy.kev)
Nh <- as.numeric(Nh[c(2, 7, 3, 4, 6, 1, 5, 8, 9)][1:6])
nh <- stratasamp(n,Nh,type = 'prop')
nh <- as.numeric(nh[2,])
st <- sampling:::strata(droplevels(data),
                        stratanames="energy.kev",
                        size=nh,
                        method="srswr")
muestra <- getdata(data,st)
media <- stratamean(muestra$duration.s ,h = as.character(muestra$Stratum), Nh)
y_bar <- media$mean
```

El error entre la media estimada y la real (error absoluto) y su error relativo son respectivamente:

$$\text{ERA}_{\bar{y}} = |493.0517 - 495.5243| = 2.4725$$

$$\text{ERR}_{\bar{y}} = \frac{|493.0517 - 495.5243|}{493.0517} \times 100\% = 0.50148\%$$

Podemos notar que, incluso con el mismo tamaño de muestra, nuestro error es significativamente mayor al caso de muestreo sistemático utilizando muestreo aleatorio simple porsteriormente. Esto es un indicio de que los datos realmente no se prestan para ser agrupados de esta manera por medio de estratos y, como veremos, tampoco como conglomerados

Analicemos las desviaciones estándar de nuestros rangos de energía respecto a la desviación estándar global de todas las duraciones:

```
data <- read.csv("hessi.solar.flare.2002to2016.csv")
sd_global <- sd(data$duration.s)
sd_estrat <- as.numeric(tapply(data$duration.s, data$energy.kev, sd))
sd_global

## [1] 433.3893

sd_estrat

## [1]  820.79855  550.31010  700.21831  474.92083 1332.69324  811.85796
## [7]  336.18589         NA   42.42641
```

Y analizamos sus diferencias:

```
sd_global - sd_estrat

## [1] -387.40925 -116.92080 -266.82901  -41.53153 -899.30395 -378.46867
## [7]   97.20341         NA  390.96289
```

Los números negativos indican que la variable representativa de la desviación estándar global es superada por las desviaciones estándar **dentro de** los estratos propuestos. Esto indica que es mala idea muestrear los estratos en lugar de el muestreo global que hicimos previamente, ya que solo estamos introduciendo mayor error.

El valor denotado por el símbolo `NA` simboliza valores que no pudieron ser calculados. En este caso es debido a que, como podemos ver a continuación

```
tapply(data$duration.s, data$energy.kev, length)

##    100-300      12-25      25-50        3-6    300-800     50-100
##        181      19364       1944       6336         12        407
##       6-12 7000-20000   800-7000
##      85695          1          2
```

existe un estrato que no contiene suficientes observaciones para ser considerada una muestra aleatoria y tener bien definida la noción de desviación estándar.

Además, en general, no deberíamos de tomar tan en cuenta los valores de aquellos estratos con menos de 100 observaciones ya que probablemente no deberían ser tomadas en cualquier muestreo por agrupaciones (ya sean estratos, como hicimos anteriormente, o conglomerados).

En todo caso, la mayoría de estratos muestran comportamiento negativo, como indicador que la variación intraestratos no favorece el modelo de muestreo estratificado sobre le modelo de muestreo sistemático previamente utilizado.

3.2 Resultados de interpretaciones

En esta sección se enlistarán y discutirán los resultados puntuales de los objetivos propuestos al inicio del informe. Como parte de ello, haremos referencia a las secciones anteriores y datos específicos extraidos mediante los muestreos.

Recordamos cuales fueron nuestros objetivos:

1. Dada una jerarquía de rangos de energía (que clasifica la naturaleza de la fulguración), encontrar las proporciones que hay de cada una de ellas.

2. Encontrar la duración media de las fulguraciones solares y encontrar si hay una correlación con la energía que producen. Es decir, responder a la pregunta:

   ¿Las fulguraciones de altas energías duran más (o menos) que las de más bajas energías?

3. Comparar la efectividad entre las técnicas de muestreo por estratificación de la muestra versus la sistematización de la muestra en base al orden natural de duraciones.

En referencia al primer objetivo:

- Las proporciones del número de eventos de fulguraciones observados en función de los rangos de energía en que pertenece fue exitosamente estimado con mucha precisión aún con un tamaño de muestra moderadamente pequeño (especialmente en un sentido porcentual respecto al total de observaciones que realmente se realizaron).

  Los valores fueron suficientemente precisos para construir un gráfico circular y poder gráficamente entender la naturaleza de los eventos de fulguraciones. Esto es especialmente cierto cuando lo emparejamos con el primer gráfico realizado en `python` sobre la distribución espacial (en geometría real del sol) que nos muestra la distribución energética y ahora también conocemos la proporción de abundancia.

  Además, estas proporciones pudieron haber sido utilizadas en principio para calcular los tamaños de muestra N_h necesarios para estratificar posteriormente (en caso de realmente no tener el total de fulguraciones de cada rango energético).

En referencia al segundo objetivo:

- La duración media fue encontrada con un error relativo muy por debajo del 1%, lo cual es sumamente satisfactorio y un indicio de que la técnica de muestreo, en este caso el muestreo sistemático con base al orden natural de la variable de duracines de las fulguraciones, fue una técnica adecuada

  Esta la comparamos además con la técnica de muestreo estratificado mediante la agrupación de rangos de energía. Esta, sin embargo, resultó ser una técnica ineficaz para estimar la media, debido a variaciones infra-estrato superiores a las variaciones globales de la data.

- En tratamiento a la prgunta encerrada en caja, podemos decir que no existe una variación clara o fácil dedeterminar entre la energía en que se encuentra la fulguración solar y la duración de la misma.

En referencia al tercer objetivo

- En cuanto a la efectividad de la técnica de muestreo (particularmente la eficiencia de muestrear estratíficamente en contra del muestreo aleatorio simple tras sistematizar respecto a un orden) se discutió brevemente en el párrafo anterior. Sin embargo, algo que notar es que utilizamos una estratificación sugerida por la estructura misma de la toma de datos. Esto usualmente es una sabia elección, sin embargo, hay mucho detrás teórico que podríamos analizar respecto al fenómeno mismo para estratificar de una manera ás inteligente.

Parte IV

CONCLUSIONES Y SECCIÓN BIBLIOGRÁFICA

"La poesía está por todas partes, pero llevarla al papel es, por desgracia, más complicado que verla."

-Vincent van Gogh

4.1 Conclusiones

En breve conclusión tenemos que, en cuanto a estimar un parámetro numérico continuo como es la duración de las fulguraciones, el muestreo sistemático se ha prestado mucho mejor a ello. Esto probablemente por la alta distinción en comportamiento que existe entre rangos de energía y, además, las grandes diferencias en tamaños de población en los estratos.

Se concluye además que, si bien la media de la duración de la fulguración fue encontrada con mucha presición respecto a la media poblacional, esta no representa del todo un dato físico tan importante como lo son las medias respecto a cada estrato.

Esto vuelve nuestro análisis de la muestra estratificada de hecho más útil en la práctica de lo que fue la sistemática.

4.2 Bibliografía

- Titov, Vyacheslav S.; Hornig, Gunnar; Démoulin, Pascal (August 2002). "Theory of magnetic connectivity in the solar corona". Journal of Geophysical Research: Space Physics. 107 (A8)
- Sitio principal de la base de datos: https://www.kaggle.com/khsamaha/solar-flares-rhessi . Extraida abril, 2019.
- Sitio web oficial de del centro de vuelo espacial de Goddard en la NASA para el experiento RHESSI: https://hesperia.gsfc.nasa.gov/rhessi3/ . Revisado en abril 2019.
- Chaves, Julio (2015). Introduction to Nonimaging Optics, Second Edition. CRC Press. ISBN 978-1482206739.
- Lohr, S. L. (2009). Sampling: design and analysis. Nelson Education.