WILDFIRES: ARSONISTS, CHILDREN, AND POWER LAWS

October 26, 2018

Jacob Delveaux
King's College London
Department of Mathematics and Natural Sciences

Contents

Abstract

Wildfire occurrences caused by children and arsonists have been investigated. Arsonists have started an order of magnitude more fires than children from 1970-2000 (23,336 vs. 1,457). Yearly fires by children appear to decrease at 0.1% per year. Power-law fits to frequency-size distributions have been analyzed and validated. Linear least squares fitting methods prove to be poor indicators of power law coefficients, failing Kolomogorov-Smirnov goodness of fit tests by more than 8 times the threshold distance. Maximum likelihood estimators prove to provide more sufficient fits of the data. Analysis of the power-law exponents indicate arsonists have higher proportions of large to small fires than children.

1 INTRODUCTION

Arsonists and children require constant supervision, are unable to handle aggression in a constructive manner, and cause significant economic/structural damage. Herein, wildfire data within the contiguous United States (U.S.) has been segmented into two partial data sets, 1) those started by arsonists and 2) by children. The aim of this research is to investigate when, where, and with what frequencies these groups start large, economically and environmentally costly fires. The data will be presented spatially and temporally; in-depth frequency-size statistics will allow direct comparison of the likelihood for each group causing extreme events. The process and applicability of fitting the distribution with a power law will be discussed in some detail. Finally, the sizes of of 50 and 100 year fires will be determined, with a final analysis on the relative danger of the two groups.

2 DATA

2.1 Description of Data

The wildfire data researched[2] is a subset derived from a database of 657,949 wildfires, reported by the United States Fire Service (USFS) and U.S. Department of Interior (DOI), ranging from 1970 - 2000 across the contiguous U.S. composed and assessed by Brown et, al. (2002)[3]. Initial consistency issues were addressed by Malamud et, al. (2005)[9], who reduced the data set to USFS reported fires with values ≥ 1 acre (0.004047 km^2), to account for spotty reporting by the DOI prior to the 1980s and incomplete records of very small fires.

2.2 Format of Data

Wildfire data is partitioned into ecoregions, each entry provides a date of occurrence (year, month, day), a latitude and longitude, a general and specific cause, a flag indicating errors with reported data, and the total burned area.

2.3 Segmented Data Sets

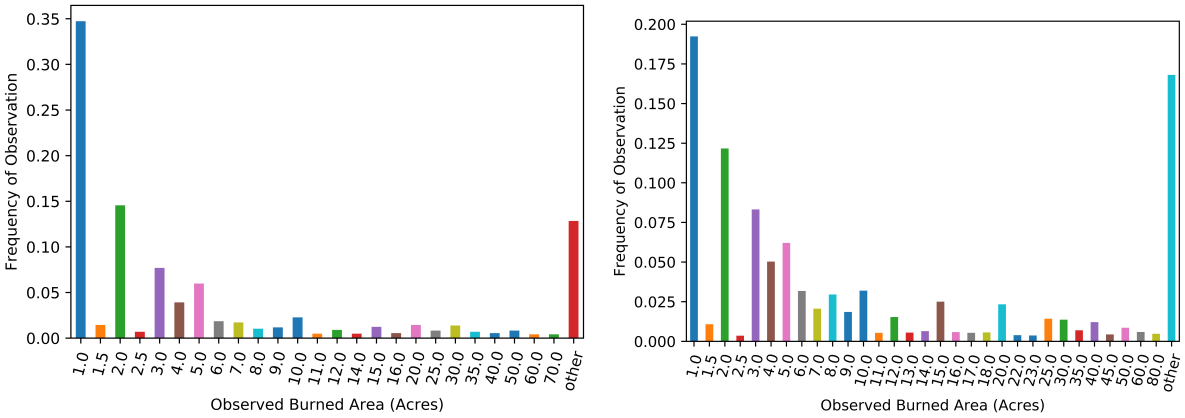
From the 88,916 fires, two segmented sets were constructed according to the general cause provided by the USFS. Those labeled with indicator 7 and 8, corresponding to fires started by arsonists (23,336 entries) and children (1457 entries) respectively, were tabulated in separate spreadsheets.

2.4 Limitations of Data

Limitations stem from the pre-processing, classification, and reporting of the data. Small fires being removed could contribute to inaccuracy of the counts and frequency-size distribution for each group. Frequency-size statistics may report a higher tendency for larger fires than expected. Cause classification by USFS leads to more confusion. An incident reported as equipment-use may have been a child using equipment, or an arsonist playing off his crime; furthermore there is the obscure case of child-arsonists. Estimating the quantitative impact of these limitations is beyond the scope of this paper.

Figure 1 represents the frequency of reported values for each data set. In the lower end, reports are frequently at integer and half-integer values, past 10 acres the reports appear rounded to the nearest 5 acres, past 40 acres this spacing increases 10 acres. This trend

demonstrates a bias in the reported acreage, leading to uncertainty in the actual total burned area.



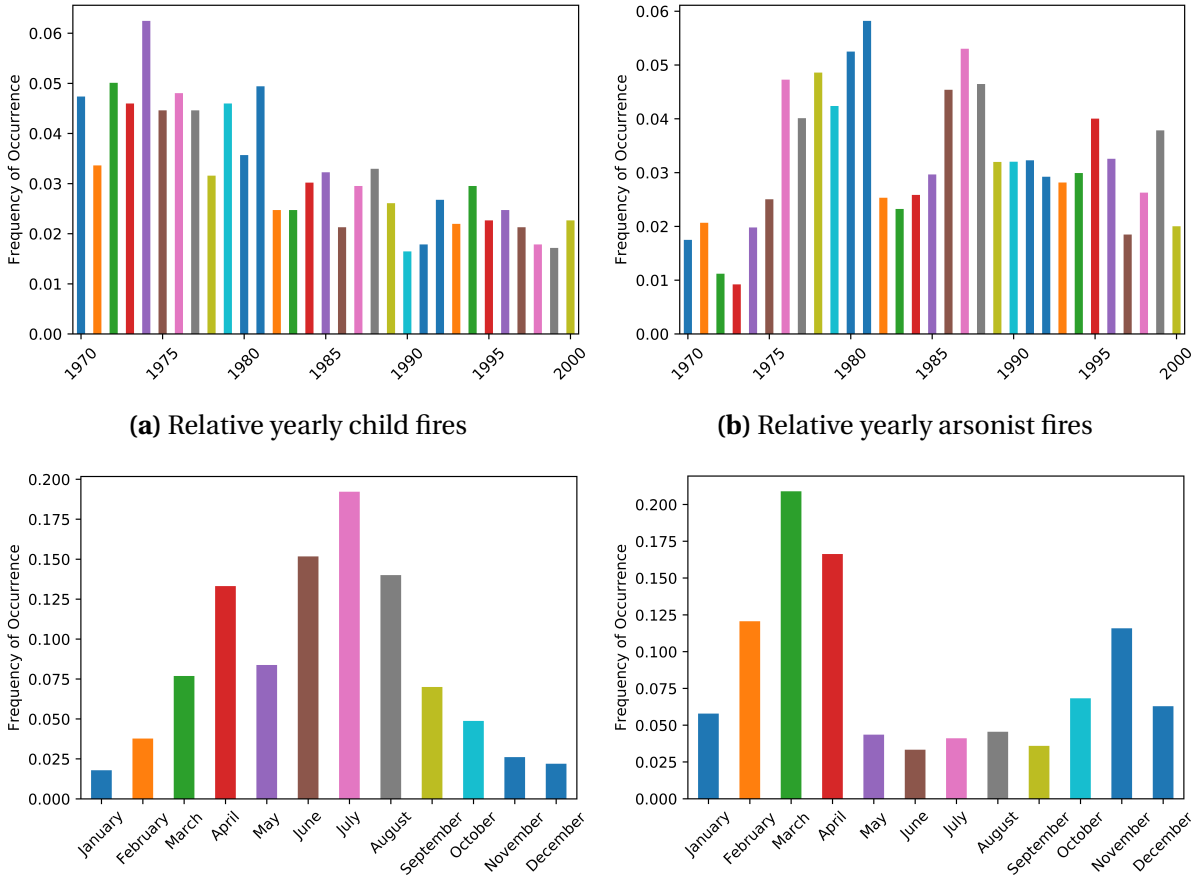
(a) Area Report Frequency: Child Data **(b)** Area Report Frequency: Arson Data

Figure 1: Reported values of wildfire burned area. Reports with frequency less than 0.3% are placed in 'other'. There is a clear bias in reporting integer values.

3 ANALYSIS

3.1 Temporal Distribution

Figure 2 illustrates the relative number of fires per year and per month for each group. A linear regression was performed to determine the trend of yearly activity for each group. The child data yielded $slope_{child}$ = -1.08 x 10^{-3} ($r^2 = 0.654$), indicating a general decrease in yearly fires, whereas $slope_{arson}$ = 1.46 x 10^{-4} ($r^2 = 0.012$) is inconclusive due to poor fit.



(a) Relative yearly child fires

(b) Relative yearly arsonist fires

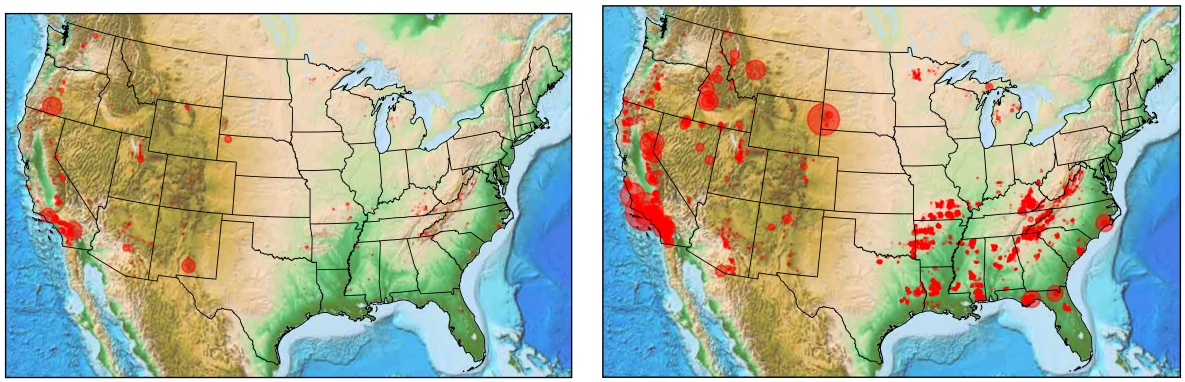
(c) Relative monthly child fires

(d) Relative monthly arsonist fires

Figure 2: Temporal distributions of fire frequencies for arsonists (b,d) and children (a,c)

3.2 Spatial Distribution

A spatial plot of the data in Figure 3 reveals higher concentration of fires along the west coast and lower east coast, with largest fires in southern California. No significant difference appears other than the sheer number of data points between groups.



(a) Spatial Distribution: Child Data **(b)** Spatial Distribution: Arson Data

Figure 3: Location of fire occurrences for a) children and b)arsonists. The size of the marker represents the total burned area of the fire.

3.3 Frequency-Size Distribution

Considerable risk stems from the cases that are rare and extreme[4], to this end, for each group a frequency-size distribution was constructed as

$$f_{NC}(A_F) = \frac{\Delta N_F}{\Delta A_F} \tag{1}$$

where A_F represents the total burned area, and ΔN_F is the count of fires within a range of burned areas ΔA_F. This distribution quantitatively presents with what relative frequency a fire of a given size occurs. Bin width is chosen to reliably represent the fundamental structure of the data, to provide the most accurate histograms[13]. Therefore, bin size increased as A_F increased, giving equal bin spacing in logarithmic coordinates to reflect the disproportionate amount of small to large fires.

A cumulative frequency-size distribution, the number of fires greater than or equal to a specified size $N_C(\geq A_F)$, is plotted adjacently to the non-cumulative distributions in Figure 4.

3.3.1 Fitting Distributions

When the cumulative and non-cumulative distributions are plotted in log-log space, a linear relationship appears, indicating a power law distribution of the form[5],

$$f_{NC}(A_F) = \alpha A_F^{-\beta} \tag{2}$$

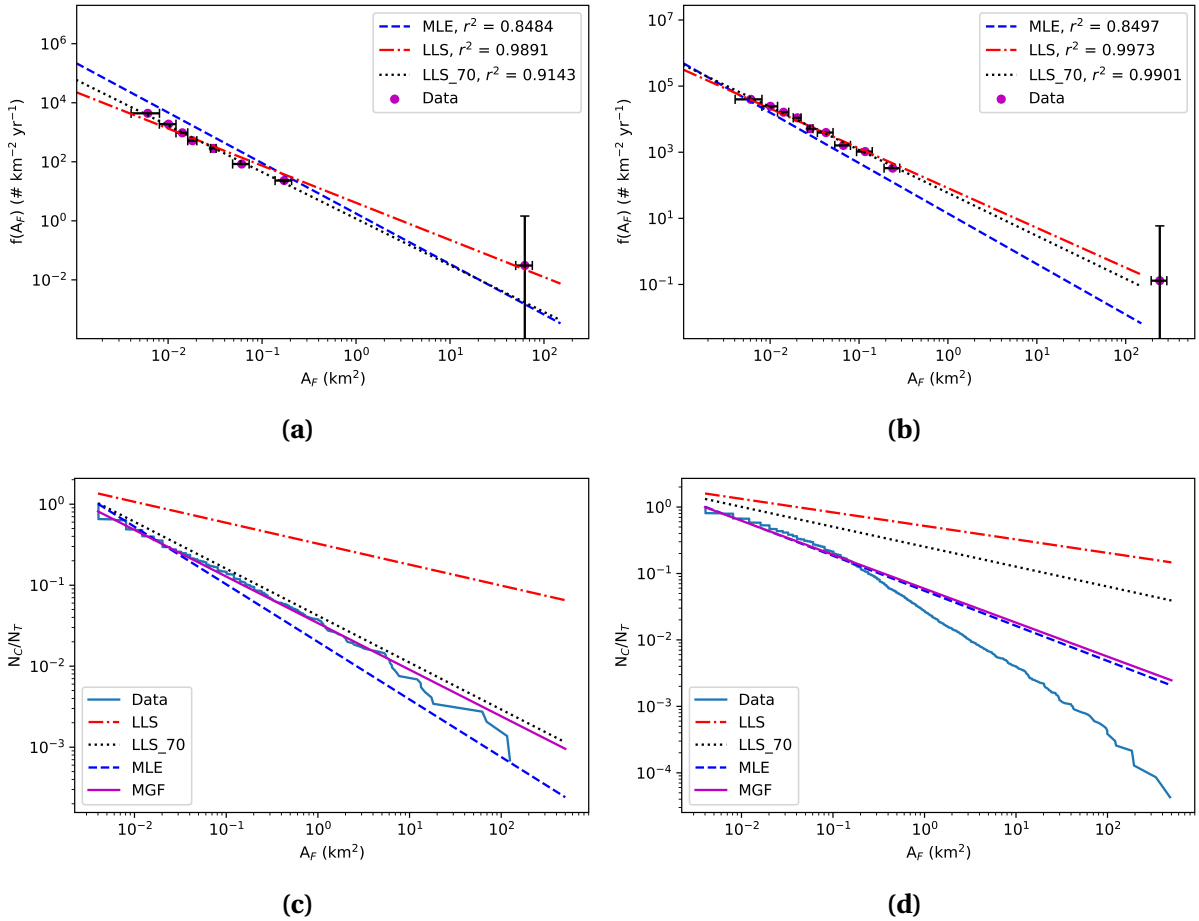


Figure 4: Non-cumulative (normalized by length of observation of 30 years) and cumulative (normalized by total number of fires) frequency-size distributions for a,c) children and b,d) arsonists. Best fit lines for power law coefficients generated by maximum likelihood estimation (MLE), linear least squares (LLS), LLS using the bottom 70% of data (LLS 70), and maximum goodness of fit (MGF). Error bars were determined by the method in Malamud et, al. (2005)[9]. Vertical error bars are $\pm 2\sigma$ with $\sigma = 2\sqrt{\Delta N_F}$. Horizontal error bars are derived from estimating the biased reports to generate 0.5 acre error on values $\leq$ 2.5 acres, and 20% of the reported value above 2.5 acres. Coefficients of determination (r^2) are provided for the non-cumulative plots, and can be found in Table 2 for cumulative.

which in log-log space becomes,

$$\log f_{NC}(A_F) = \log \alpha - \beta \log A_F \tag{3}$$

where α and β are constants of the non-cumulative frequency f_{NC}.

Least squares estimation (Linear Least Squares - LLS), arguably the most well known method of fitting linear data, minimizes the squared distance between the best fit line and individual data points. However, due to inaccuracies caused by large variation in the tail this method produces biased results[7]. Thus, two fits of LLS were performed, one for the whole data set and another for the first 70%. Maximum likelihood estimation (MLE) provides constants that would make the observed data most likely to be the data that occurs when drawing from the distribution[11], and is shown to be more accurate in predicting power law behavior than LLS[7,14]. The MLE exponent β is given by Deluca and Corral (2013)[6] as,

$$\beta = 1 + \frac{1}{\ln(g/a)} \tag{4}$$

where g is the geometric mean of the data, and a is the minimum value (0.004047 km^2). This estimation is valid in the case of a non-truncated distribution, however in the case $\frac{b}{a} \to 0$ with b as the maximum value, it provides a good approximation. The authors also determine the coefficient.

$$\alpha = \frac{\beta - 1}{a^{1-\beta} - b^{1-\beta}} \tag{5}$$

Table 1: Best fit constants and confidence intervals for power law fit to non-cumulative frequency-size distributions

Data Subset	MLE $log\alpha$	MLE β	LLS $log\alpha$	LLS β	LLS 70% $log\alpha$	LLS 70% β
Arson	1.14 ± 0.10	1.53 ± 0.01	1.91 ± 0.02	1.20 ± 0.03	1.77 ± 0.03	1.30 ± 0.03
Child	0.247 ± 0.172	1.71 ± 0.72	0.610 ± 0.348	1.26 ± 0.86	0.072 ± 0.269	1.56 ± 0.88

95% confidence intervals (1.96σ) were determined by fitting 1000 samples of the data and taking $2\sigma = \frac{|max-mean|+|min-mean|}{2}$. This is valid assuming the distribution of generated constants is Gaussian.

The power law constants for the cumulative distribution, $N_C(\geq A_F)$ are merely obtained by integrating over the non-cumulative distribution,

$$N_C(\geq A_F) = \int_{A_F}^{\infty} f_{NC}(A'_F)\, dA'_F = \int_{A_F}^{\infty} \alpha A_F'^{-\beta}\, dA'_F = \frac{t\alpha}{\beta - 1} A_F^{1-\beta}, if\, \beta > 1 \tag{6}$$

with the t our normalization factor, the number of years in the data set.

3.3.2 Testing Validity of Fits

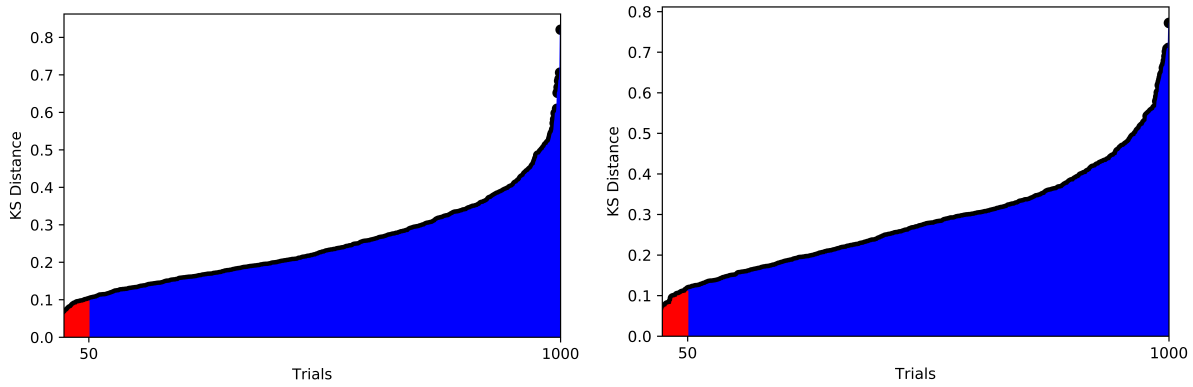
The validity of the LLS and MLE derived power laws were determined by the coefficient of determination (r^2) and the Kolmogorov - Smirnov (KS) goodness of fit test,

$$D = \sup_x |F_E(x) - F_t(x)| \tag{7}$$

with D as the distance statistic, between the fitted model $F_t(x)$, and the empirical cumulative distribution $F_E(x)$, taking the supremum over all data points. If this distance surpasses a threshold, D_R, the hypothetical distribution is rejected [10]. The threshold was determined by a bootstrapping method described by Babu et, al. (2004) [1, 12] as,

$$D_B = \sup_x |(F_{Es}(x) - F_{ts}(x)) - (F_E(x) - F_t(x))| \tag{8}$$

where the subscript 's' denotes a sample of the data. Threshold distances can be obtained from tables [10], however, the re-sampling preserves the underlying distributions present in the data and gives a more accurate threshold in the convergence limit [1] (See Babu et, al. (2004) pg. 68). The 95% confidence interval (CI) is illustrated in Figure 5.



(a) Child Data: $D_R = 0.0978$ **(b)** Arson Data: $D_R = 0.1033$

Figure 5: Resulting KS distances over 1000 trials of sampled data, biased by the empirical data as in Eqn. 7. The shaded red area represents the cutoff distance (95% CI) for not rejecting a model

An additional fit, dubbed maximum goodness of fit MGF, provides parameters for α and β based on minimizing the KS distance statistic on the cumulative distribution function.

Table 2: Coefficient of determination and KS goodness of fit for cumulative frequency-size distributions

Data Subset	MLE r^2	MLE KS	LLS r^2	LLS KS	LLS 70% r^2	LLS 70% KS	MGF r^2	MGF KS
Arson	0.937	**0.0937**	N/a	0.840	0.211	0.586	0.930	**0.0870**
Child	0.939	0.185	N/a	0.707	0.948	0.194	0.990	**0.0926**

MGF $\alpha_{child} = 0.034 \pm 0.001$, MGF $\beta_{child} = 0.575 \pm 0.001$, MGF $\alpha_{arson} = 0.059 \pm 0.001$, MGF $\beta_{arson} = 0.511 \pm 0.001$. MGF values obtained by searching parameter space for those that minimize KS distance. Red underlined values represent fits that have been rejected by the KS test, whereas green bolded values have not been rejected. The r^2 for LLS was not determined, as it clearly misrepresented the data (see Figure 4).

3.4 Large Event Return Period

Finally, the 50 and 100 year recurrence interval for wildfires is determined the relation,

$$RI(\geq A_F) = \frac{t+1}{N_C(\geq A_F)} \tag{9}$$

with recurrence interval $RI(\geq A_F)$ (year fires^{-1}) determined by the years of observation t plus one over the total number of fires observed of greater or equal size $N_C(\geq A_F)$.

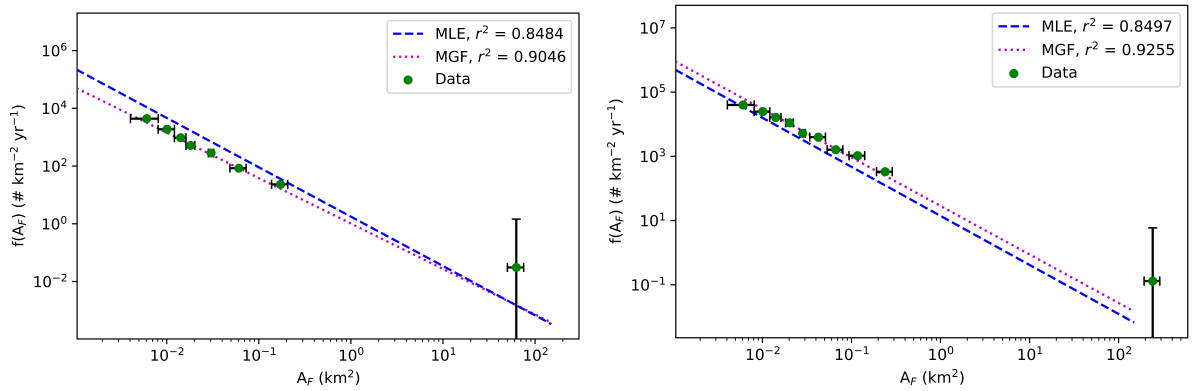
4 Discussion and Conclusion

4.1 Temporal Inferences

From Figure 2, the high activity of children in summer months likely corresponds to reduced supervision when outside of school. The yearly decrease may, in part, correspond to the internet revolution, where children start fires on online forums rather than in nature. The low frequency of arson cases in the summer months could be a case of mis-classification, where uncaught arsonists have their crimes attributed to natural forest fires. On the other hand, summer vibes and sunshine could cause arsonists to be more content. These trends may be better interpreted by a social scientist.

4.2 Frequency Distribution Fits

From the initial non-cumulative distribution, it appears the LLS and LLS of the bottom 70% best fit the data based on r^2 determination (see Figure 4). However, transitioning to the cumulative distribution, the failure of LLS to represent the data in both r^2 and KS distance in Table 2 is apparent. The LLS 70% method performed far better, giving credence to the claim the problem lies in the fluctuation of the tail[7, 14]. MLE and MGF appear to best represent the power law structure of the data. However, the MLE for the child case is rejected by KS test, indicating it is not a sufficient fit. Furthermore, although some proponents claim MGF to be a powerful tool[8] the statistical significance is unclear (see Luceño (2006) pg.905). Figure 7 provides the MGF fit in non-cumulative space, giving some validation to its legitimacy in this case.



(a) Child Data **(b)** Arson Data

Figure 6: Comparison of MLE and MGF fitting the non-cumulative distribution function. MGF_{NC} $\beta_{child} = 1.575 \pm 0.001$, MGF_{NC} $\beta_{arson} = 1.511 \pm 0.001$

4.3 Recurrence Intervals

The 50 and 100 year recurrence intervals are obtained by rearranging Eqn. 9 and substituting in Eqn. 6 to give,

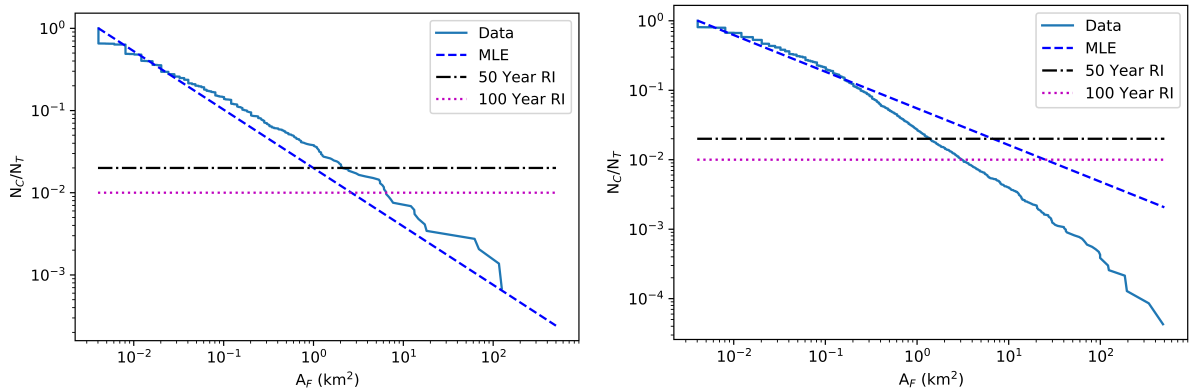
$$A_F = \left(\frac{\frac{1-\beta}{\alpha t} + \frac{1-\beta}{\alpha}}{RI(\geq A_F)} \right)^{\beta - 1} \tag{10}$$

where α and β are determined by MLE fit.

Table 3: 50 and 100 year fire sizes based on MLE estimate

Data Subset	50 year fire (km^2)	100 year fire (km^2)
Arson	17.85 ± 1.28	24.76 ± 1.87
Child	6.794 ± 1.93	8.303 ± 2.11

95% CI was determined from the error in α and β of the MLE. However, these estimates should be taken lightly, as this assumes that MLE fits the model precisely.



(a) Child Data **(b)** Arson Data

Figure 7: 50 and 100 year return intervals graphically represented. Horizontal lines indicate probability of 1/50 and 1/100, interception with the data indicates what size fire occurs with this yearly probability. Here, the inaccuracy of estimating return period with the MLE is clear.

4.4 Group Danger and β

Although both groups are dangerous in their own right, there is more risk for arsonists starting fires than children in all locations and any time of the year. Furthermore, taking the average of all β values obtained from each fit, we have $\beta_{arson,avg} = 1.34$ and $\beta_{child,avg} = 1.51$. The lower value of β indicates a higher ratio of large to small fires. Arsonists therefore not only start more fires, but larger fires than children. In the extreme case, the largest fires for each

group over the 30 year time period is $\text{size}_{arson,max} = 483km^2$ and $\text{size}_{child,max} = 125km^2$. Lastly, there is a 2% chance for an arsonist to start a fire of size ~$18km^2$ each year, whereas with the same probability a child would start a fire only of size ~$7km^2$.

4.5 Conclusion

The temporal, spatial, and frequency analysis of each group provides excellent insight into the relative danger of arsonists and children starting wildfires. Arsonists have conclusively been found to be far more dangerous in this regard than children. Statistical measures of the frequency-size distribution illustrate the difficulty in finding fitting methods that provide consistent results. Future research should inquire further into the spatial distribution of each group, to answer why the clustering along the west coast and south-east, and lack of fires in the Midwest.

REFERENCES

[1] Babu, G Jogesh and CR Rao (2004), "Goodness-of-fit tests when parameters are estimated." *Sankhya*, 66, 63–74.

[2] Brown, Timothy J (2004), "Electronic database of U.S. Department of Agriculture Forest Service wildfires over the time period 1970-2000. obtained by personal communications from Timothy J Brown (Univ. of Nevada, Reno) to Bruce D. Malamud (King's College London), in 2004."

[3] Brown, Timothy J, Beth L Hall, Charlene R Mohrle, and Hauss J Reinbold (2002), "Coarse assessment of federal wildland fire occurrence data." *Report for the National Wildfire Coordinating Group, CEFA Report*, 02–04.

[4] Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano (2013), "Catastrophic natural disasters and economic growth." *Review of Economics and Statistics*, 95, 1549–1561.

[5] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman (2009), "Power-law distributions in empirical data." *SIAM review*, 51, 661–703.

[6] Deluca, Anna and Álvaro Corral (2013), "Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions." *Acta Geophysica*, 61, 1351–1394.

[7] Goldstein, Michel L, Steven A Morris, and Gary G Yen (2004), "Problems with fitting to the power-law distribution." *The European Physical Journal B-Condensed Matter and Complex Systems*, 41, 255–258.

[8] Luceño, Alberto (2006), "Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators." *Computational Statistics & Data Analysis*, 51, 904–917.

[9] Malamud, Bruce D, James DA Millington, and George LW Perry (2005), "Characterizing wildfire regimes in the United States." *Proceedings of the National Academy of Sciences*, 102, 4694–4699.

[10] Massey Jr, Frank J (1951), "The Kolmogorov-Smirnov test for goodness of fit." *Journal of the American statistical Association*, 46, 68–78.

[11] Myung, In Jae (2003), "Tutorial on maximum likelihood estimation." *Journal of mathematical Psychology*, 47, 90–100.

[12] Praestgaard, Jens Thomas (1995), "Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions." *Scandinavian Journal of Statistics*, 305–322.

[13] Wand, MP (1997), "Data-based choice of histogram bin width." *The American Statistician*, 51, 59–64.

[14] White, Ethan P, Brian J Enquist, and Jessica L Green (2008), "On estimating the exponent of power-law frequency distributions." *Ecology*, 89, 905–912.

Word count
Words in text: 1672
Words in headers: 59
Words outside text (captions, etc.): 358
Number of headers: 21
Number of floats/tables/figures: 10
Number of math inlines: 57
Number of math displayed: 10
Subcounts:
text+headers+captions (#headers/#floats/#inlines/#displayed)
106+14+0 (2/0/0/0) _top_
137+1+0 (1/0/0/0) Section: Introduction
0+1+0 (1/0/0/0) Section: Data
89+3+0 (1/0/2/0) Subsection: Description of Data
37+3+0 (1/0/0/0) Subsection: Format of Data
45+3+0 (1/0/0/0) Subsection: Segmented Data Sets
150+3+35 (1/1/0/0) Subsection: Limitations of Data
0+1+0 (1/0/0/0) Section: Analysis
54+2+29 (1/1/6/0) Subsection: Temporal Distribution
41+2+31 (1/1/0/0) Subsection: Spatial Distribution
508+8+186 (3/4/26/8) Subsection: Frequency-Size Distribution
40+4+0 (1/0/3/1) Subsection: Large Event Return Period
0+3+0 (1/0/0/0) Section: Discussion and Conclusion
93+2+0 (1/0/0/0) Subsection: Temporal Inferences
141+3+20 (1/1/9/0) Subsection: Frequency Distribution Fits
27+2+57 (1/2/2/1) Subsection: Recurrence Intervals
120+3+0 (1/0/9/0) Subsection: Group Danger and β
84+1+0 (1/0/0/0) Subsection: Conclusion

Figure 8: Verification of word count, 1572 excluding abstract