Forest Science, Vol. 35, No. 2, pp. 319–328.

Do One Percent of Forest Fires Cause Ninety-Nine Percent of the Damage?

DAVID STRAUSS
LARRY BEDNAR
ROMAIN MEES

ABSTRACT. A relatively small number of forest fires are responsible for a very high proportion of the total damage. The proportion due to the fraction p of largest fires, when plotted against p, is a measure of variability of fire sizes that is especially sensitive to the important extreme events. We find the theoretical form of the plot for several commonly used distributions and show how the results can be used in the analysis of empirical plots from actual fire size distributions. Some simple graphical methods are suggested for the fitting of truncated Pareto and lognormal distributions to empirical data. We use the plots to compare the firesize heterogeneity of several fire climate regions of the western United States; the proportion of area burned by the 1% of largest fires ranges from 80%–96%. We also compare patterns of the larger fires for Southern California and Baja California. FOR. SCI. 35(2):319–328.

ADDITIONAL KEY WORDS. Large fires, Pareto distribution, Lorenz curve.

IT IS AN ALL TOO FAMILIAR FACT that a small number of forest fires are responsible for the burning of huge areas of wildlands. The expression "1% of the fires do 99% of the damage" is frequently heard among fire management personnel. The amount of damage caused by different fires varies enormously; our purpose in this paper is to present some statistical methods for the measurement and display of this variability.

Quantification of the variability can be helpful in policy decisions. For example, Minnich (1983) argued that, because of differing fire management strategies in the United States and Mexico, the variability of fire sizes in the United States would be greater. Since fire size has an inviolable lower limit, such increased variance probably results from the occurrence of more large fires in the United States. This implies that the frequency of large fires can be reduced by modification of fire management strategies. One method of assessing the variability is to examine the proportion of total damage (in this case burned area) caused by the largest fires. If there is a difference in variability between Mexico and the United States, the proportion of burned acreage due to, say, the largest 1% of fires would be higher in the United States than in Mexico. We shall take up this question in the Applications section. Awareness of characteristic regional variability could aid long-range dispatch of suppression forces among several regions. Similarly, an awareness of the proportion of damages due to fires of different size classes can help in the allocation of prevention and suppression resources between the classes.

The data discussed in this paper are the sizes of fires, not their suppres-

David Strauss is with the Department of Statistics, University of California, Riverside, CA 92521 and Larry Bednar and Romain Mees are with the USDA Forest Fire Laboratory, Riverside, CA 92507. The authors are grateful to Barry Arnold, Merlise Clyde, Gerald Walton, and anonymous referees for a number of helpful suggestions and criticisms. Manuscript received August 18, 1987.

sion costs or associated economic loss. Acreage burned is of course not the only measure of the damage caused by a fire, and in many contexts economic or other measures will be more appropriate. We emphasize that the methods described would apply equally to other measures of damage, and our focus on acreage burned is merely for purposes of exposition.

The statistical issues in treatment of variability in fire damage closely parallel those arising in economics in connection with income distributions. The measurement of income inequality has an extensive literature (see Arnold 1983 for a review). Since Pareto's classic work around 1900, the statistical distribution bearing his name has become the standard for the fitting of empirical income data, and indeed in many other contexts where the distribution is heavily skewed to the right ("heavy-tailed distributions"). We shall borrow some economic terminology here, referring for example to the "degree of inequality" of fire sizes, and use methods somewhat related to those known from economics. It turns out, however, that the inequality in fire sizes is so extreme that even the Pareto distribution is inappropriate to fit the data, and we will be lead to consider a truncated version of the Pareto distribution instead. Further, the preferred method of describing inequality in income distributions, the so-called Lorenz function (see the Theory section), will be less convenient in the context of fire sizes, and we will propose a related measure. This is the *extreme proportion function.* It specifies the proportion of the total acreage burned that is caused by the largest 100 $p\%$ of the fires. The plot of this against p is our primary method for describing fire size inequality.

We note that in the context of fire sizes, just as in economics, one seldom has a random sample from any population. Indeed the various data sets we have considered here represent a complete census of the larger fires in a given region and time period; this, too, is often the case with income data. In such circumstances one is interested in descriptive statistics rather than in questions of statistical inference. The fitting of theoretical distributions to data is then performed as an aid to summarizing the data and should not be accompanied by inferential methods such as significance testing. Consequently some of our fitting methods are informal and chosen mainly for their simplicity. We shall, however, indicate some procedures that may be useful when formal inferential methods are appropriate. Our methods in effect weight the fires according to their size. This seems appropriate in the present context because (1) the larger fires are much more important, and (2) for heavy-tailed distributions one can generally only hope to fit a theoretical distribution to the upper portion of the data (cf. the econometric literature summarized by Arnold 1983).

THEORY

Let x denote the size of a fire, in appropriate units, and suppose that x has a density function $f(x)$ and a distribution function $F(x)$. We define the extreme proportion function, $EP(p)$, to be the proportion of all the area burnt that is attributable to the largest 100 $p\%$ of the fires. These are the fires with areas larger than ξ_{1-p}, the $(1-p)$th quantile of the distribution. We will write $q = 1 - p$. For technical convenience we shall denote constants of proportionality by c, it being understood that the quantities c are not necessarily equal. By absorbing various constants into c we effect a considerable simplification of the algebra; as we shall see, it will be easy to evaluate c whenever required.

The proportion of burnt area due to fires of sizes between x and $x + dx$ can be seen to be $cxf(x)dx$. Thus

$$EP(p) = c\int_{\xi_q}^{\infty} xf(x)dx, \tag{1}$$

provided that the integral exists. Where necessary one may evaluate c by noting that $EP(1) = 1$. By expressing EP as a function of p rather than x we make it invariant under changes of units, as can easily be verified; this scale invariance simplifies formulae and is convenient when we want to compare different populations.

Consider the Pareto distribution with distribution function

$$F(x) = 1 - (x/\sigma)^{-\alpha}, \qquad x > \sigma, \tag{2}$$

where $\sigma > 0$ is a scale parameter and $\alpha > 0$ determines the "weight" of the tail; smaller values of α correspond to heavier tailed distributions. As noted previously, the Pareto is widely used as a model for heavy-tailed data. For the fire size data, however, it is preferable to consider a truncated Pareto distribution

$$F(x) = \begin{cases} \dfrac{1 - (x/\sigma)^{-\alpha}}{1 - b^{-\alpha}} & \sigma < x \leq b\sigma \\ 1 & b\sigma < x. \end{cases} \tag{3}$$

This corresponds to the placing of an upper bound $b\sigma$ on the potential size of fires; within the range $(\sigma, b\sigma)$ the density function for (3) is proportional to the untruncated case (2). The parameter b may be regarded as the upper bound measured in units of σ.

We now find the EP function for (3). Because EP is scale invariant we may take $\sigma = 1$ without affecting the result. From (3), the probability density function is $f(x) = cx^{-\alpha-1}$ for x in $(\sigma, b\sigma)$ and $f(x) = 0$ otherwise. Substitution into (1) gives

$$\begin{aligned} EP(p) &= c\int_{\xi_q}^{b} x \cdot x^{-\alpha-1}\, dx \\ &= c(b^{1-\alpha} - \xi_q^{1-\alpha}) \end{aligned} \tag{4}$$

provided that $\alpha \neq 1$, which will be assumed here. Now from (3),

$$F(\xi_q) = q = \frac{1 - \xi_q^{-\alpha}}{1 - b^{-\alpha}} \tag{5}$$

Substituting for ξ from (5) into (4), simplifying, and using the fact that $EP(1) = 1$, we obtain

$$EP(p) = \frac{[pb^{\alpha} + q]^{1-1/\alpha} - 1}{b^{\alpha-1} - 1}, \quad b > 1,\ \alpha > 0,\ \alpha \neq 1. \tag{6}$$

The ratio b of greatest to least possible size of forest fires is exceedingly large, and it is natural to consider the limiting value of (6) as b tends to infinity. In many applications α will exceed 1; with income data, for example, $\alpha \approx 1.5$ is typical. In the case $\alpha > 1$ we obtain the limit

$$EP(p) = p^{1-1/\alpha} \qquad 0 \leq p \leq 1, \alpha > 1 \tag{7}$$

for the regular Pareto distribution (2). Forest fire sizes, however, turn out to have such a heavy-tailed distribution that values of α less than 1 are typical. For example, Robertson (1972) fitted a Pareto to the size distribution of the 20,000 largest Southern California fires occurring in a 10-year period. He concluded that a reasonable fit could be obtained with $\alpha \approx 0.5$. But when α is less than 1, (6) tends to 1 for *all* p as b tends to infinity. The interpretation is that for such values of α the Pareto distribution (2) is so heavy tailed that it predicts 100% of the damage to be due to fires larger than *any* specified threshold. This is clearly unacceptable. The only conclusion is that standard Pareto (2) cannot adequately fit the distribution of fire sizes, at least with respect to the EP function.

The same calculations leading to (6) can be performed for other distributions; several cases are shown in Table 1. In addition to the Pareto, the lognormal distribution is also widely used for right-skewed data and is worth considering in the fire context. One can use the *EP* functions in Table 1 to obtain simple estimators for the parameters of a distribution and as a quick check on whether a given distributional form is reasonable for the data. We shall see some examples in the next section. Note that *EP* is related to the Lorenz curve $L(u)$ widely used to describe income inequality. In fact $L(u) = 1 - EP(1 - u)$, the proportion due to the fraction u of the population with the *lowest* income. In our context, where the upper tail is of primary interest and that tail is exceptionally heavy, *EP* is the more useful of the two.

We note finally that *EP* is easily modified for the case where size data are only available for the largest observations; this is quite common with fire data. Let S denote the fires on which data are available, and suppose these largest fires comprise a proportion π of the total number. Write $EP^*(p)$ for the fraction of area burned by the fraction p of largest fires, both fractions here being computed with respect to S only. Then

$$EP^*(p) = \frac{\text{acreage burned by fraction } \pi p \text{ of largest fires}}{\text{acreage burned by fraction } \pi \text{ of largest fires}}.$$

If we divide numerator and denominator by the total acreage burned, we obtain $EP(\pi p)$ and $EP(\pi)$ respectively. Thus

$$EP^*(p) = EP(\pi p)/EP(\pi), \qquad 0 < \pi < 1. \tag{8}$$

For example, in the case of Pareto with $\alpha > 1$ we find from (7) that

$$EP^*(p) = EP(p). \tag{9}$$

To state it in words: for the Pareto case the proportion due to the fraction p of largest fires is unchanged if we restrict attention to fires of size above an arbitrary threshold. We note that although equations (8) or (9) enable one to compare two "left-censored" distributions, the practical interpretation of the results becomes difficult if the two censoring points are unequal. In that case the comparison is perhaps not to be recommended.

APPLICATIONS

Comparison of Some National Forest Climate Regions

Figure 1 shows the plot of $EP(p)$ against p for three fire climate regions (Schroeder et al. 1964) in the western United States. The data for each fire

TABLE 1. Extreme proportion (EP) due to the fraction p of largest fires

Distribution	Definition	*EP(p)*
Pareto	$F(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}, x > \sigma, \alpha > 0$	$p^{1-1/\alpha}$ if $\alpha > 1$ 1 if $\alpha \leq 1$
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right\}, x > 0$	$1 - \Phi\{\Phi\ (q) - \sigma\}$*
Truncated Pareto	$F(x) \propto 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}, \quad \sigma < x < b\sigma, \ \alpha > 0, \alpha \neq 1, \ b > 1$	$\frac{(pb^\alpha + q)^{1-1/\alpha} - 1}{b^{\alpha-1} - 1}$
Half normal	$Y \quad N(0,\sigma^2); X = \lvert Y \rvert$	$p\left\{1 + \sqrt{\frac{\pi}{2}}\,\Phi^{-1}\left(1 - p/2\right)\right\}$*
Exponential	$f(x) = \lambda e^{-\lambda x}, x > 0$ $= 0$ otherwise	$p(1 - \log p)$
Uniform	$f(x) = \frac{1}{\theta}, 0 < x < \theta$ $= 0$ otherwise	$p(2 - p)$

* *Note:* $\Phi(t)$ is the distribution function of the standard normal.

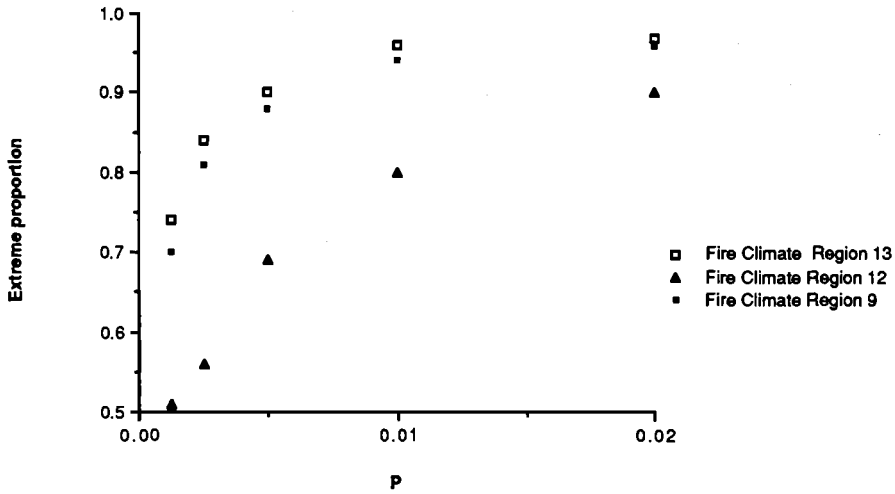

FIGURE 1. Plot of the extreme proportion against p, for three climate regions of the National Forest System (1970–1984). Region 9 is the area east of the Cascades, west of the Rockies. Region 12 is coastal Washington and Oregon. Region 13 is northern California.

climate region consists of a complete census of fires reported in USDA Forest Service jurisdiction during the period 1970–1984, except that sizes less than 0.25 ac are recorded as zero. The number of fires in the regions range from about 6,000 to 36,500. As can be seen, it is only a slight overstatement to say that 1% of the fires are responsible for 99% of the burned area; the actual proportions are 96%, 94%, and 80% for regions 13, 9, and 12 respectively. Note that the curves do not cross, at least away from the vicinity of $p = 0$. This means we have a clear ordering of the three regions with respect to inequality of fire sizes, with fire climate region 13 showing the greatest degree of inequality.

The *EP* function can be used to examine the fit of various distributions to the size data. Consider, for example, a Pareto distribution, with $\alpha > 1$ so as to give a sensible form for *EP*. It follows from (7) that

$$\log[EP(p)] = (1 - 1/\alpha)\log(p).$$

The actual plots of $\log[EP(p)]$ against $\log(p)$, not shown here, do not resemble straight lines; they are markedly concave. Thus the Pareto distribution does not provide a fit that is acceptable for our present purposes.

It is perhaps worth comparing this procedure with more traditional graphical methods for the Pareto. The most common is a plot of $\log[1 - F(x)]$ against $\log x$, which will be a straight line if model (2) holds. One typically finds with fire data that such plots are in fact reasonably linear, at least away from the extreme upper tail region (cf. Robertson 1972). For our purposes, the drawback of such plots is that they give insufficient weight to the important largest fires.

The lognormal distribution is perhaps the chief alternative to the Pareto in the fitting of heavy-tailed data. Most methods for examination of the lognormal fit test for the normality of the log-transformed data. In the present context, where the extreme fires are of special importance, use of the *EP* function may be preferable. A quick, if rather crude, check for lognormality is obtained if one rewrites the expression for the lognormal *EP* in Table 1 as

$$\sigma = \Phi^{-1}(1 - p) + \Phi^{-1}[EP(p)].$$

where Φ^{-1} is the inverse distribution function for the standard normal. If the expression on the right-hand side were indeed approximately constant for different choices of p, this would provide a rough estimate of σ and support the log-normal model. For the data sets of the present example the expressions tend to increase systematically with p. This suggests that our data have too heavy a right tail for the lognormal. Of course this might not be the case if economic cost or other measures of damage are employed.

As noted previously, a truncated Pareto model is worth considering. Formal maximum likelihood estimation of the three parameters is algebraically very cumbersome because the data are censored from below: in our data, the fires of size less than 0.25 ac—which are the great majority—do not have their exact size recorded. Moreover, such an estimation again would suffer from the undue weight placed on the less important small fires. In the present example, where the data cannot be regarded as a random sample, formal estimation methods are inappropriate in any event, and a simple ad hoc method may be desirable. Again, the *EP* function may be used. As an example, we fitted a truncated Pareto to the data from fire climate region 9. The theoretical form of the *EP* function is given in Table 1, the shape parameter and ratio of largest to smallest fires being α and b respectively. Using a nonlinear least squares procedure in the SAS statistical package, we found that the values giving the best fit to the empirical *EP* curve are $\alpha = 0.90$ and $b = 700{,}000$. Figure 2 shows the empirical and fitted truncated Pareto *EP* curves.

In cases where the data can be legitimately regarded as a random sample from some population one may wish to assign confidence intervals to the estimates of $EP(p)$ for selected values of p. To do this one could assume a given distributional form for $F(x)$, estimate its parameters and the desired quantiles, and estimate their reliability using standard inferential methods (Hogg and Craig 1978, Chapter 11). Such a procedure will be tedious to

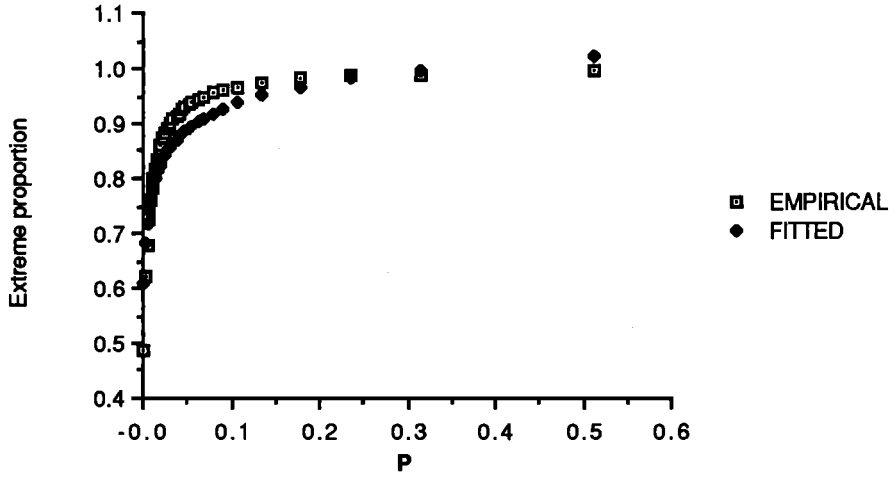

FIGURE 2. Theoretical *EP* values (diamonds) for climate region 9, based on a truncated Pareto distribution with $\alpha = 0.90$, and $b = 70{,}000$ (see text). Squares denote the actual *EP* values for various values of p.

TABLE 2. Frequency of fire events by size class. From Minnich (1983)

Observation period	Size Class (thousands ha)										
	0.04 to 0.1	0.1 to 0.2	0.1 to 0.4	0.4 to 0.8	0.8 to 1.6	1.6 to 3.2	3.2 to 6.4	6.4 to 12.8	12.8 to 25.6	25.6 to 51.2	>51.2
					Baja California						
1972–1980	167	84	61	29	19	17	4	4	1	0	0
					Southern California						
1971–1980	110	57	38	19	26	16	9	7	2	0	0

carry out and the results highly sensitive to the assumed distributional model. A simpler method, and one that makes no distributional assumptions about $F(x)$, is the bootstrap (see, for example, Efron and Gong 1983). We illustrate this with results obtained by bootstrapping $EP(0.01)$ using the data from fire climate region 9. The data contained records for 36,401 fires during the period 1970–1984, and $EP(0.01)$ was computed for 500 independent samples of 36,401 fires drawn randomly and with replacement from this data. The mean of these 500 values was 0.94, and the sample standard deviation was 0.010. Using a normal approximation, we obtain (0.92, 0.96) as an approximate 95% confidence interval for the true $EP(0.01)$.

Comparison of Pattern of Large Fires for Baja California and Southern California

Table 2 was extracted from Minnich (1983), who argued that the fire management policies adopted in southern California resulted in a more unequal distribution of fire sizes than obtained in Mexico. Data for Mexico was only available for the period 1972–1980, and thus here we have only taken southern California data for the period 1971–1980, even though information for earlier decades is available. We have taken the sizes of fires in a given class to be the geometric mean of the end points. Minnich does not give the numbers of fires of size less than 40 ha in his two data sets, and thus we can only compare the regions with respect to fires larger than this.

Figure 3 shows the plots of $EP(p)$ against p, each proportion being for fires larger than 40 ha. Both curves are constrained to pass through (0,0) and (1,1). The curves are very similar, suggesting that for Minnich's 1970s data the degrees of size inequality of large fires are in fact much the same in both regions. It must be pointed out that the pattern for southern California in some earlier decades showed much greater inequality, with several fires larger than 25,000 ha; evidently the great heterogeneity over time makes the drawing of firm conclusions difficult. The inherent instability of extreme

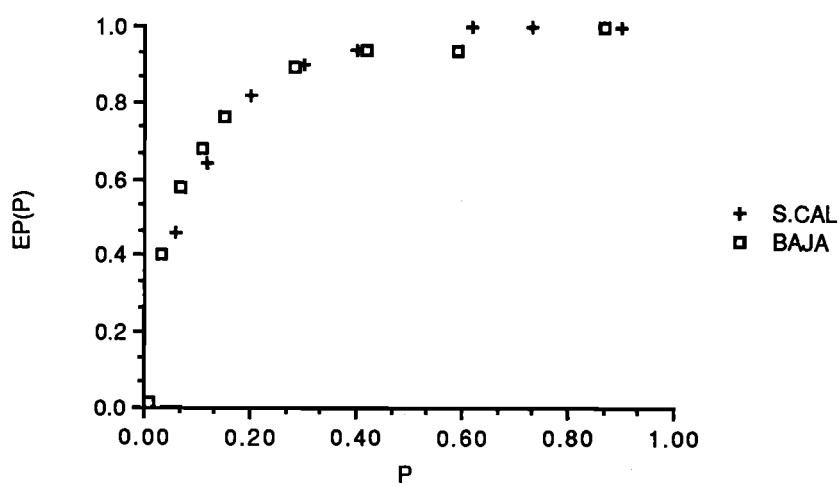

FIGURE 3. *EP* functions for fires larger than 40 ha, during the 1970s for southern California (one symbol) and for Baja California (other symbol). Data from Minnich (1983).

events such as wildfires surely contributes to this heterogeneity and should lead to cautious interpretations on the part of researchers.

LITERATURE CITED

ARNOLD, B. C. 1983. Pareto distributions. Internat. Coop. Publ. House, Fairland, MD.

EFFRON, B., and G. GONG. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37:36–48.

HOGG, R. V., and R. T. CRAIG. 1978. Introduction to mathematical statistics. Ed. 4. Macmillan, New York.

MINNICH, R. A. 1983. Fire measures in Southern California and Northern California. Science 219:1287–1294.

ROBERTSON, C. A. 1972. Analysis of forest fire data in California. Tech. Rep. No. 11, Dep. Stat., Univ. of California, Riverside.

SCHROEDER, M. J., et al. 1964. Synoptic weather types associated with critical fire weather. Pac. Southwest For. & Range Exp. Stn. 492 p. Available from NCIS, #AD-449 630/3.