Wildfires in Portugal

1761867

October 2017

Abstract

In this article I investigate the statistical properties of wildfires in Portugal from 1980 to 2005. I will show that area burnt exhibits an asymmetric frequency distribution with heavy tails on extreme events. Under the assumption that the frequency size distribution follows a power law model, I determine the parameters using maximum likelihood estimation and I quantify what an extreme event is in our context. I apply this methodology to two tasks: a) we compare different causes of large wildfires, showing that extreme events have a very characteristic behaviour with respect to their causes. b) we infer from the model the events with a recurrent time of 100 years.

1 Introduction

Many quantities that we encounter in everyday life, such as the life of a light bulb, or the height of male population are observed to take values not far from their average. For instance, the waiting time associated to a Poissonian process follows an exponential decaying probability density function (pdf), such that extremes events are very unlikely. On the contrary, crisis and booms are observed in financial markets, similarly extreme natural events such as floods, avalanches, and drought happen. It has been shown that these phenomena follow a frequency size distribution with heavy tails, that is often modelled as a power law pdf [6], although this is not a unique choice [8]. This approach has a social relevance, since predicting the likelihood of an extreme event is crucial both to policymakers and private entities (*e.g.* insurance companies). Moreover, the statistical properties of these rare events represents a benchmark to build realistic models of the dynamics (*e.g* natural cellular automata models [7])

Mediterranean countries are highly affected by wildfires, with high economical, social and environmental impacts [10]. Climate, topography, and distribution of tree are key factors to ease the develop of fires, and the combined effect with exceptional weather condition can give rise to extreme large wildfires [5]. In this report I investigate the statistical properties of wildfires in Portugal using a record from 1980 to 2005 [9]. In section 2 I discuss the proprieties of the database and his weakness, in particular with respect to the changes over time of the cause investigation efficacy. In section 3 I infer a power law model from

Table 1: Definition of main quantities used in the text

Name	**Symbol**
number of fires	N_F
number of fires per year	N_{FY}
number of fires per month	N_{FM}
area burnt	A_F
non-cumulative frequency size density	$f(A_F)$
cumulative frequency-size distribution	$F(A_F)$
normalised cumulative frequency-size distribution	$\hat{F}(A_F)$
probability density function (Pdf)	$p(x)$
cumulative density function (CDF)	$P(x)$
scaling parameter power law Pdf (see Eq. (1))	α
lower bound power law Pdf (see Eq. (1))	a

the data available and I discuss the considerations adopted in the choice of the parameters. I use the model to compute the event with 100 year recurrence time. Finally, I compare two sets of data with different labels.

1.1 Quantity definitions

In this report I make use of the non-cumulative frequency size density

$$f(A_F) = \frac{\delta N_F}{\delta A_F}$$

where δN_F is the number of fires in a bin of size δA_F. The cumulative frequency-size distribution $F(A_F)$ represents the number of values greater or equal than A_F. The normalised analogues are the normalised non-cumulative frequency-size density

$$\hat{f}(A_F) = f(A_F)/N_F$$

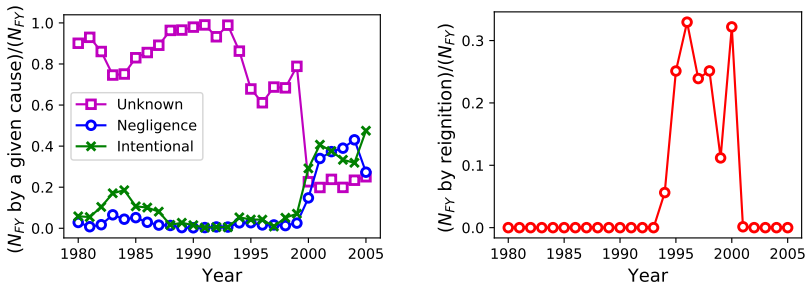
and the normalised cumulative frequency-size distribution

$$\hat{F}(A_F) = F(A_F)/N_F$$

where N_F is the total number of data points, these experimental quantities can be immediately compared to a probabilistic model. A probabilistic model is defined by the probability density function $p(x)$ or the cumulative density function (CDF)

$$P(x) = \int_x^{+\inf} p(t)dt = Prob[y \geq x]$$

.



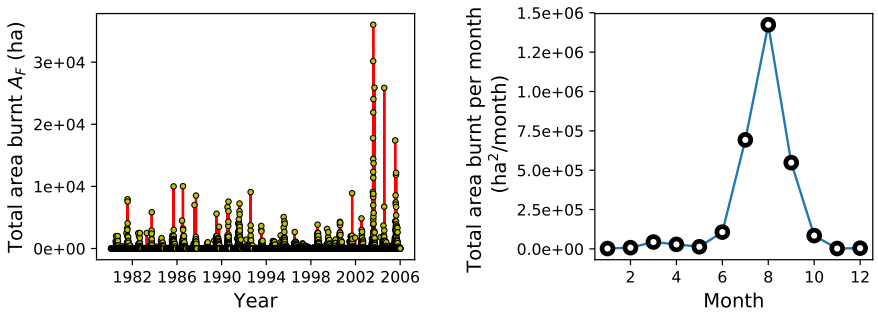
(a) (b)

Figure 1: 1a Number of fires in a given year N_{FY} whose cause was classified as "Unknown" (square), "Negligence" (circle), "Intentional" (x) divided by the total number of fires in the same period. "Unknown", "Negligence" , "Intentional" are the top 3 causes of larger area burnt A_F. The $N_{FY}(cause)/N_{FY}$ is given as a function of the year, 1980-2005. After 1999 the fraction of fires reported as unknown fell, whereas the fraction of Negligence and Intentional rose consistently. This change coincided with an increasing number of investigations [3]. 1b Same quantity as in 1a in the case of reignition wildfire. Reignition was only recognised as a fire cause for the period 1994–2001. Data from [9].

2 Data

2.1 Background proprieties and general limits of the data record

Data are provided by the "Autoridade Florestal Nacional", that is the current Portuguese Forest Service. This database, which encompasses more than 450000 wildfires over 25 years, reports the fires have affected the natural environment but not urban building. For each given wildfire, when available, several attributes are reported, but in this article I refer to: total area burnt A_F (not-burned area excluded), the causes of the fire, and the date of ignition. Database show lack of data, since data collection as well as the fire investigation accuracy was not homogeneous over the time. The minimum recorded area has changed toward smaller values in later years, and from data analysis it can be assumed that the accuracy in the recording for $0.1 < A_F < 1$ ha grows over time [9]. Hence, in order to handle homogeneous data, I filtered the data-set under the condition $A_F > 1$ ha. Moreover, the investigation strategies witnessed multiple refinement over time, whereas different methods for fire labelling were employed over the period considered (see **Figure 1**).



(a) **(b)**

Figure 2: 2a Time series of total area burnt A_F. Size frequency is asymmetric with a periodic behaviour due to season effect. 2b Sum of the total area burnt A_F in each month respect to months. Higher value are observed during summer.Data from [9].

2.2 Statistical proprieties of A_F

Wildfires in Portugal are strongly influenced by several environmental and anthropogenic factors [9]. Due to the Azores anticyclone, the distribution of precipitation exhibits a marked seasonal character[11].

Exploring the raw data of A_F over time Figure 2a, one concludes that A_F does not distribute symmetrically over a central value, as one would expect for a Gaussian distribution, but there are several large events. Furthermore, A_F shows a periodic behaviour that is highlighted by the plot of total area burnt A_F respect to months (Figure 2b): the largest fraction of wildfires occur during the dry summer months. I do not analyse inter-annual variation of A_F because the time length of the dataset is too short for a proper analysis.

3 Analyses

3.1 Model

Power law is defined as:

$$\begin{aligned} p(x) &= \frac{\alpha-1}{a^{1-\alpha}}x^{-\alpha} \qquad \text{Pdf} \\ P(x) &= \frac{1}{a^{1-\alpha}}x^{1-\alpha} \qquad \text{CDF} \end{aligned} \tag{1}$$

both for $0 < a < x < \infty$ and $\alpha > 1$.Taking the logarithm both sides, the two parameters to assign are the slope α and the left minimum a, with $a > 0$ in order to guarantee integrability. The power law distribution is a good model to capture the limit behaviour of large-size event distribution, but it unable to describe the small size behaviour. Hence the left limit a represents our apriori

knowledge of the starting point of the heavy-tails distribution. In this case I do not have a clear way to infer apriori such a limit. Let me assume for now I know a,then I could determine α using a linear fit in the log-log space, but it has been shown [2] that a more robust approach involves the Maximum Likelihood Estimator (see [4] for the derivation):

$$\alpha_e = 1 + \frac{1}{\log(g/a)} \tag{2}$$

where $g(x)$ is the geometric mean of $\{x : x > a\}$. The standard deviation of α_e is:

$$\sigma_e = \frac{\alpha_e - 1}{(1 - 1/N)\sqrt{N-2}} \tag{3}$$

In order to quantify the goodness of fit with respect to the empirical data, I use the Kolmogorov–Smirnov distance [1] that is defined as follows:

$$D = \max_{x>a} \left| P(x) - \hat{F}(x) \right|$$

where $P(x)$ and $\hat{F}(x)$ are the CDF of the model and the normalised cumulative distribution of empirical data respectively. Both has to be defined such that $F(a) = P(a) = 1$

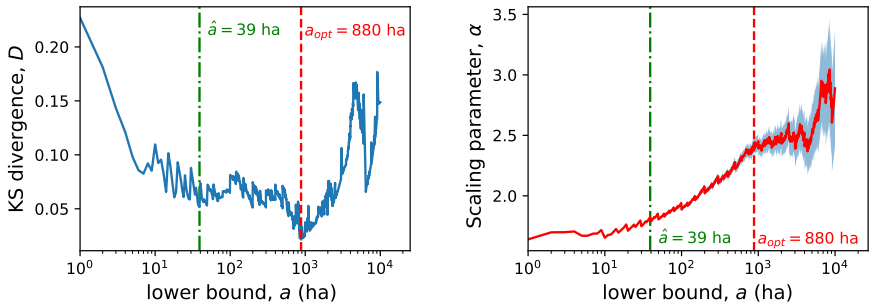
Assign the lower bound a The above discussion relays on the choice of a, but I still to assess a clear way on how to compute it. As suggested by [2] I compute α at different value of a, than I consider the value a that minimize the KL divergence, this is obtained at a_{opt}= 880 ha. This choice is confirmed by the trend of the parameter α with respect to a. One observe that α_e stabilises after a_{opt}. On the other hand,this mathematical result is not supported by any physical intuition from the real problem, but I believe that a reasonable estimate of a should be less than 100 ha, hence I consider the minimum in 1< a <100 ha and the result is $\hat{a}$ = 39 ha. Moreover, with this choice I expect that the rounding-off bias at low values of A_F has not severe effects on the estimation.

I plot the cumulative frequency distribution, $F(A_F)$ and non-cumulative frequency density $f(A_F)$ for the total area burnt A_F in Figure 4. The Scaling parameter $\alpha_e = 1.83 \pm 0.01$ is large enough respect to 1 that the cumulative distribution does not show a problematic behaviour at large values of area burnt.

3.2 Estimation of the size of a fire for a 100 year return period

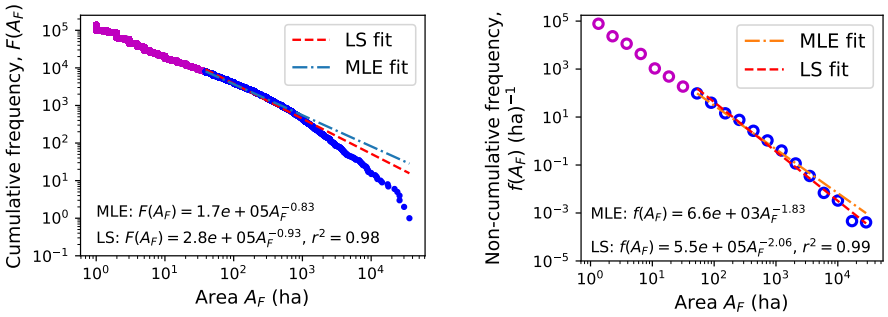
I want to determine the size of a wildfire, A_{100}, whose average recurrence time is 100 years, therefore the probability that A_{100} happens at least one in a year is $Prob_1[A_{100}] = 1/100$ and the complementary event is $Prob_1[\bar{A}_{100}] = 0.99$. The entire data interval of the fitted model is 25.8 years . The probability not to have any event of size A_{100} in 25.8 years is

$$Prob_{25.8}[\bar{A}_{100} = 0.99^{25.8} \approx 0.772$$



(a) **(b)**

Figure 3: 3a Plot of the Kolmogorov-Smirnov distance D with respect to lower bound a. The global minimum of D is observed at $a_{opt} = 880$ (ha) (dashed line), while in the interval $1 < a < 100$ ha the local minimum is at $\hat{a} = 39$ (dash dotted line) ha. 3b Plot of scaling parameter α with respect to the lower bound. Global and local minimum, a_{opt} and $\hat{a}$ represented as before. The shaded area represent the error on α computed using the MLE estimation, σ_e see Eq. (3). Data from [9].



(a) **(b)**

Figure 4: 4a Cumulative frequency distribution $F(A_F)$ of the total area burnt A_F. Blue points have been used to produce the fit, whereas magenta points have not been used for the fit. Fit region is $A_F > 39$ ha. Least square (LS) fit (dot-dashed line) performed using a linear interpolation to $\log(F(A_F) = -\alpha \log(A_F) + B$. Maximum likelihood estimates(MLE) $\alpha_e = 1.83 \pm 0.01$ (dashed line) is computed from Eq.(2), the fitting function is given from the CDF Eq. (1), multiplied by the number of data points in the fitting region. Kolmogorov-Smirnov divergence, $D = 0.051$. 4b. Non cumulative Frequency density, $f(A_F)$ of the total area burnt A_F. Bin size δA_F increase with A_F such that the bin width is almost the same with logarithm coordinates. LS and MLE parameters computed as in 4a. Data from [9].

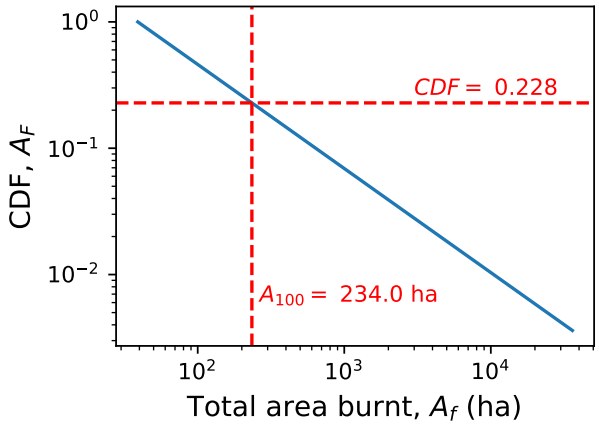


Figure 5: Cumulative density function CDF,$P(A_F)$ of the total area burnt A_F. Graphical estimation of the area A_{100} that have a recurrence time of 100 years. A_{100} is the value such that $CDF(A_{100}) = 1 - 0.99^t$ where $t = 25.8$ is the temporal range in years of data used to build the model. Model parameters are $\alpha_e = 1.83 \pm 0.01$, $a = 39$ ha.Data from [9].

The probability to have at least one event of size A_{100} is :

$$Prob_{25.8}[A_x] = 0.228 \tag{4}$$

From the CDF of Eq. (1) with model parameters $a = 39\,\text{ha}$, $\alpha_e = 1.83 \pm 0.01$ one can invert and obtain:

$$A_{100} = a * Prob_{25.8}[A_{100}]^{\frac{1}{1-\alpha}}$$

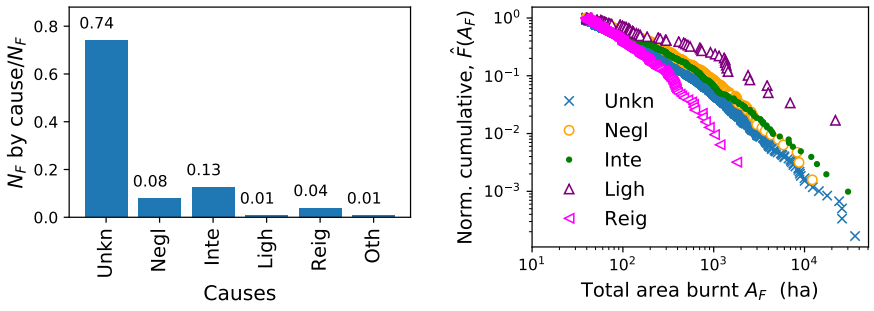
and the extimate of the error over A_{100} is:

$$\sigma_{A_{100}} = \left| \frac{\partial A_{100}}{\partial \alpha} \right| \sigma_e$$

then $A_{100} = 234 \pm 5$ ha

3.3 Causes

In this subsection I use the techniques discussed above to compare the statistical properties of wildfires with respect to their causes. I use the lower bound parameter a obtained above. This analyses presents severe limitations due to non-homogeneous methodology in the way data were collected, as discussed in Section 2. In particular, the largest fraction of data are labelled with "Unknown" cause, but if this kind of fires belonged systematically to other cause, this would bias extensively our results. Moreover, the number of fires in most



(a) **(b)**

Figure 6: 6a Number of fires N_F due to a given cause divided by the total number of fires.The top 5 causes are plotted, namely "Unknown", "Negligence", "Intentional", "Lightning", "Re-ignition". Class "Other" accounts for other causes, therefore the Sum of the height in the bar chart is equal to 1. Only fires with size greater than $a = 39$ ha are considered. $N_F = 8019$ 6b Normalised cumulative distribution of the total area burnt by causes, $\hat{F}(A_F) = Prob[\text{fire greater then} A_F]$. For large A_F the fraction of re-ignition fires is the fastest decreasing, whereas wildfires due to natural causes exhibit the slowest decrease. Intentional and negligence show a similar trend.

of the categories is too low to allow the creation of a good model on top of the data.

In Figure 6 I plot the fraction of fires and the normalised cumulative distribution of the total area burnt, $\hat{F}(A_F)$ with respect to top five causes. Although lighting accounts for only a tiny fraction of fires, the decaying trend is much slower than in the other cases. Therefore, if one is interested in large events, lightning has to be taken into account as a reasonable cause.On the other hand, the number of lightning wildfires is too low to fit a model on the data, hence it represents only a qualitative indication of the trend.

I compare the Unknown and Intentional wildfires, these are the only categories with a large statistics, 5941, 10140 records respectively (Figure 7). Unsurprisingly, the model fitted on top of data labelled as Unknown gives similar results to the case of setting with all data. On the other, the scaling parameter of intentional fires is 1.74 ± 0.02 that is smaller than the same value for the unknown, 1.85 ± 0.01.

4 Conclusion

In this report I show how to build a power law model from the data. This is not straightforward, since one cannot rely on the brute force of computer but the results has to be compatible with real world phenomena.Therefore I present a way to conjugate physical intuition on the problem with mathematical

Table 2: Comparison of the model prediction for two causes: "Intentional", "Unknown". Taking into account errors, different behaviour is observed.

Cause	**Scaling parameter** α	**KS divergence** D	**100 years recurrent event**,A_{100}
Intentional	1.74 ± 0.02	0.057	290 ± 18 ha
Unknown	1.85 ± 0.01	0.052	222 ± 5 ha

results. Finally, the analysis of wildfires with respect to their causes display that classification methodology can lead to counter-intuitive results. In fact, classes that are very unlikely can contribute extensively in the rare events regime if the distribution of these events presents heavier tail respect to more frequent classes.

References

[1] Rémy Chicheportiche and Jean-Philippe Bouchaud. Weighted kolmogorov-smirnov test: Accounting for the tails. *Physical Review E*, 86(4):041115, 2012.

[2] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[3] Defesa da floresta contra incendios.

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

[5] Anne Ganteaume and Marielle Jappiot. What causes large fires in southern france. *Forest Ecology and Management*, 294:76–85, 2013.

[6] Bruce D Malamud. Tails of natural hazards. *Physics World*, 17(8):25, 2004.

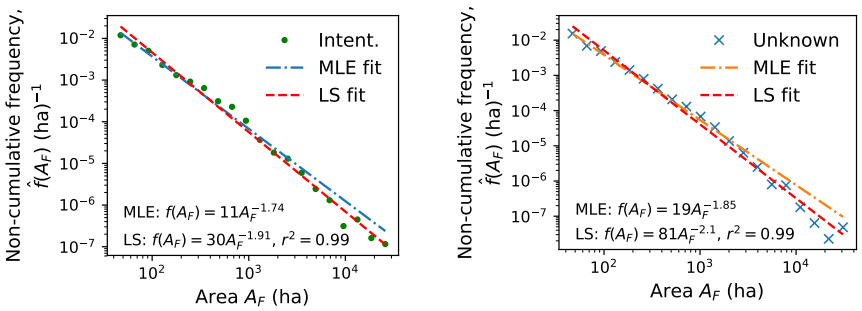
[7] Bruce D Malamud and Donald L Turcotte. Cellular-automata models applied to natural hazards. *Computing in Science & Engineering*, 2(3):42–51, 2000.

[8] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[9] MG Pereira, BD Malamud, RM Trigo, and PI Alves. The history and characteristics of the 1980-2005 portuguese rural fire database. *Natural Hazards and Earth System Sciences*, 11(12):3343, 2011.

[10] Jesús San-Miguel-Ayanz, Tracy Houston Durrant, Roberto Boca, Giorgio Libertá, Francesco Boccacci, Margherita Di Leo, Jorge López Pérez, Ernst Schulte, Abdelhafid Benchikha, Mohamed Abbas, et al. Forest fires in europe, middle east and north africa 2015. *JRC Techincal Report*, 2016.

[11] Ricardo M Trigo, CARLOS C DaCAMARA, et al. Circulation weather types and their influence on the precipitation regime in portugal. *International Journal of Climatology*, 20(13):1559–1581, 2000.



(a) **(b)**

Figure 7: 7a Normalised frequency density $\hat{F}(A_F)$ (number of fires per unit bin divided by the bin size and the total number of fires) for wildfires officially recorded as "Intentional". Fit region is $A_F > 39$ ha. Least square (LS) fit (dot-dashed line) performed using a linear interpolation to $\log(F(A_F) = -\alpha \log(A_F) + B$. Maximum likelihood estimates(MLE) $\alpha_e = 1.74 \pm 0.02$ (dashed line) is computed from Eq.(2), the fitting function is given from the PDF Eq. (1). Kolmogorov-Smirnov distance $D = 0.057$. 7b Normalised frequency density $\hat{F}(A_F)$, for wildfires officially recorded as "Unknown". Same method as above. Maximum likelihood estimates(MLE) $\alpha_e = 1.85 \pm 0.01$, Kolmogorov-Smirnov distance $D = 0.052$.Data from [9].