Statistical Considerations of Solar Flare Data: Fall 2020 Report

Joshua D. Ingram *New College of Florida*

Solar flares are impulsive releases of energy that tend to occur in active regions located in the solar corona, occurring as a result of the reconnection of the sun's magnetic field lines. The energies of flares appear to follow a power-law distribution, but due to sensitivity limitations of satellites at low energies, flares are not detectable at the left end of the distribution and the power-law turns over. Given this limitation, we utilize the maximum product of spacings (MPS) method to simultaneously estimate the power-law and the energy range over which it occurs in our observed data. We utilize the GOES database covering solar cycles 23 and 24 in our analysis, reporting the power-law exponent and bounds for the total energy, peak flux, and duration for the aggregate, by-cycle, and by-year subsamples. Additionally, we find that the distribution of flare counts throughout the cycles are overdispersed and do not abide by the Poisson distribution's assumptions, reporting that the Negative Binomial distribution seems to be a much better fit for the count data.

Introduction

About every 11 years the sun completes a full *solar cycle* after its poles flip. This is a cycle of increasing and decreasing activity, with the most active phase of the cycle being known as the *solar maximum* and the least active known as the *solar minimum*. We typically characterize the phase of the cycle by the number of sunspots that occur, but this can also be accomplished by observing the number of solar flares occuring in the corona of the sun. *Figure 1* displays the solar cycle through a histogram of the frequency of solar flares observed by the GOES Satellites throughout solar cycles 23 and 24.

These solar flares are bursts of light and radiation that occur in the sun's active regions as a result of the reconnection of the sun's magnetic field lines. They are high energy events observable at all wavelengths between radio and γ-ray on the electromagnetic spectrum, where each flare differs in their magnitudes of total energy release, peak flux, and duration. There is even a trend in the intensity of flares, classified by their peak flux, throughout the solar cycle (see *Figure 2*). The lowest classification of flares are known as A-class flares, with their peak flux being between $1e^{-5}$ and $1e^{-4}$ $ergs/s/cm^2$ at earth. The least common but most intense flares are defined as X-class flares, with their peak flux being greater than 0.1 $ergs/s/cm^2$ at earth. See *Table 1* for a list of flare classifications.

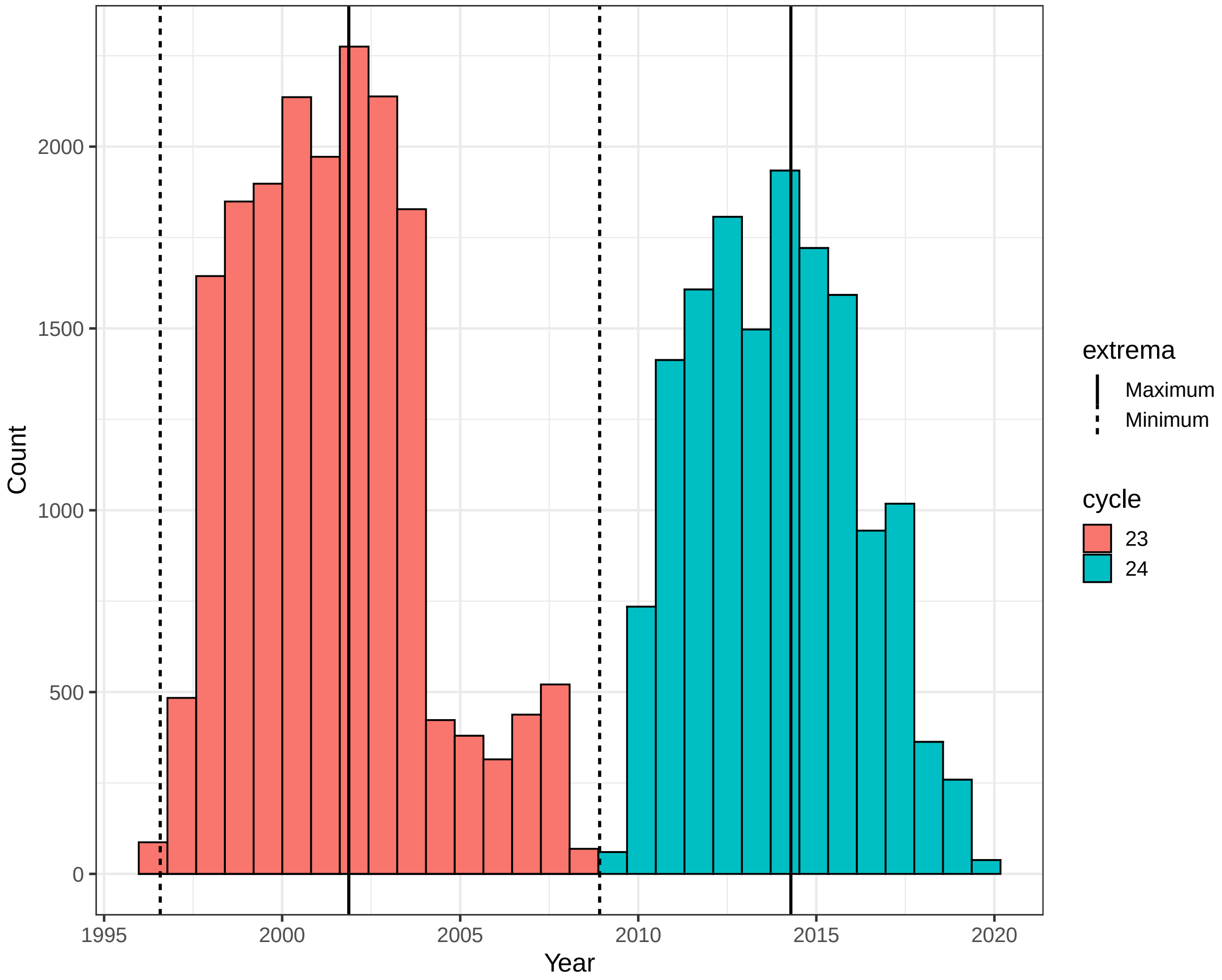

Figure 1: Distribution of Flare Counts

Table 1: Flare Classifications

Class	Peak Flux Range
A	1e-5 to 1e-4 ergs/s/cmˆ2 at earth
B	1e-4 to 1e-3 ergs/s/cmˆ2 at earth
C	1e-3 to 0.01 ergs/s/cmˆ2 at earth
M	0.01 to 0.1 ergs/s/cmˆ2 at earth

Class	Peak Flux Range
X	> 0.1 ergs/s/cm^2 at earth

As seen in the two figures above, the total number of flares occuring increase as the solar maximum is approached. Additionally, the number of flares in each class differ, with B and C-class flares being the most common. The A, M, and X-class flares are the least common, with their frequencies increasing near the solar maximum. The reason for these differing frequencies by class can be explained by the distribution that flare properties follow, which is the power-law distribution. However, under a power-law distribution we would expect that the low energy A-class flares would be the most frequent flares in our observations, followed by B-class flares and so on.

This disparity in the number of low energy flares observed and the number of low energy flares expected is caused by sensitivity limitations of satellities at low energies. As the intensity of a flare decreases, it is harder to detect these flares because of the weakening contrast between the regular "background" emission of the corona and the emission of the flare. Thus, the lower the intensity of a flare, the less likely it is that we detect that flare. This also results in another detectability problem caused by the nature of the solar cycle. That is, as the number of high energy flares increase it is even less likely that we detect lower energy flares. This can be seen in *Figure 2* where there is an unexpected dip in the number B-class flares near the solar maximum, where high intensity flares are most common.

The nature of the solar cycle, the distribution of flare properties, and the limitations of satellites leads us to several important questions in solar physics that can be answered by utilizing statistical metholdologies. In the following sections, we discuss how flare counts are distributed throughout the solar cycle, present the Maximum Product of Spacings method to address the detectability problem, and estimate the power-law distribution of several solar flare properties.

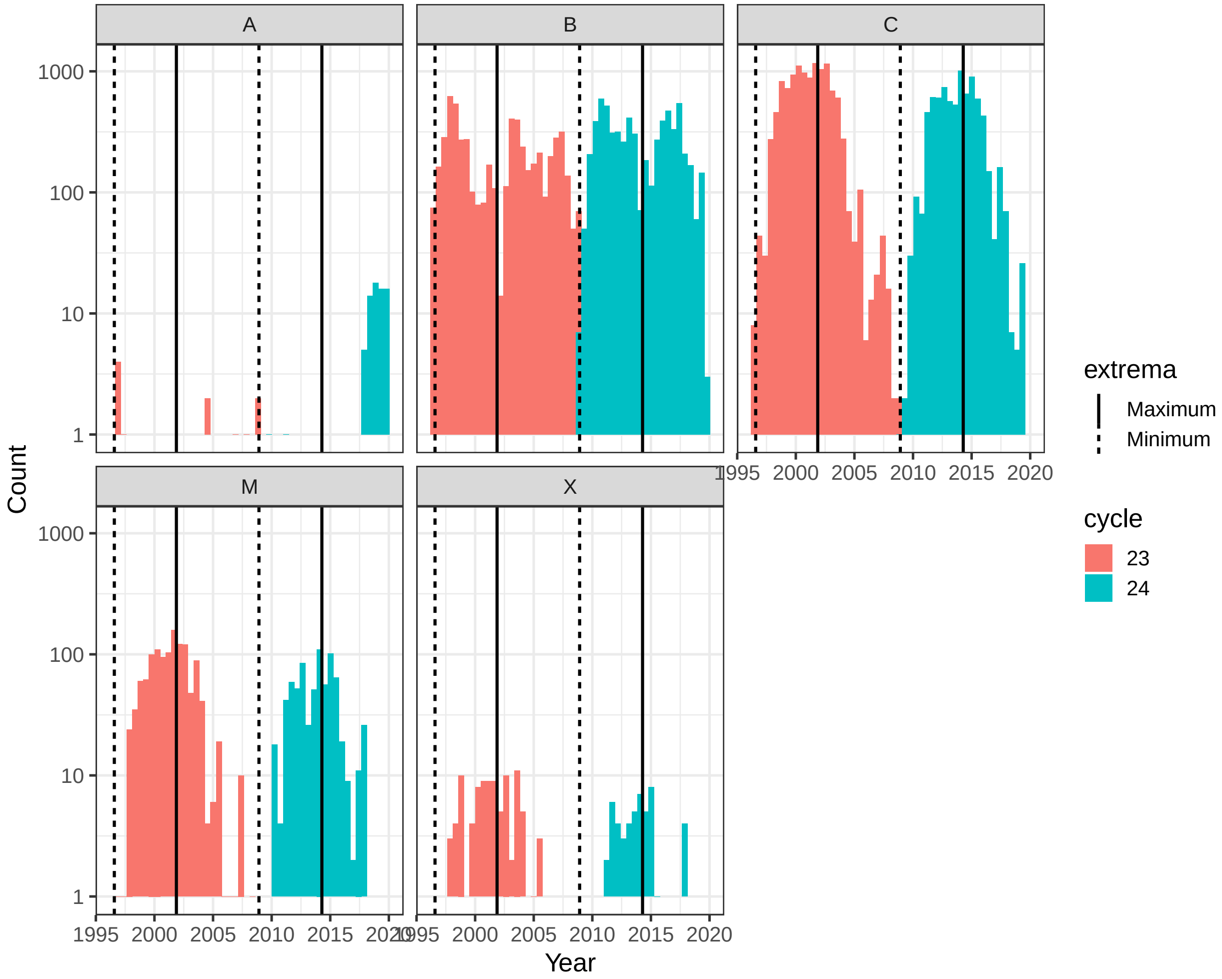

Figure 2: Distribution of Flare Counts by Class

Data

Observations

The data utilized for our analysis was collected from the Geostationary Operational Environmental Satellites (GOES) database. The GOES satellites are operated by NASA and the National Oceanic and Atmospheric Administration (NOAA), collecting data on the earth's atmosphere and on flares occuring on the sun. There are several versions of the GOES satellites, each equipped with a Solar X-Ray Imager (SXI), with seventeen of the satellites being active at some point after their launch. Our solar flare data has been collected by GOES-7 through GOES-16 satellites, with the earliest obervation occuring in July of 1996 and the latest occuring in December of 2019.

Each observation in our data is an individual flare, with *Table 2* containing the definitions and units of the measured properties on each of the observed flares.

Table 2: Variable Definitions

Variable	Definition	Units
Gevtnum	flare event number	
Garreg	AR region number, if known	
Gstart	flare start time	YYYY-MM-DDTHH:MM:SS
Gpeak	flare peak time	YYYY-MM-DDTHH:MM:SS
Gstop	flare stop time	YYYY-MM-DDTHH:MM:SS
Gduration	duration of the flare	sec
Gflrtotalenergy	flare total energy at sun	ergs
Glfxpeak	peak flux derived from GOES class at earth	ergs/s/cm^2
class	alphabetical GOES class assigned to flare peak	

Wrangling

Before continuing on with our analysis, it was necessary to properly clean and prepare our data. We removed any observations containing NA for the selected variables seen in *Table 2*. If any observation had an unreasonable value due to a recording error, such as a negative duration value, we outright removed these observations from the dataset. Before any processing of the GOES data, there were 38,114 observed flares. After filtering out bad observations and missing data, the processed data had 33,445 observations.

We would expect that each observed flare would have its own unique energy value, but rounding and recording limitations cause a "discretization" in our observations. This results in a clustering of flares at specific energy values. See *Figure 3* below for a visualization of this problem when the number of bins is set to 1000.

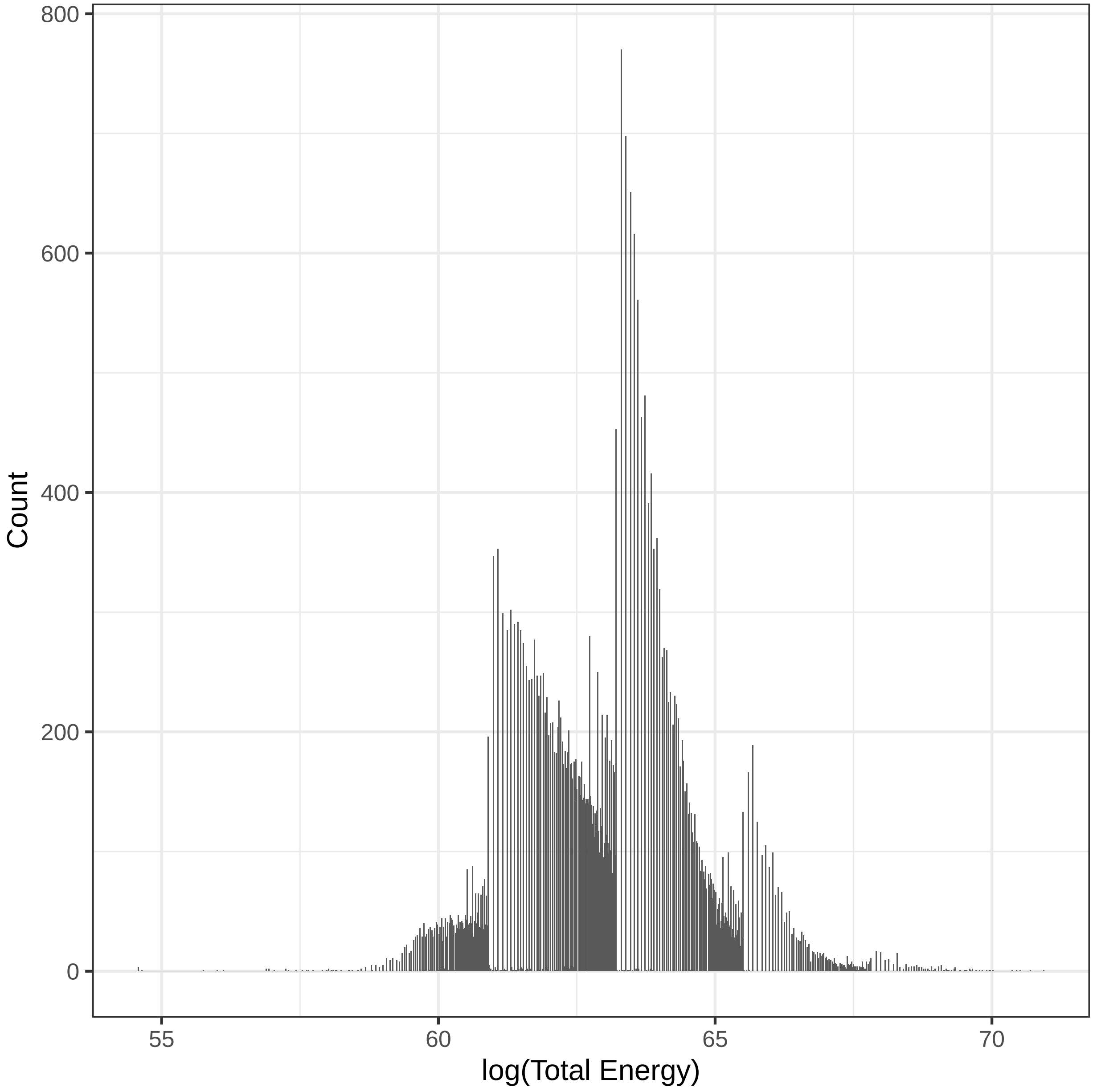

Figure 3: Discreteness casused by rounding

We found this had the potential to lead to less precise and biased estimates given by the maximum product of spacings algorithm. To address this, we jittered the filtered data before fitting our models to allow for more unique values. We accomplished this by uniformly distributing the observations over a small neighborhood of the discrete energy value. An example of the jittered distribution can be seen in *Figure 4*.

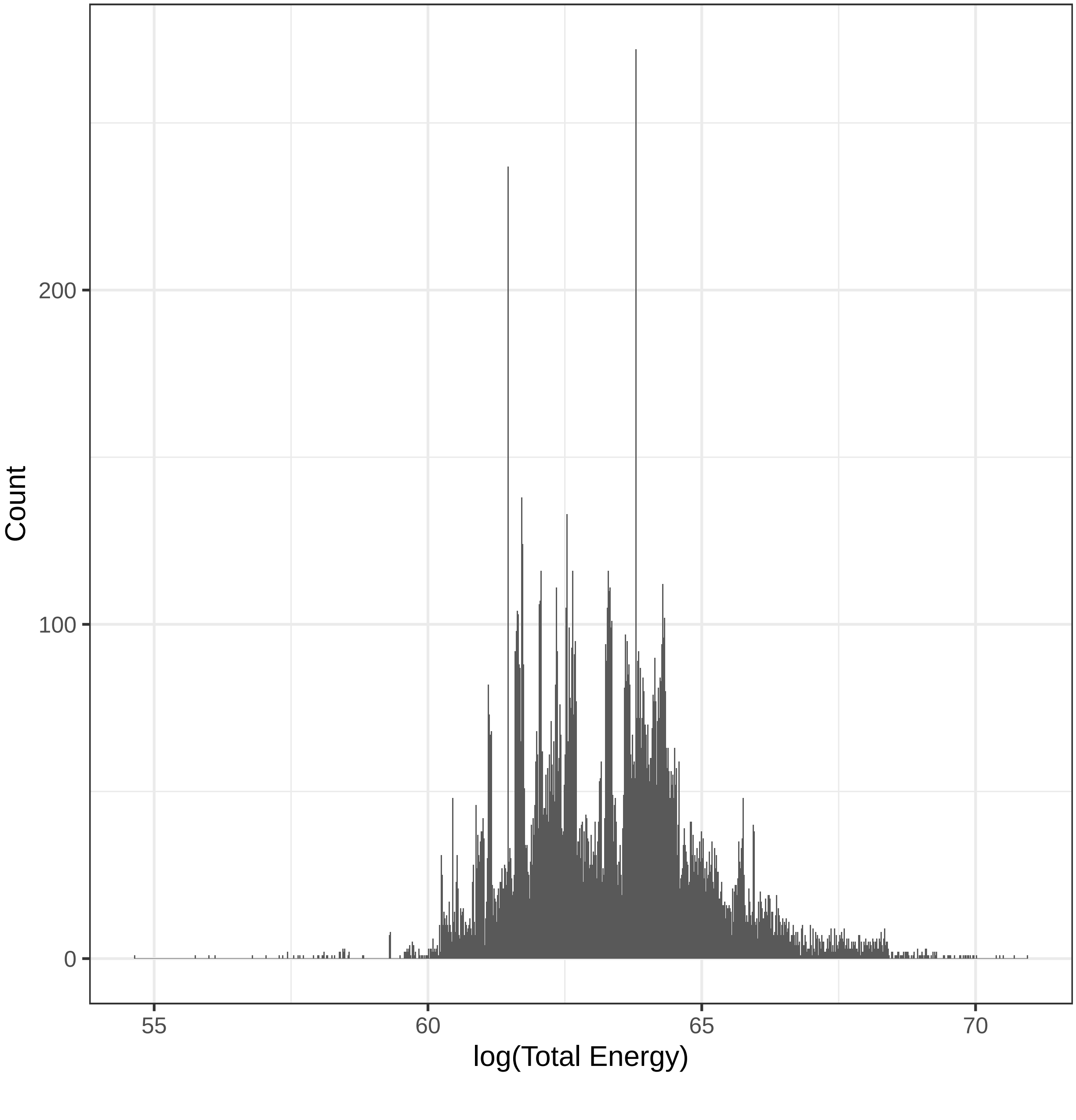

Figure 4: Jittered distribution resolves discreteness

Power-Law Distribution

The total energy release, peak flux, and duration of solar flares follow a power-law over a wide range of values. The power-law distribution is given by

$$f(x) = \frac{\alpha - 1}{x_{min}} (\frac{x}{x_{min}})^{-\alpha}$$

where

$$x_{min} > 0 \ , \ \alpha > 0$$

The distribution of the log values, $y = \log(x)$, is then given by an exponential distribution

$$f(y) = \lambda e^{-\lambda(y - \log(x_{min}))}$$

where

$$\lambda = \alpha - 1$$

However, due to the sensitivity limitations of satellities discussed in the introduction, we experience a detectability problem that causes the power-law distribution to turn over at the left end. For the total energy of a flare, we would expect that the power-law distribution is followed even at low flare energy levels, but we have a significant amount of flares missing at the left end because they were not detected. See *Figure 5* for a visualization of this problem.

The detectable flares still follow a power-law, but this is only within a bounded region of the observed distribution. In order to estimate this power-law without selecting an arbitrary starting point, we propose the Maximum Product of Spacings method to determine the bounded region and power-law within.

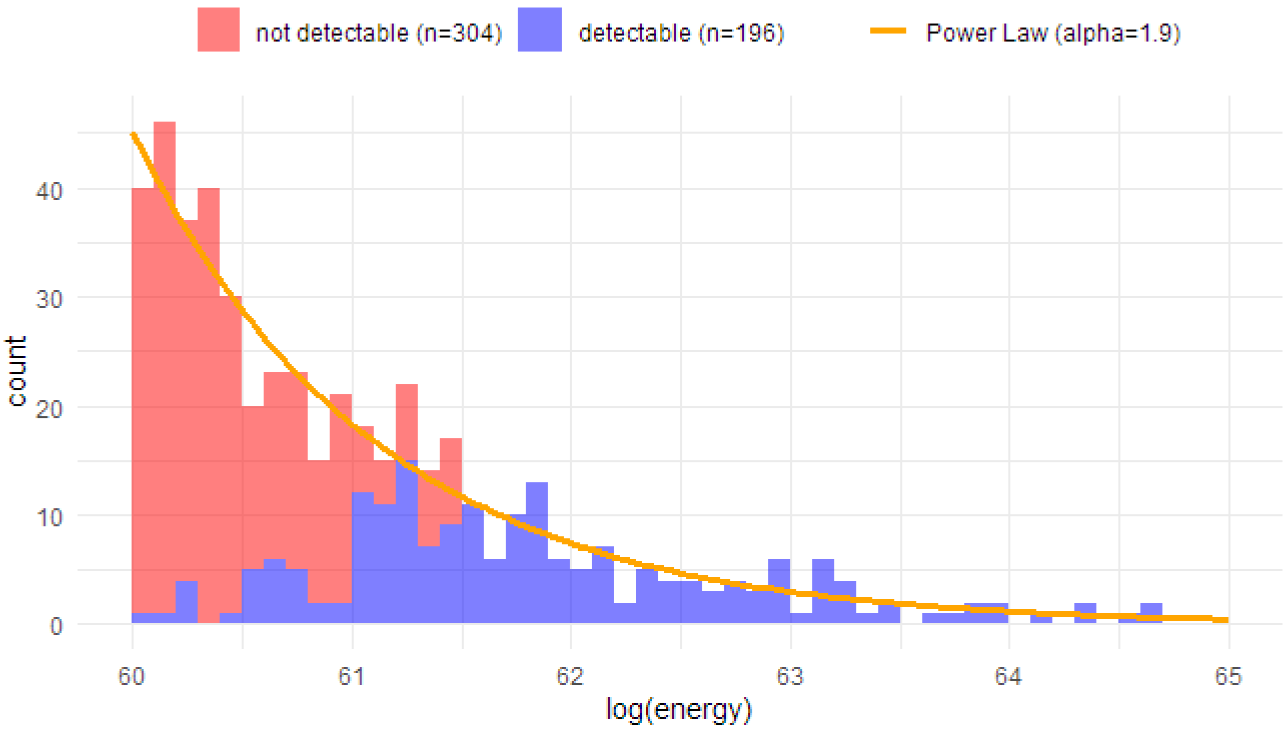

Figure 5: Visualization of the Detectability Problem

Maximum Product of Spacings

As opposed to utilizing the Maximum Likelihood Estimation (MLE) method and "eyeballing" the point at which the power-law starts for the detectable flares, we use the Maximum Product of Spacings (MPS) method because it allows us to simulatenously estimate the left end, right end, and the exponent α of the power-law (see *Figure 6*). As opposed to the MLE, which uses the likelihood function, MPS maximizes the product of the differences between the cumulative distribution evaluated at the order statistics[1]. Simply stated, MPS is given by the following

$$max\Pi_{i=0}^{n}F(x_{(i+1)}) - F(x_{(i)})$$

Before implementing the MPS algorithm on our solar flare data, we tested its robustness

[1] Note that the MPS algorithm fits to the exponential form of the power-law, so it returns the parameter λ, where $\alpha = \lambda + 1$.

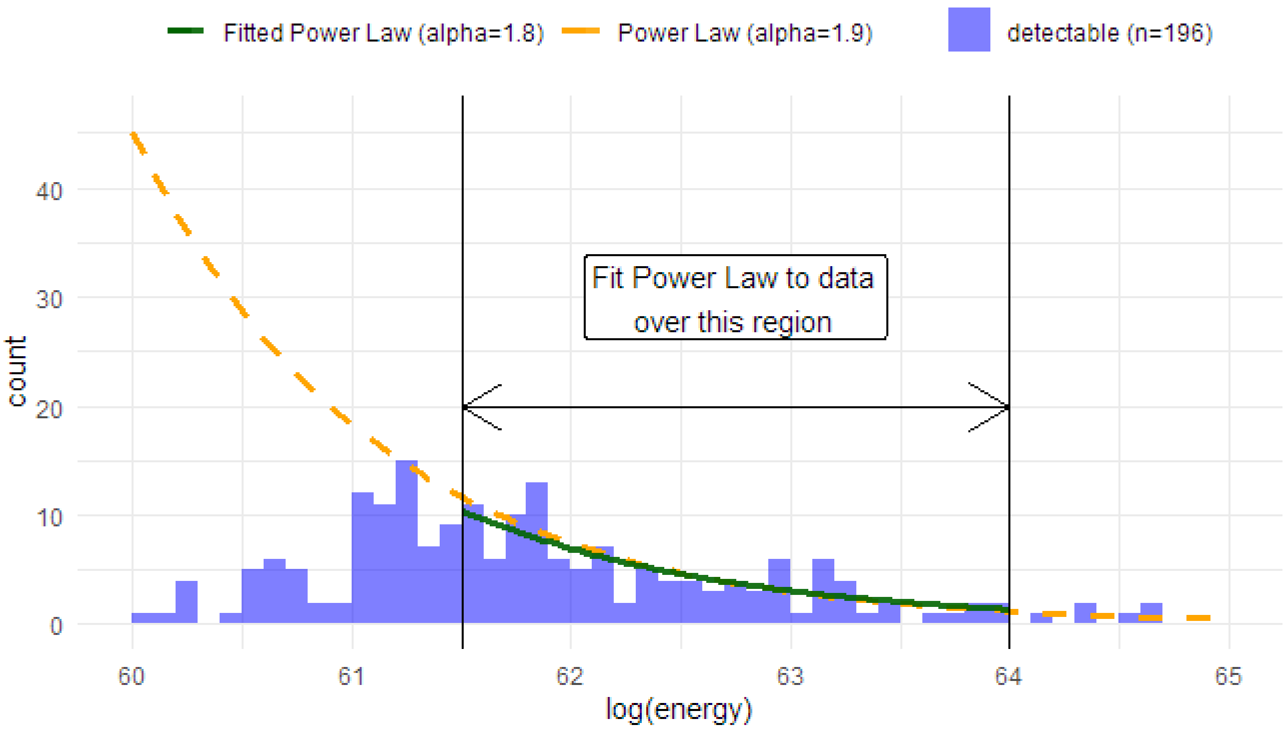

Figure 6: Bounded Region Where Observed Power-Law Occurs

on 15 combinations of simulated model parameters and a range of sample sizes. *Figure 7* shows the performance of the algorithm through a boxplot of the difference between the fitted and true λ as the sample size increases. We observe that as the sample size increases, the fitted values approach the true values of α. *Figure 8* gives the runtime of the algorithm as the sample size increases, increasing as N^2.

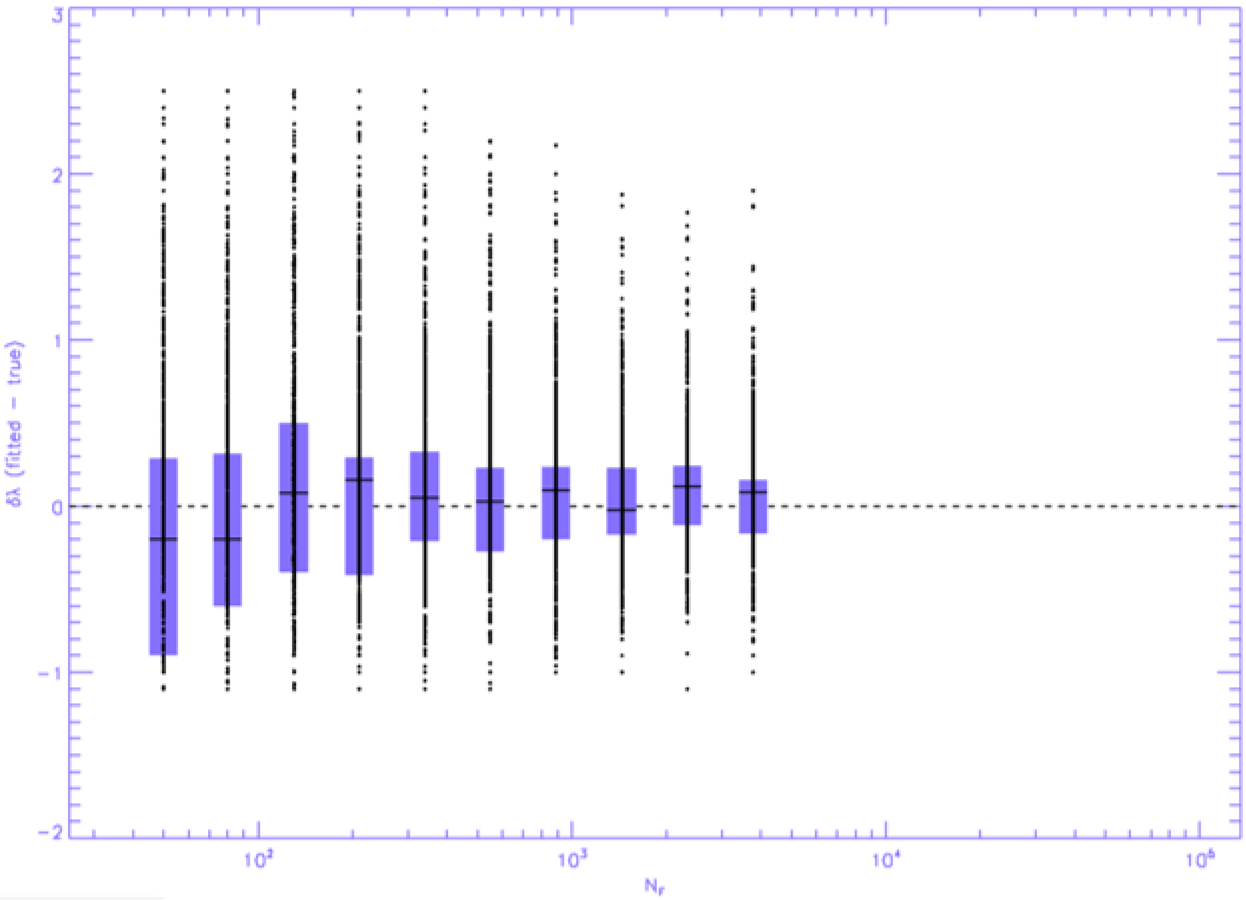

Figure 7: Fits Approach True Value as Sample Size Increases

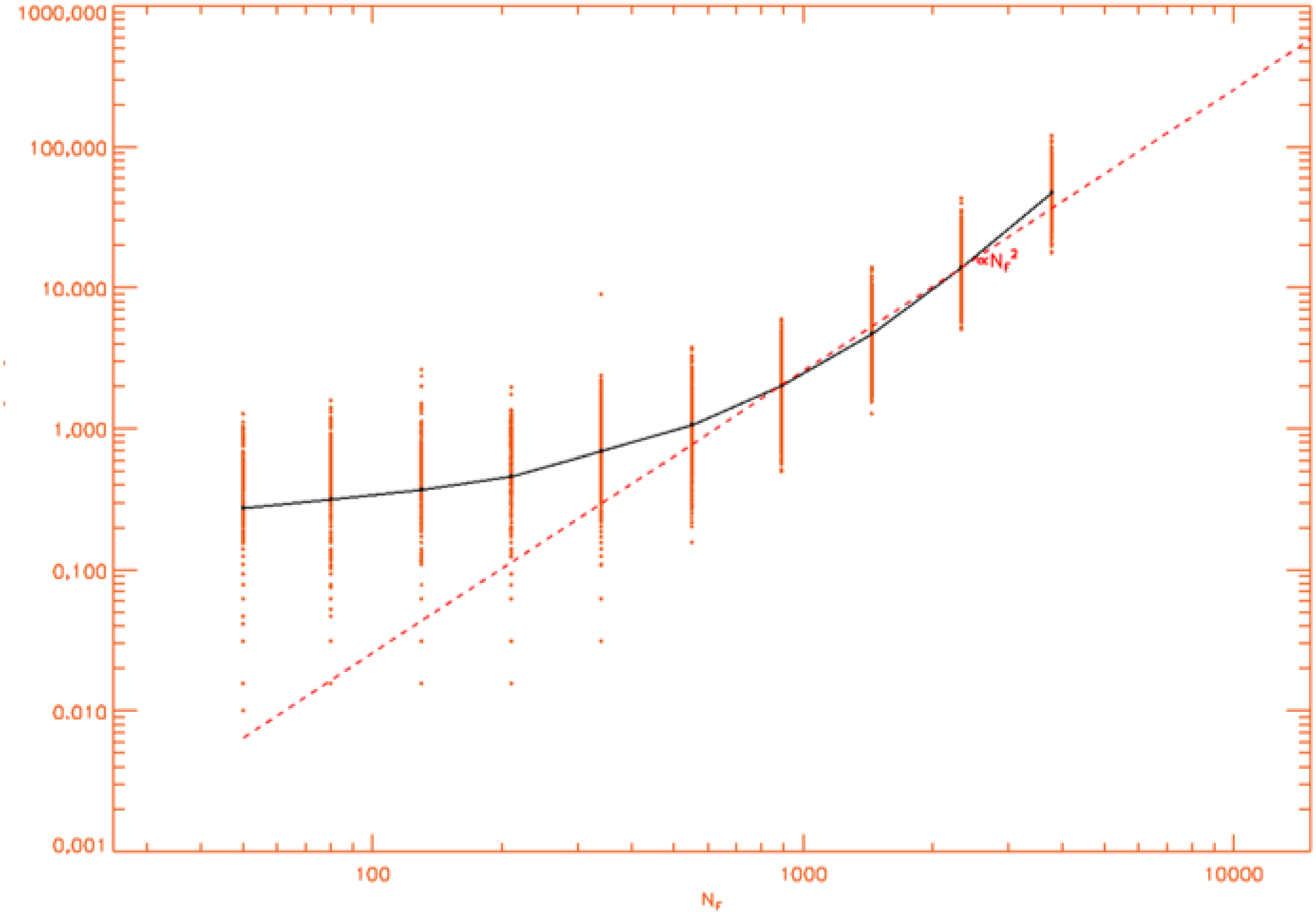

Figure 8: Runtime Increases as N^2

Monte Carlo Simulations

In addition to the value given by the MPS algorithm after fitting to our original data, we report error bars with the estimate to give a range of uncertainty by using Monte Carlo simulations. We first take a bootstrap sample from our original sample. The size of the sample is determined by a Poisson random variables with the mean equal to the observed sample size. This allows for variation in our estimates that could be caused by differing sample sizes. Next, we add randomly distributed noise, with a standard deviation of 10%, to each sampled value in the bootstrap sample to account for systematic uncertainties. After this, we fit the power-law distribution using the MPS algorithm. This process is repeated 100 times to obtain a bootstrap sampling distribution. We then take the mode of the sampling distribution as the point estimate for our parameters and report the error bars as the 68% highest density interval.

Distribution of Flare Energies

After fitting the power-law distribution to our data using the MPS algorithm and obtaining error bars from the Monte Carlo simulations, we received the estimates for the aggregate, by-cycle, and by-year subsamples for total energy, peak flux, and duration. These results can be found in the following sections.

Total Energy

Aggregate and By-Cycle

The estimated α for the aggregate total energy data is similar to previous estimates, but the by-cycle estimates suggest that they may differ by a significant amount.

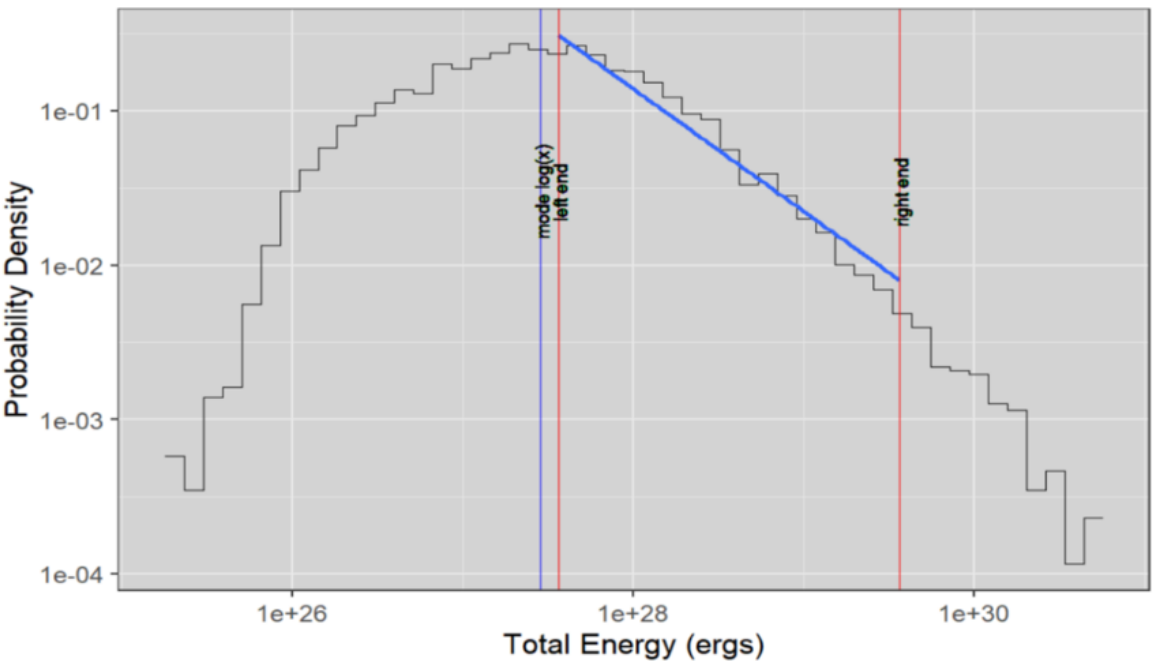

Figure 9: Aggregate Total Energy Distribution

Table 3: Aggregate and By-Cycle Total Energy Fitted Values

Cycle	Alpha	Mode of ln(x)	ln(Left End)	ln(Right End)
Aggregate	1.794005	63.220000	63.47	68.07
23	1.767075	63.220000	63.47	67.81
24	2.068505	2.068505	63.20	68.78

By-Year

We observe a possible trend appearing within each cycle for the fitted α values. The power-law seems to steepen as the solar maximum is approached and flattens as it gets closer to the solar minimum[2].

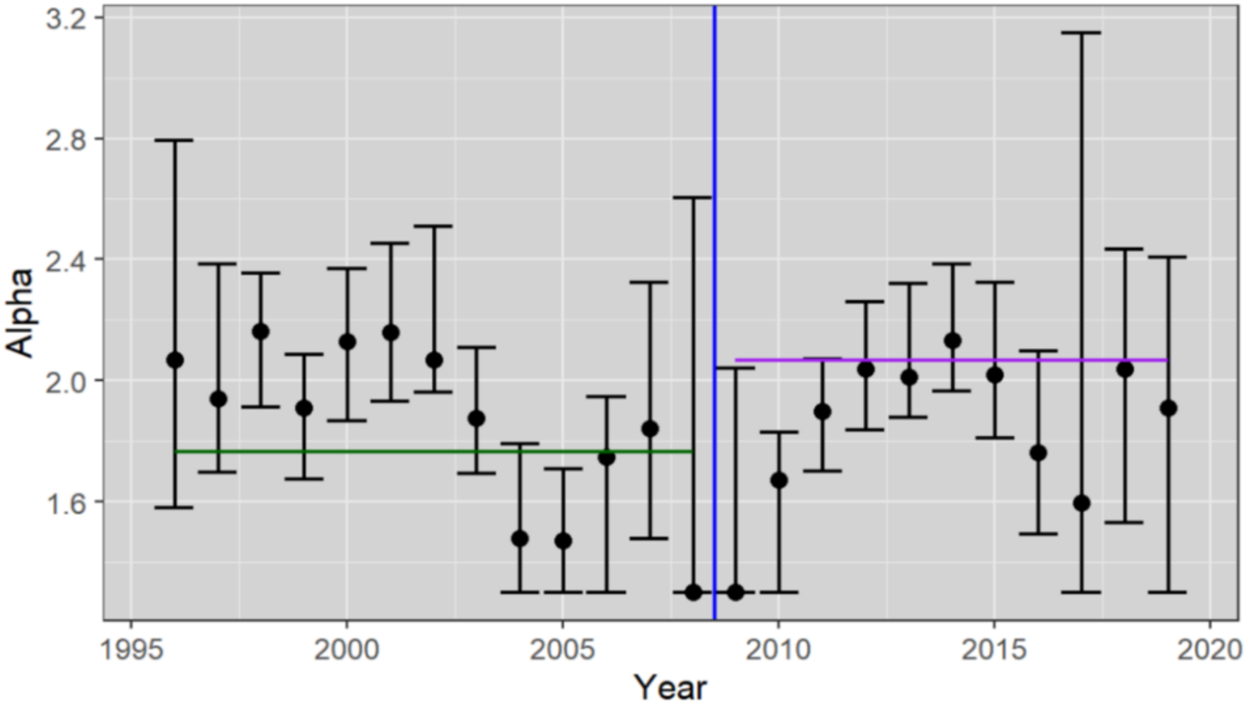

Figure 10: By-Year Total Energy Fitted Alpha

Peak Flux

Aggregate and By-Cycle

The fitted α values are consistent across solar cycles and the power-law region is found to contain C-class and M-class flares within our observed data. This could suggest that the probability that C and M-class flares are detectable is near 1 since the fitted region follows a power-law according to MPS.

[2] Note that the horizontal colored lines represent the α estimates for the by-cycle values and the horizontal blue line represents the end of cycle 23 and the start of cycle 24.

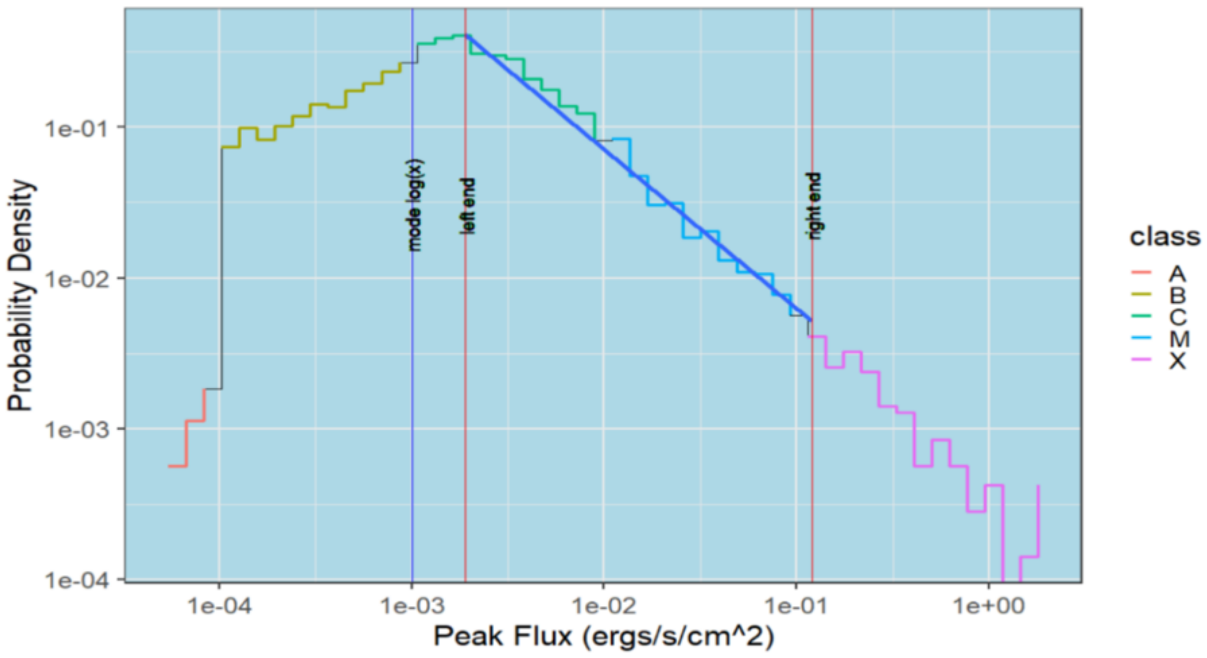

Figure 11: Aggregate Peak Flux Distribution

Table 4: Aggregate and By-Cycle Peak Flux Fitted Values

Cycle	Alpha	Mode of ln(x)	ln(Left End)	ln(Right End)
Aggregate	2.055357	-6.91	-6.27	-2.12
23	2.083280	-6.91	-6.21	-2.32
24	2.038141	-6.93	-4.61	-2.38

By-Year

Unlike Aschwanden et al. 2012, we do not observe an obvious trend in α by year throughout the two cycles for peak flux.

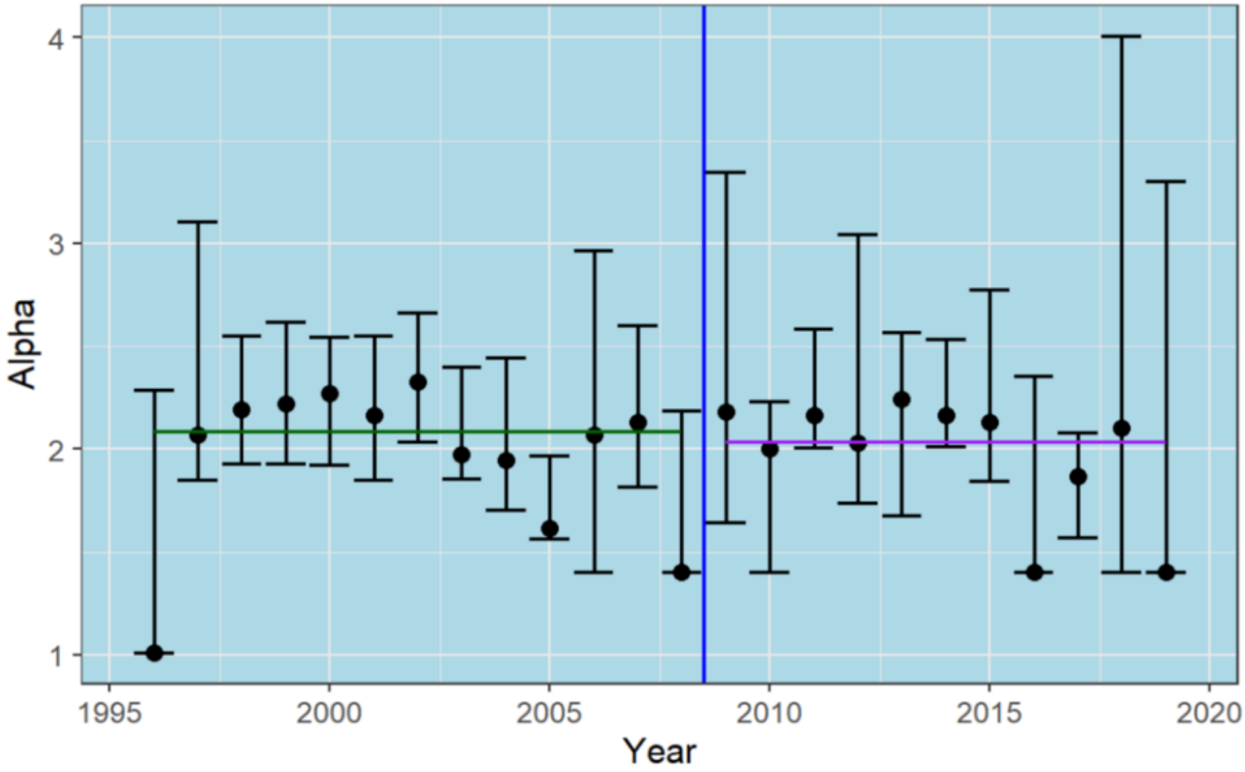

Figure 12: By-Year Peak Flux Fitted Alpha

Duration

Aggregate

The power-law for duration is substantial and does not go out of our expectations.

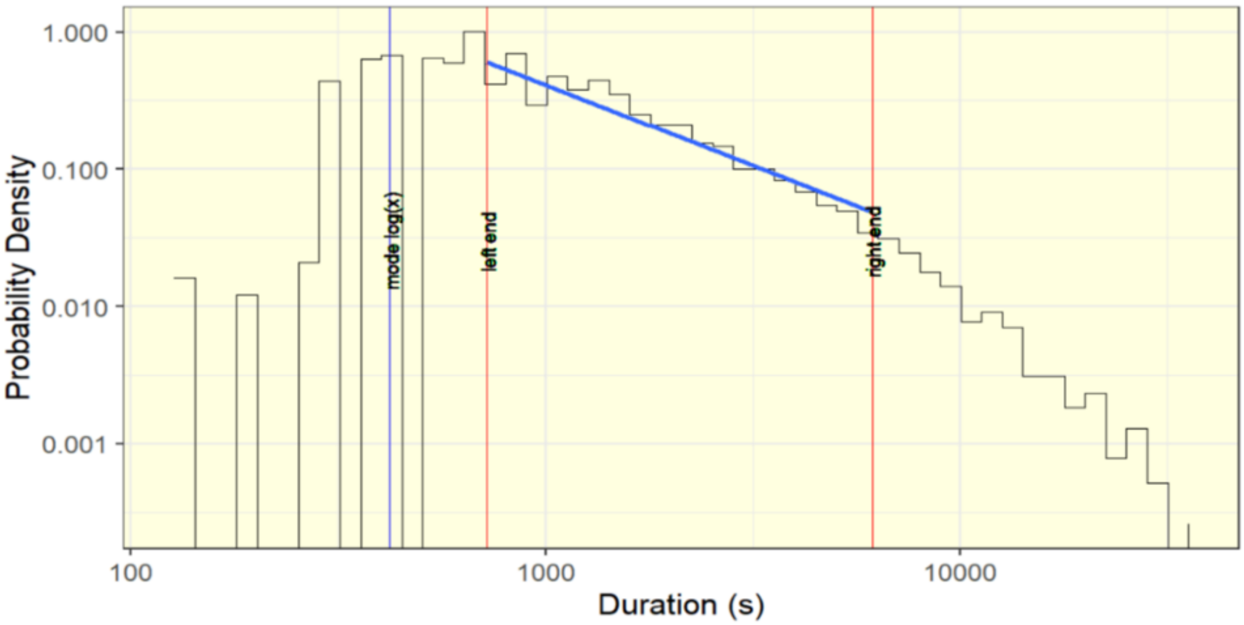

Figure 13: Aggregate Duration Distribution

Table 5: Aggregate Duration Fitted Values

Cycle	Alpha	Mode of ln(x)	ln(Left End)	ln(Right End)
Aggregate	2.180884	6.043652	6.582649	8.719979

By-Year

The results for α by year suggests that duration varies by considerable amounts throughout each year, although the error bars are rather wide in some instances.

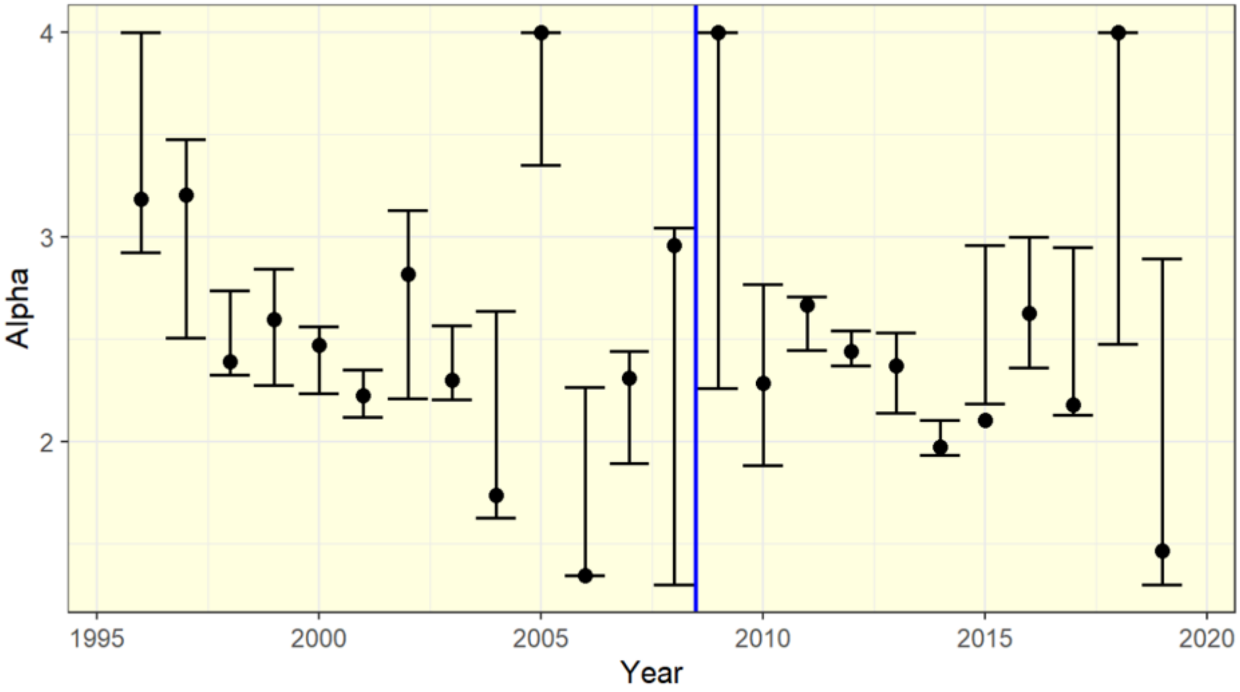

Figure 14: By-Year Duration Fitted Alpha

Distribution of Flare Counts

After obtaining the results from all the power-law fits of the solar flare properties, we began to investigate the goodness of fit of the Poisson distribution that was used to model the number of solar flares in our Monte Carlo Simulations. It was important that we understand how to model the number of flares that occur each year given the importance of the solar cycle and the effects of variation in sample size on our fits given by the Maximum Product of Spacings algorithm.

The number of flares occuring is clearly temporal, specifically being dependent on the phase of the solar cycle (see *Figure 1*). Because of this, we cannot draw from a fitted distribution to the aggregegate count data. Instead, it would be more appropriate to subset the data by some unit of time, such as by year, and model the counts for each subset.

When we subset our data by year and take the *rate of counts per 30 days*, we get the sample statistics for each year as seen in *Table 6*.

Table 6: Flare Counts per 30 Days

year	count	mean	var	sd
1996	233	46.6	1678.30	40.97
1997	1125	86.5	5439.27	73.75
1998	2217	170.5	1344.77	36.67
1999	2382	183.2	1977.19	44.47
2000	2631	202.4	3128.76	55.94
2001	2670	205.4	5389.42	73.41
2002	2676	205.8	5001.31	70.72
2003	2357	181.3	2750.23	52.44
2004	509	39.2	276.14	16.62
2005	520	40.0	847.17	29.11
2006	496	38.2	1116.81	33.42
2007	559	39.9	2186.38	46.76
2008	88	11.0	103.71	10.18
2009	237	21.5	691.87	26.30
2010	1243	95.6	2701.09	51.97
2011	2183	167.9	3615.91	60.13
2012	2154	165.7	2951.56	54.33
2013	2017	144.1	6045.46	77.75
2014	2226	171.2	2935.86	54.18
2015	2019	155.3	1767.23	42.04

year	count	mean	var	sd
2016	1246	95.8	2287.14	47.82
2017	1079	83.0	4768.17	69.05
2018	333	25.6	718.76	26.81
2019	245	20.4	711.54	26.67

This table illustrates that it is not appropriate to fit a distribution to the aggregate data, no matter the "rate" for flare counts, because there is so much variation in the total counts by year as a result of the temporal dependency. If we fit a distribution for each year, we need to select a rate for our counts. For example, the number of flares that occur "per 30 days" or "per 7 days" in a specified year.

A prevalent issue that can be observed in the table is the discrepancy between the mean and the variance for each year. This cautions us in our selection what distribution to use to model our counts, which was initially the Poisson distribution.

The graph below shows the trend in flare counts for each year by binning the counts by month. For years like 2004 and 2008, there is no noticeable issues with the count data, but in most other years we observe the counts changing throughout the individual year. The high variance is a result of these changing counts within each year, especially in cases with extreme outliers.

We could potentially address this in several ways, one of which is changing our unit of time for which we model the flares. Rather than fit a distribution for each year (which is arbitrary in terms of the sun), we could segment by the phase of the solar cycle. If this segmentation is done properly, the trend may be "washed out." We compare the Poisson distribution and the Negative Binomial distribution for the 30-day rate in the following sections.

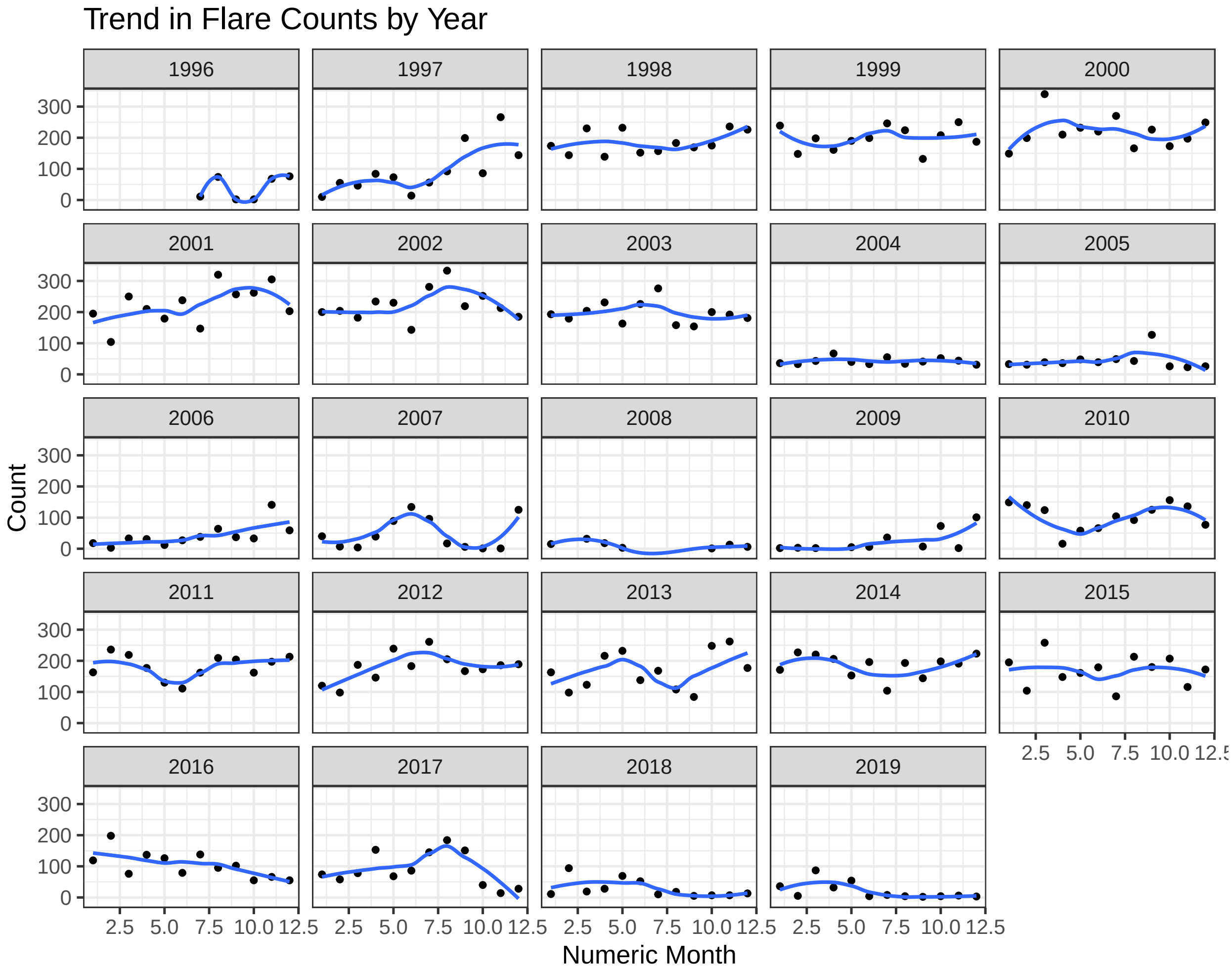

Figure 15: Trend in Flare Counts by Year

Poisson Distribution

For a discrete random variable that follows a Poisson distribution, the distribution takes on the functional form

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

This distribution models the number of times an event occurs within some interval of time. The Maximum Likelihood estimate is simply given by the mean

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

Thus the $E[X] = \lambda$, which is also the variance: $V[X] = \lambda$. This property of the Poisson distribution should be considered, especially since the variance is much larger than the mean of flare counts for each year (see *Table 6*).

Negative Binomial Distribution

Given that we are observing a sample variance much greater than our mean, the Poisson should be used with caution. Instead, we can fit a Negative Binomial distribution to our by-year data. This distribution has two parameters, the mean and the dispersion parameter, allowing for greater variance. We use Negative Binomial Distribution Family Function `negbinomial()` from the VGAM package in R to obtain our estimates.

The Negative Binomial distribution takes on the form

$$P(X = k|\mu, s) = \binom{k+s-1}{k}(\frac{\mu}{\mu+s})^k(\frac{s}{\mu+s})^s$$

Where μ is the mean and s is an index parameter. The dispersion parameter is given by $\frac{1}{s}$. We can derive the variance from μ and s, where

$$V[X] = \mu + \frac{\mu^2}{s}$$

Results

Conclusion

By utilizing the Maximum Product of Spacings method, the fits of the exponents and bounds of the power-law regions were found simultaneously and were statistically demonstrated to follow power-laws. We confirm that the exponent for the total energy power-law is similar to previous estimates and does not exhibit a detectable trend over the past two cycles. However, we do not find a noticeable trend in the estimated exponents throughout the solar cycles for peak flux. This outcome is different than what was found by Aschwanden et al. 2012.

We also found that at any given time, the observed flare counts are overdispersed by more than twice the expected amount under the Poisson distributional assumptions. We proposed a negative binomial distribution to take into account the variability in counts and found that it is a much more appropriate fit. We plan to continue analysis on the distribution of flare counts, as this is vital to obtaining our error bars. Moving forward, we plan to fit the power-law distribution using MPS to all active regions and orders within a sequence of flares. Additionally, it will be important to finalize the analysis of the flare counts distribution and to continue to test the MPS algorithm. Once our results are finalized, we will then utilize the RHESSI satellite database to compare our results to the GOES data.

References

Freeland, ASchwanden; 2012. “Automated Flare Statistics in Soft X-Rays over 37 Years of Goes Observations: The Invariance of Self-Organized Criticality During Three Solar Cycles.” *The Astrophysical Journal.*

Kashyap, et al. 2020. “Impulsive Energy Deposition into Coronar Through Self-Organized Criticality.” *BAAS.*

Pasachoff, Leon Golub; Jay. 2010. “The Solar Corona.” In, 2nd Edition, 1–20. Cambridge University Press.

Shao, Hahn. 1999. “Maximum Product of Spacings Method: A Unified Formulation with Illustration of Strong Consistency.” *Illinois Journal of Mathematics.*