Análisis exploratorio de datos para llamaradas solares

Sebastian Forero Mazo, Santiago Ortiz Guarín y Alejandro Jimenez Garzon

Introducción

Procesos en una estrella:
Las estrellas son sistemas físicos en los cuales se presentan una serie de interacciones fundamentales que dependen de su composición interna. Estos procesos son normalmente conocidos como cadenas de fusión, en muchos aspectos estos fenómenos dieron nacimiento al universo tal como lo conocemos hoy y describen la dinámica principal detrás de la evolución de estos cuerpos celestes. En el caso del Sol, se tiene una estrella de tipo POP II. Estas estrellas se caracterizan primero por componerse de elementos livianos de la tabla periódica tales como el hidrógeno y el helio, además de producir elementos pesados. Las reacciones mas comunes dentro de este tipo de estrellas son las cadenas protón-protón.

¿Donde se presentan estas interacciones?
Los procesos nucleares no son homogéneos en todo el volumen de la estrella entonces se tiene un sistema complejo con un constante equilibrio entre la expansión del plasma de fusión y la fuerza gravitacional dirigida al núcleo de la estrella. Básicamente en la capa superficial se tienen los procesos de fusión de los iones presentes como si fuese una sopa hirviendo, el producto final de una cadena protón-protón más probable es el helio, el cual al ser más pesado que el resto de los iones en el plasma es atraído por la fuerza gravitacional del núcleo.

Erupciones solares, una liberación energética
El plasma de fusión tiende a acumular valores de energía en regiones activas de la superficie solar, las cuales liberan una acumulación de energía magnética en forma de erupciones, estas erupciones son captadas por el RHESSI (Ramaty High Energy Solar Spectroscopic Imager) como picos de intensidad de un espectro de longitudes de onda (de los rayos Gamma γ al espectro visible).

Base de datos
La base de datos de la misión contiene 116.143 eventos solares medidos entre Febrero de 2002 y Marzo de 2018. Las características listadas en la base de datos se explican a continuación:

- flare: Número que denota la erupción solar.
- start.date: Fecha de inicio de la erupión.
- start.time: Hora de inicio.
- peak: Hora a la que se presenta el máximo.
- end: Hora de finalización
- duration.s: Duración de la erupción.
- peak.c/s: Recuento de picos por segundo
- total.counts: Recuento total de picos
- energy.kev: Las erupciones se dan en cierto rango de energía (E_{min},E_{max}). (KeV)
- x.pos.asec: La posición X por segundo de arco
- y.pos.asec: La posición Y por segundo de arco
- radial: Unidad em segundo de arco
- active.region.ar: Región activa del Sol
- flag.i: $i = 1, 2, 3, 4, 5$ La base de datos contiene una serie de banderas que clasifican el evento solar según parámetros de caracterización explicados más adelante.

Las banderas pueden tomar diferentes valores que describen cierta condición o característica del evento.

- a0: In attenuator state 0 (None) sometime during flare
- a1: In attenuator state 1 (Thin) sometime during flare
- a2: In attenuator state 2 (Thick) sometime during flare
- a3: In attenuator state 3 (Both) sometime during flare
- An: Attenuator state (0=None, 1=Thin, 2=Thick, 3=Both) at peak of flare
- DF: Front segment counts were decimated sometime during flare
- DR: Rear segment counts were decimated sometime during flare
- ED: Spacecraft eclipse (night) sometime during flare
- EE: Flare ended in spacecraft eclipse (night)
- ES: Flare started in spacecraft eclipse (night)
- GD: Data gap during flare
- GE: Flare ended in data gap
- GS: Flare started in data gap
- NS: Non-solar event
- PE: Particle event: Particles are present
- PS: Possible Solar Flare; in front detectors, but no position
- Pn: Position Quality: P0 = Position is NOT valid, P1 = Position is valid
- Qn: Data Quality: Q0 = Highest Quality, Q11 = Lowest Quality
- SD: Spacecraft was in SAA sometime during flare
- SE: Flare ended when spacecraft was in SAA
- SS: Flare started when spacecraft was in SAA

Tratamiento de datos

Limpieza de datos preliminar

Se define un nuevo índice con un DataTime que contiene la información presente en las columnas start.date y start.time, luego se eliminan estas dos columnas junto con las columnas end y flare. end no es necesaria ya que se tiene la duración del evento, que es un tipo de información más útil a la hora de analizar los procesos y flare solo denota las llamaradas con cierta secuencia de números.

A partir del método isnull de pandas se determina que solo las tres últimas columnas del DataFrame, las cuales corresponden a flags, tienen datos nulos, estos datos fueros sustituidos por una cadena con un espacio para construir una nueva columna que contenga todas las banderas del evento en una sola celda y posteriormente se eliminan las columnas flag.i.

Dado que la columna energy.kev tiene asociado valores de tipo string, se utiliza la libreria re para separar los valores de energía mínimo y máximo, así mismo como la energía promedio de todo el evento.

Análisis de datos

Lo anterior es útil para hacer un análisis exploratorio para los diferentes rangos de energía en los que se dan las erupciones, un recuento de los eventos que se dieron en diferentes rangos de energía se muestra a continuación.

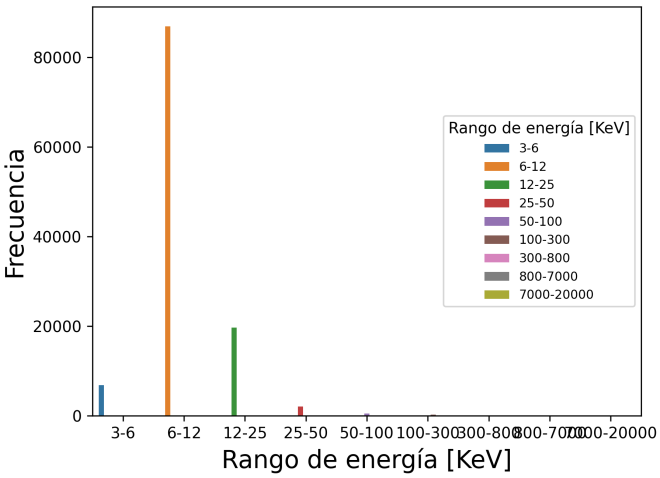

Figura 1: Resultados para el conteo de eventos con diferentes rangos de energía.

Hipótesis: Los datos cumplen una relación lineal entre las variables. Se puede pensar que algunas características del evento como el conteo de picos dependen linealmente (directamente proporcional) de la energía del evento.

Una forma de observar la relación entre las variables de la base de datos es vía una matriz de correlación.

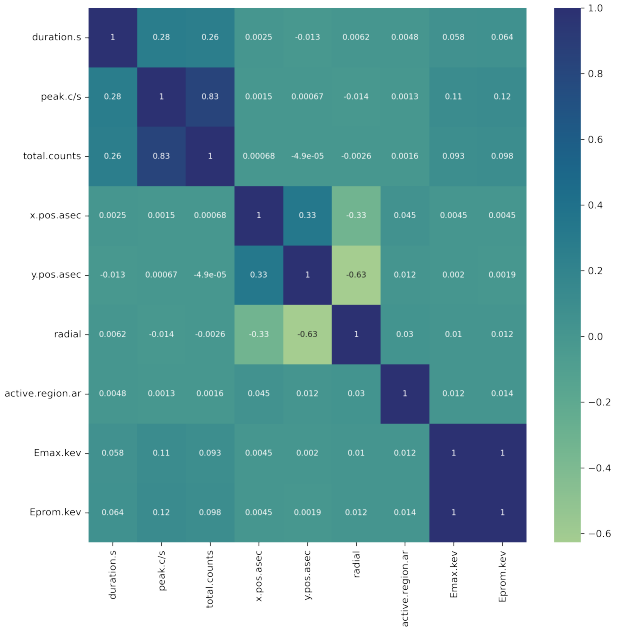

Figura 2: Matriz de correlación para los datos.

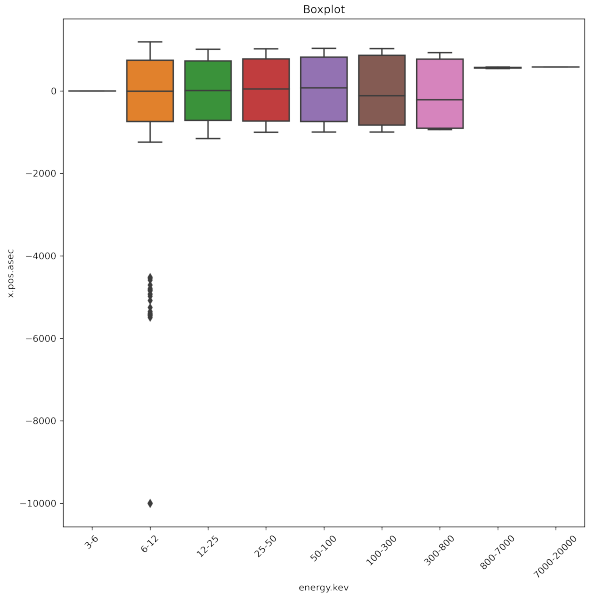

Figura 3: Diagrama de cajón y bigote para la posición x para diferentes rangos de energía.

De la figura 2 se observa que las variables no siguen una relación lineal entre ellas y por tanto puede pensarse que los datos siguen algún otro tipo de relación.

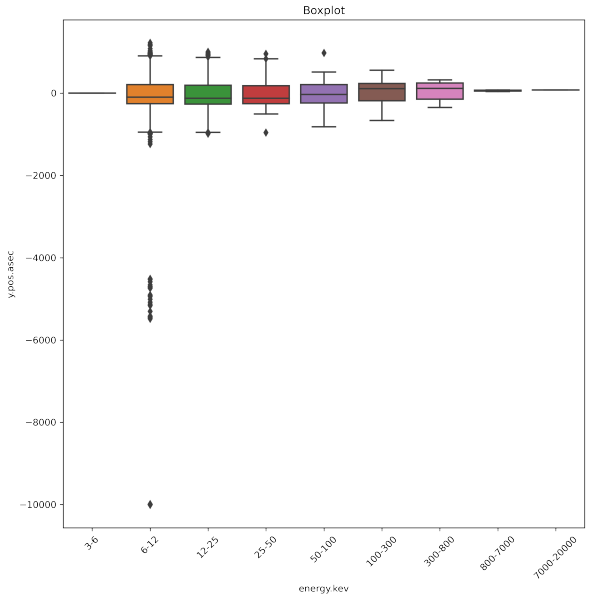

Figura 4: Diagrama de cajón y bigote para la posición y para diferentes rangos de energía.

Los diagramas de cajón y bigote son útiles a la hora de observar el comportamiento de la distribución de los datos, para el caso de las posiciones x,y para los diferentes rangos de energía se observa que no se siguen distribuciones simétricas y se tienen outlayers que no definen el comportamiento de los datos.

Con el objetivo de centrar el análisis de los eventos en eventos que se acerquen al comportamiento estadístico de los datos, se eliminan todos los datos aquellos que contengan las flags NS, P0 Y PE, luego que grafican nuevamente los diagramas de cajón y bigote para cada rango de energía.

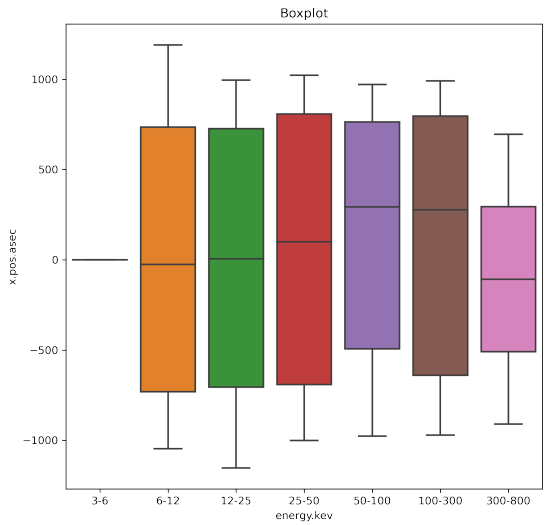

Figura 5: Diagrama de cajón y bigote para la posición x para los datos filtrados.

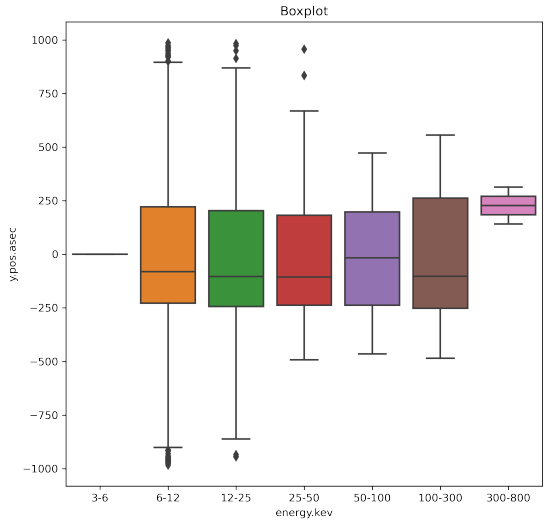

Figura 6: Diagrama de cajón y bigote para la posición y para los datos filtrados.

A partir de la función jointplot de Seaborn se contruye un gráfico de scatter para las posiciones x, y. Además de los respectivos histogramas que muestran la distribución para diferentes energías.

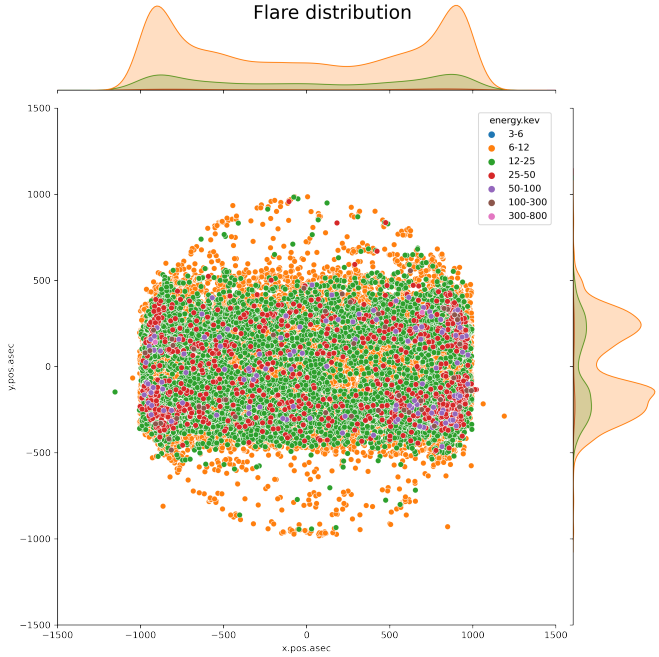

Figura 7: Gráfico de dispersión e histogramas para las posiciones y diferentes energías (datos diltados).

De donde se observa un mapa de eventos que tienen determinado rango de energía. Además se tiene que las llamaradas se producen con mayor frecuencia en la zona central del Sol y se tiene una mayor cantidad de eventos que corresponden a energías entre 12 y 25 keV.

Se desea analizar si la duración de los eventos sigue algún tipo de distribución simétrica mediante un gráfico de cajón y bigote.

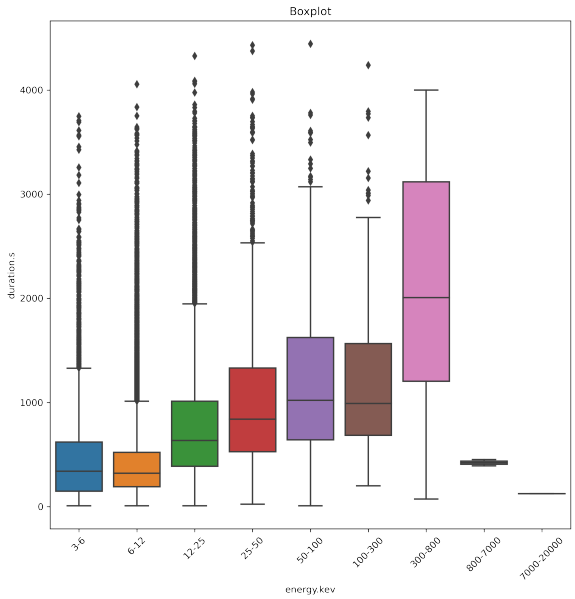

Figura 8: Diagrama de cajón y bigote para el logaritmo de la duración de los eventos.

Se tiene una distribución no simétrica con diferentes outlayers. AL graficar la función de distribución de probabilidad se muestra cómo se organizan los datos de duración.

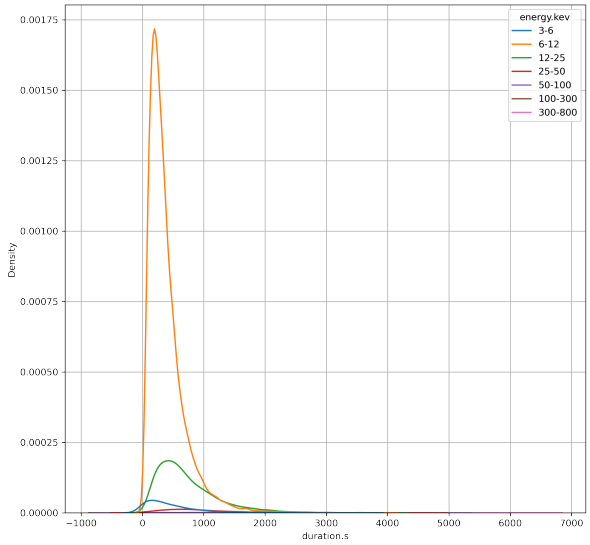

Figura 9: Distribución de la duración para diferentes rangos de energía.

Se verifica la distribución no simétrica de los datos, se analizará ahora la distribución para el logaritmo de la duración.

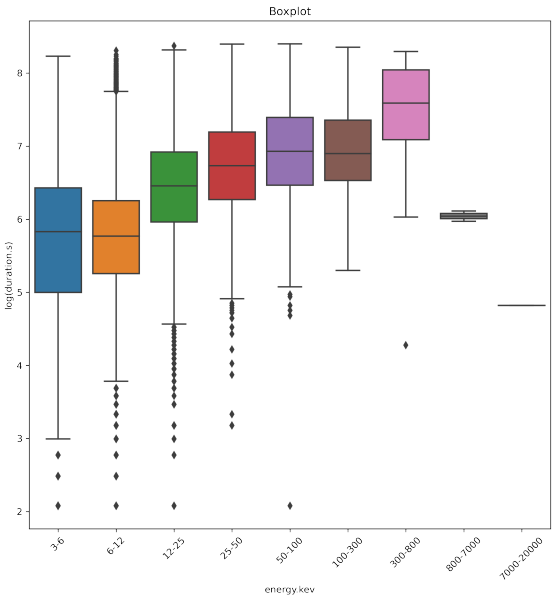

Figura 10: Diagrama de cajón y bigote para la duración de los eventos.

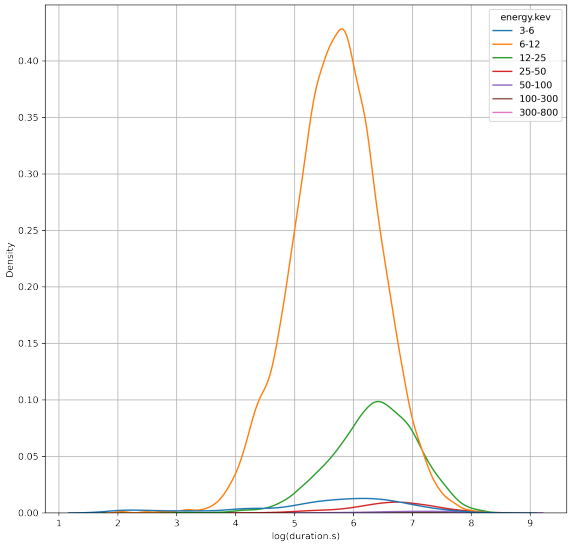

Figura 11: Distribución del logaritmo de la duración para diferentes rangos de energía.

Para las figuras 9 y 10 se realiza un ajuste de los datos de la duración en una escala de logaritmo natural, a la cual se le realizo un análisis con una distribución normal la cual se puede apreciar con detalle en el notebook adjunto. De la aproximación se obtuvo para el logaritmo de la duración la distribución observada en la figura 12.

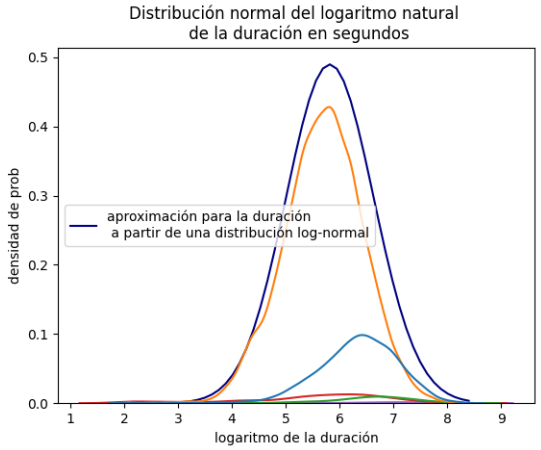

Figura 12: Aproximación del logaritmo de la duración para y logaritmo de la duración para diferentes rangos de energía.

También se trabajo para la mitad de los datos con respecto a la posición de arco en Y pues al tenerse en cuenta el punto 0 como el punto central del detector, al observar como se distribuyen los datos en la figura 7 se tiene que se podría encontrar alguna distribución para una mitad de los datos posicionales tanto de x como de y. en la siguiente figura se tiene la aproximación de la distribución normal de la variable y.pos.asec en escala logarítmica.

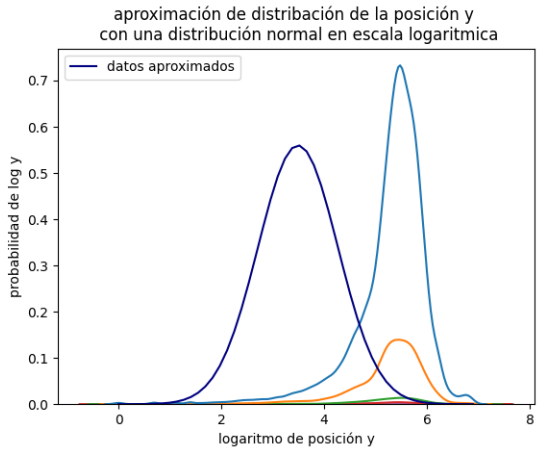

Figura 13: Aproximación del logaritmo de la duración para y logaritmo de la duración para diferentes rangos de energía.