On the Distribution of Wildfire Sizes

Frederic Paik Schoenberg[1] Roger Peng[1] James Woods[2]

[1] Department of Statistics, University of California Los Angeles, Los Angeles, CA 90095–1554.

[2] Department of Geography, California State University Long Beach, Long Beach, CA 90840.

Corresponding author: Frederic Paik Schoenberg

phone: 310-794-5193

fax: 310-472-3984

email: frederic@stat.ucla.edu

Postal address: UCLA Dept. of Statistics

8142 Math-Science Building

Los Angeles, CA 90095–1554, USA.

Summary.

A variety of models for the wildfire size distribution are examined using data on Los Angeles County wildfires greater than 100 acres between 1950 and 2000. In addition to graphs and likelihood criteria, Kolmogorov-Smirnoff and Cramer-von Mises statistics are used to compare the models. The tapered Pareto distribution appears to fit the data quite well and offers some advantages over the untapered Pareto distribution, while alternatives including the lognormal, half-normal, exponential, and extremal distributions fit poorly. The size distribution appears to be quite stable over the examination period, though inspection of the transformed wildfire sizes for the tapered Pareto reveals some limited trend in the residuals, indicating a very slight gradual decrease in the average fire size in Los Angeles County over these fifty years.

Key words: wildfires, Pareto, tapered Pareto.

1 Introduction.

A problem fundamental to the analysis of wildfires is the characterization of the wildfire size distribution. A variety of models for this distribution have been prescribed, including different forms of the Pareto (Robertson, 1972; Montroll and Shlesinger, 1982; Strauss et al., 1989; Alvarado et al., 1998), extremal (Moritz, 1997; Alvarado et al., 1998), and other distributions such as the lognormal, half-normal, exponential, and uniform (e.g. Strauss et al., 1989).

The relative fits of these models to data have not been adequately scrutinized to date. Fortunately, wildfire recordings from Los Angeles County, California comprise a sufficiently rich and complete dataset enabling the discrimination between these models based on goodness-of-fit.

Fire histories based on burn maps on file at Los Angeles County Department of Public Works (LACDPW), Watershed Management Unit include fires ranging from 1 acre to greater than 50,000 acres, and date back as far as 1877. The dataset, in its complete form, is a compilation of several thousand wildfire recordings from various recording agencies, including LACDPW, the Los Angeles County Fire Department (LACFD), the Santa Monica Mountains National Recreation Area, the Ventura County Fire Department (VCFD), and the California Department of Forestry and Fire Protection. Data on most of the fires were initially mapped by LACFD officials in paper format and subsequently digitized using a geographic information system and catalogued by LACDPW.

There are several different possible measures of a wildfire's size, including the area it burns, its maximum temperature, the amount of environmental or monetary damage it

causes, etc. Our analysis focuses exclusively on burn area. While the LACDPW catalog contains brush fires of all different sizes, some as small as 1 acre, it certainly is not contain all small fires. For instance, LACFD and VCFD maps contain only "significant" fires, i.e. fires burning at least 100 acres or where significant structures are damaged or destroyed. LACFD officials informed us that they believe that the catalog contains *all* major fires (greater than 100 acres) since 1950.

The structure of this paper is as follows. In Section 2, we outline various models for the wildfire size distribution. Comparison of the fit of the models is presented in Section 3. Section 4 investigates the stationarity of the wildfire size distribution over time, and conclusions and areas for further analysis are discussed in Section 5.

2 Models.

Numerous models have been used to describe the distribution of wildfire sizes. For instance, Strauss et al. (1989) list several possibilities, including the lognormal, half-normal, and exponential distributions. Yet another set of alternatives are the extremal distributions, especially the Frechet distribution, which were originally used to describe the distributions of minima and maxima (Johnson et al. 1995) but are fit to wildfire areas by Moritz (1997) and Alvarado et al. (1998).

The model most commonly used for the size distribution of wildfires seems to be the Pareto. The large tails of the wildfire distribution have been widely observed (e.g. Bruce, 1963; Johnson and Gutsell, 1994). as has the power-law decrease in frequency with area,

suggesting the Pareto distribution:

$$F(x) = 1 - (a/x)^{\beta}, a \leq x < \infty. \tag{1}$$

The parameter β in the Pareto model is the slope of the decrease in the survivor function $1 - F(x)$ with x, when plotted on log-scale. The lower truncation point a, sometimes called the *completeness threshold*, represents the lower limit on the sensitivity of the records; typically small wildfires are absent from datasets due to limitations on resources and/or recording instrumentation. For instance, as mentioned in the introduction, the LACDPW dataset is believed to be complete for fires greater than 100 acres (0.15625 square miles) for the period 1950-2000, although there are some observations of fires before 1950 and/or far below this cutoff size. In what follows we shall refer only to the subset of 548 Los Angeles County wildfires since 1950 burning greater than 100 acres.

Observations that the upper tail of the wildfire distribution generally decays to zero more rapidly than the Pareto have led to the use of the truncated Pareto:

$$F_{trunc}(x) = F(x)/F(\gamma), a \leq x < \gamma, \tag{2}$$

Where F is as in (1). Additional justification for this truncated form is based on the desire for a distribution with finite mean and variance (Strauss et al., 1989; Alvarado et al., 1998).

However, one problem with the truncated Pareto is that a hard upper threshold does not agree with basic ecological principles nor with elementary understanding of how fires burn, their interaction with humans, and the error in their recording. The maximum likelihood estimate of γ for the truncated Pareto is the maximum observed fire size in the dataset, and a model which posits a sudden drop to zero in the frequency of observation of fires greater than the biggest observed seems absurdly unrealistic. In addition, the maximum likelihood

estimate of γ is biased downward, and this bias may be substantial even for datasets consisting of thousands of events (Kagan and Schoenberg, 2001). Hence although the truncated Pareto may offer satisfactory empirical fit, its predictive value is limited especially at the upper end of the size spectrum, and it is precisely this portion of the distribution that is typically most of interest in wildfire research.

In the field of seismology, similar considerations led to the use of the tapered Pareto distribution (Jackson and Kagan, 1999; Vere-Jones et al. 2000; Kagan and Schoenberg, 2001). The tapered Pareto, which was originally introduced by Vilfredo Pareto himself (see pp. 305-306 of Pareto, 1897) has cumulative distribution function

$$F_{tap}(x) = 1 - (a/x)^{\beta} exp\left(\frac{a-x}{\theta}\right), a \leq x < \infty. \quad (3)$$

For the density, characteristic function, moments, and other properties of the tapered Pareto see Kagan and Schoenberg (2001). The parameter θ in the tapered Pareto distribution is interpreted as an upper threshold above which the fire frequency begins to decay especially rapidly.

Details on the moments, estimates, and other properties of the distributions listed in this Section are given by Johnson et al. (1995).

3 Estimation and comparison of fit.

We fit each of the models discussed in Section 2 by maximum likelihood. The parameter estimates converged quickly in every case. Both Nelder-Mead and Newton-Raphson optimization procedures were used with various starting values; for every model the differences between parameter estimates across choices of optimization routine and starting values were

Table 1: Models and parameter estimates

Model	$F(x) =$	Estimate (SE)	Estimate (SE)
Lognormal	$\Phi\left(\frac{log(x)-\mu}{\sigma}\right)$	$\hat{\mu} = -0.216(0.0608)$	$\hat{\sigma} = 1.42(0.0430)$
Half-normal	$2\Phi\left(x/\sigma\right) - 1$	$\hat{\sigma} = 6.16 \ \ (0.132)$	
Exponential	$1 - exp(-\lambda x)$	$\hat{\lambda} = 0.317(0.0135)$	
Frechet	$exp\left[-\left(\frac{\delta}{x-a}\right)^{\beta}\right]$	$\hat{\delta} = 0.140(0.0140)$	$\hat{\beta} = 0.454(0.0133)$
Pareto	$1 - (\frac{a}{x})^{\beta}$	$\hat{\beta} = 0.610(0.0260)$	
Truncated Pareto	$\left(1 - (\frac{a}{x})^{\beta}\right) / \left(1 - (\frac{a}{\gamma})^{\beta}\right)$	$\hat{\beta} = 0.532(0.0320)$	$\hat{\gamma} = 75.8(5.03)$
Tapered Pareto	$1 - (\frac{a}{x})^{\beta} \, exp\left(\frac{a-x}{\theta}\right)$	$\hat{\beta} = 0.548(0.0277)$	$\hat{\theta} = 29.8(7.26)$

imperceptible. Table 1 lists the models, their cumulative distribution functions $F(x)$, and the maximum likelihood parameter estimates. Data used to fit the models were in units of square miles. In Table 1, Φ denotes the standard normal cumulative distribution function. Standard errors (SEs) of the parameter estimates given in parentheses in Table 1 were obtained via the diagonal of the inverse of the Hessian of the log-likelihood function (for details see e.g. Chapter 13.8 of Wilks, 1962). Alternative standard errors were obtained by Monte Carlo simulation; in each case the results were rather similar to those shown in Table 1. For instance, for the tapered Pareto distribution the standard errors for β and θ based on simulation were 0.0288 and 7.80, respectively, whereas those based on the Hessian were 0.0277 and 7.26.

Figure 1 presents the empirical survivor function and estimated survivor functions based on the various models. It is immediately evident that the lognormal, half-normal, and ex-

Table 2: Goodness-of-fit Statistics

Model	AIC	K-S (p-value)	C-vM (p-value)
Lognormal	857.2	2.97 ($p < .0001$)	0.0470 ($p = 0.0003$)
Half-normal	2666	12.9 ($p < .0001$)	1.08 ($p < 0.0001$)
Exponential	1180	9.18 ($p < .0001$)	0.353 ($p < 0.0001$)
Frechet	755.3	1.92 ($p = 0.0017$)	0.234 ($p < 0.0001$)
Pareto	703.0	0.812 ($p = 0.490$)	0.0223 ($p = 0.0043$)
Truncated Pareto	689.1	0.387 ($p = 0.997$)	0.00138 ($p = 0.876$)
Tapered Pareto	691.0	0.467 ($p = 0.973$)	0.00269 ($p = 0.602$)

ponential models fit very poorly. The Frechet and Pareto models approximate the empirical size distribution closely for small areas but fit poorly for large areas, and the tapered Pareto and truncated Pareto models appear to fit quite well.

These observations are confirmed by the goodness-of-fit statistics presented in Table 2. The Akaike Information Criterion (AIC) is a likelihood-based statistic which includes a penalty for each parameter in a model (Akaike, 1983). When comparing the performance of models with varying numbers of parameters, the AIC is useful for avoiding overfitting. The AIC is computed as twice the negative loglikelihood plus twice the number of estimated parameters; lower AIC indicates better fit. We also computed Kolmogorov-Smirnoff (K-S) and Cramer-von Mises (C-vM) test statistics and corresponding p-values for each model; these represent the maximum difference and integrated squared difference, respectively, between the fitted distribution and the empirical distribution. The poor fit of all the models except the tapered Pareto and truncated Pareto is clearly exhibited.

The K-S and C-vM tests exploit the deviations in the fitted distribution functions from the empirical distribution function. The test statistics are, respectively,

$$\sqrt{n}\sup_x |\hat{F}(x) - F_n(x)|, \tag{4}$$

and

$$\int_{-\infty}^{\infty} \left(\hat{F}(x) - F_n(x)\right)^2, \tag{5}$$

where $\hat{F}$ and F_n denote the estimated and empirical distribution functions. Deviations between these two functions can also be represented graphically by Q-Q plots. Such plots, along with 95%-confidence bands based on Monte Carlo simulation, are shown in Figures 2 and 3 for the tapered Pareto and truncated Pareto distributions, respectively. For comparison, a Q-Q plot for the Pareto distribution is shown in Figure 4: the poor fit at the high end of the wildfire area spectrum is immediately evident. The Q-Q plots for the lognormal, half-normal, exponential and extremal distributions (not shown) displayed similarly pronounced (or even more flagrant) departures.

Both the truncated and tapered Pareto models appear to provide satisfactory fit based on Table 2 and Figures 1, 2 and 3. However, as previously noted, the maximum likelihood estimate of the upper cutoff area γ is always biased downward, introducing substantial bias into the estimates of the crucial high end of the area distribution. There is no such systematic problem in the estimation of the tapered Pareto model, for although the maximum likelihood estimate of θ in the tapered Pareto can be biased as well (Kagan and Schoenberg, 2001), the more gradual decline in density $f(x)$ for the tapered Pareto ensures that small deviations in estimates of θ do not result in such substantial differences in $f(x)$ as compared with the truncated form. This and the disagreement of the truncated Pareto model with basic

physical principles as discussed in Section 2 may suggest the use of the tapered Pareto model instead.

4 Inspection of stationarity

In the previous section it was shown that the tapered and truncated Pareto distributions fit well to the size distribution of Los Angeles County wildfires from 1950 to 2000. One may inquire whether the distribution of wildfire sizes is stable over this time period or whether it changes significantly over time.

Figure 5 shows a plot of the transformed wildfire sizes for the estimated tapered Pareto distribution versus time. For each observed wildfire area, x_i, the quantity $U_i = \hat{F}_{tap}(x_i)$ is shown, where $\hat{F}_{tap}$ is the tapered Pareto distribution function with maximum likelihood parameter estimates given in Table 1. If the fitted distribution were correct and the areas were i.i.d. then the U_i would be independent random variables uniformly distributed between 0 and 1. At a glance, Figure 5 shows no flagrant departures from this hypothesis.

The hypothesis that the distribution is stationary over time may be examined with respect to various alternatives. One alternative is that some linear trend exists, and this may be tested simply by regression of U_i against t_i, where t_i is the time of fire i. The results of this regression are shown in Table 3. The downward slope indicates that fires have been getting smaller, on average, as time has progressed. Although the slope of -0.00298 is statistically significantly different from zero, this trend in Figure 5 is less than overwhelming and appears to be largely attributable to the scarcity of very large fires between 1996 and 2000 and the lack of many small fires between 1960 and 1967.

Table 3: Linear regression of U_i on t_i

	Estimate	SE
Intercept	6.39	1.78
Slope	-0.00298	0.000902

A second alternative hypothesis is that the points in Figure 5 are clustered, i.e. that for certain time periods, certain size ranges are especially likely. This hypothesis is relevant to the question of whether or not temporal changes in the fuel age mosaic due to prescribed burning have significantly altered the frequency of certain sizes of wildfires in California, a question discussed in great depth by Keeley and Fotheringham (2001). The L-function described by Ripley (1981) is useful for investigating this hypothesis, and is shown in Figure 6 along with 95% confidence bounds obtained via simulation. For any distance $z > 0$, L is defined via

$$L(z) = \sqrt{\hat{K}(z)/\pi}, \tag{6}$$

where $\hat{K}(z)$ is the boundary-corrected proportion of points in Figure 5 within z of each other; see Chapter 8.3 of Ripley (1981) for details. Note that before calculating $\hat{K}(z)$ the time axis of Figure 5 was rescaled to stretch from 0 to 1 so that the axes would be commensurate. Under the null hypothesis of stationarity and uniformity, the L-function in Figure 6 should approximate the line $L(z) = z$. From Figure 6 general agreement with this hypothesis can be seen. In order for departures to be more easily identified Figure 7 shows $L(z) - z$ versus z; i.e. Figure 7 displays the residuals from Figure 6. From Figure 7 one can see that there is some minimal but statistically significant clustering in Figure 5 for small distances z in the

range 0 to 0.4.

Note that the clustering in the L-function may be due in part to heterogeneity in the *number* of wildfires over time. For instance, examination of Figure 5 reveals some clusters of events in 1980, 1981 and 1984. However, this appears to be due simply to the fact that there are unusually many fires occurring in these years, rather than nonstationarity in the fire size distribution.

5 Conclusions

The distribution of sizes of wildfires recorded in Los Angeles County, California, between 1950 and 2000 is shown to approximate a tapered or truncated Pareto distribution quite closely. The lognormal, half-normal, exponential and extremal distributions fit poorly, and the ordinary Pareto distribution significantly overpredicts the frequency of fires at the high end of the size spectrum.

The distribution of wildfire sizes appears to change little over time, though there is some limited trend corresponding to fewer larger fires in the later part of the 20th century compared to the 1950s and 1960s. Despite some minimal clustering of fire sizes in time, no other substantial signs of non-stationarity in the size distribution are readily apparent.

Further work should be done to verify whether the tapered Pareto model fits well to the distribution of wildfire sizes in other regions. In addition, some modification to the tapered Pareto distribution may be desired to deal with the case where data are available on *some* wildfires below the minimal cutoff size level a and where exclusion of these fires is not desired. For instance, it may be that some form of tapering of the *lower* end of the Pareto distribution

is appropriate for such purposes.

Acknowledgements.

This material is based upon work supported by the National Science Foundation under Grant No. 9978318. The authors thank the LACDPW and LACFD (especially Mike Takeshita and Frank Vidales) for their generosity in sharing their data with us.

References.

Akaike H. 1983. Information measure and model selection. *Bulletin of the International Statistical Institute* **50**(1): 277–291.

Alvarado E, Sandberg D, Pickford S. 1998. Modeling large forest fires as extreme events. *Northwest Science* **72**: 66–75.

Bruce D. 1963. How many fires? *Fire Control Notes* **24**(3): 45–50.

Jackson D, Kagan Y. 1999. Testable earthquake forecasts for 1999. *Seismological Research Letters* **70**(4): 393–403.

Johnson E, Gutsell S. 1994. Fire frequency models, methods and interpretations. *Advances in Ecological Research* **25**: 239–287.

Johnson N, Kotz S, Balakrishnan N. 1995. *Continuous Univariate Distributions*, 2nd edition. Wiley, New York.

Kagan Y, Schoenberg F. 2001. Estimation of the upper cutoff parameter for the tapered Pareto distribution. *Journal of Applied Probability* **38A**, Festschrift for David Vere-Jones, D. Daley, editor: 158–175.

Keeley JE, Fotheringham CJ. 2001. History and management of crown-fire ecosystems: a summary and response. *Conservation Biology* **15**: 1561–1567.

Montroll E, Shlesinger M. 1982. On 1/f noise and other distributions with long tails. *Proceedings of the National Academy of Sciences of the United States of America, Applied*

Mathematical Sciences **79**: 3380–3383.

Moritz M. 1997. Analyzing extreme disturbance events: fire in Los Padres National Forest. *Ecological Applications* **7**(4): 1252–1262.

Pareto V. 1897. *Cours d'Économie Politique*, Tome Second, Lausanne, F. Rouge, quoted by Pareto, V. (1964), Oevres Complètes, Publ. by de Giovanni Busino, Genève, Droz, volume 2.

Ripley B. 1981. *Spatial Statistics.* Wiley, NY.

Robertson C. 1972. *Analysis of forest fire data in California.* Technical Report No. 11, Department of Statistics, University of California, Riverside.

Strauss D, Bednar L, Mees R. 1989. Do one percent of forest fires cause ninety-nine percent of the damage? *Forest Science* **35**(2): 319–328.

Vere-Jones D, Robinson R, Yang W. 2001. Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysics Journal International*, **144**: 517–531.

Wilks S. 1962. *Mathematical Statistics.* Wiley, New York.

List of Figures and their Captions:

Figure 1: Empirical and estimated survivor functions.

Figure 2: Q-Q plot for the tapered Pareto distribution.

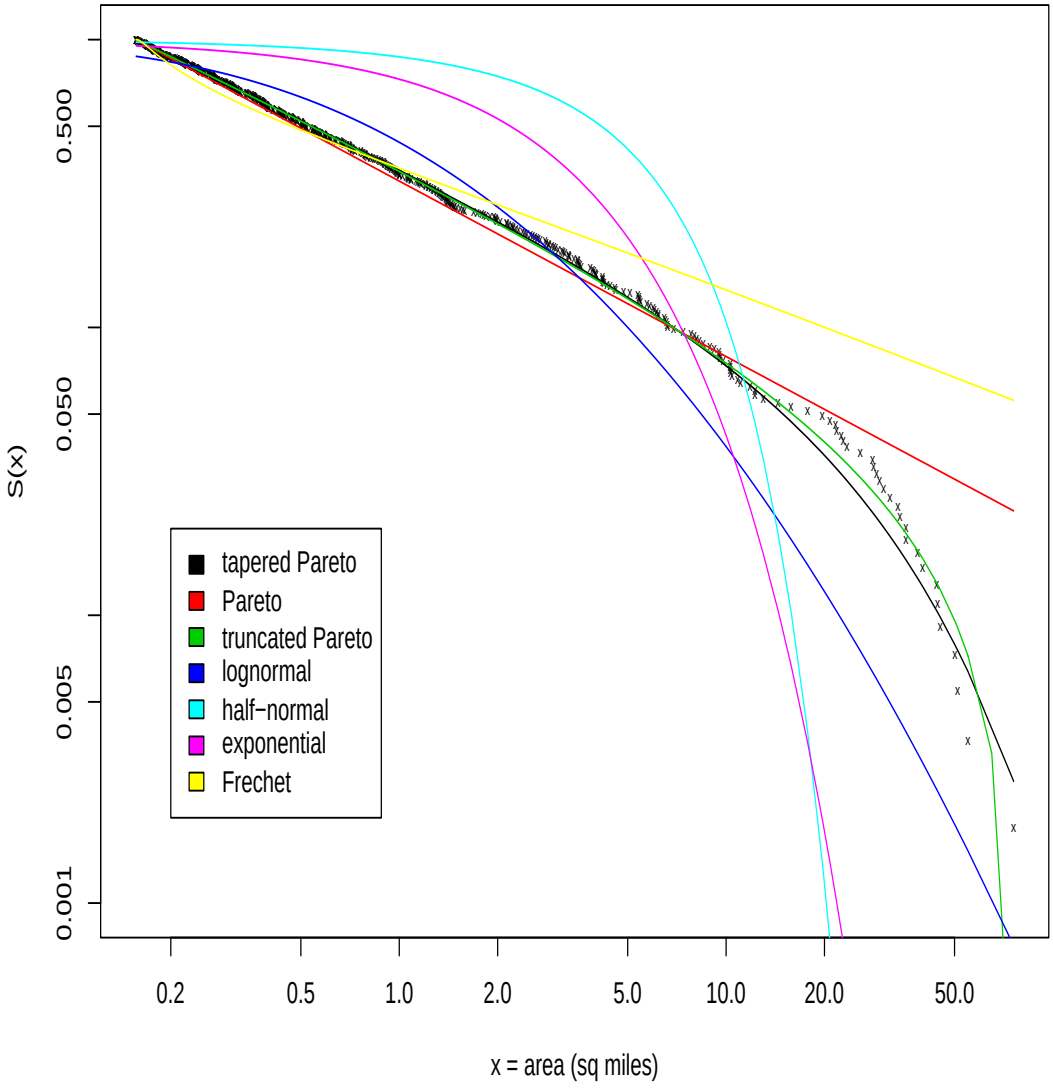
Figure 3: Q-Q plot for the truncated Pareto distribution.

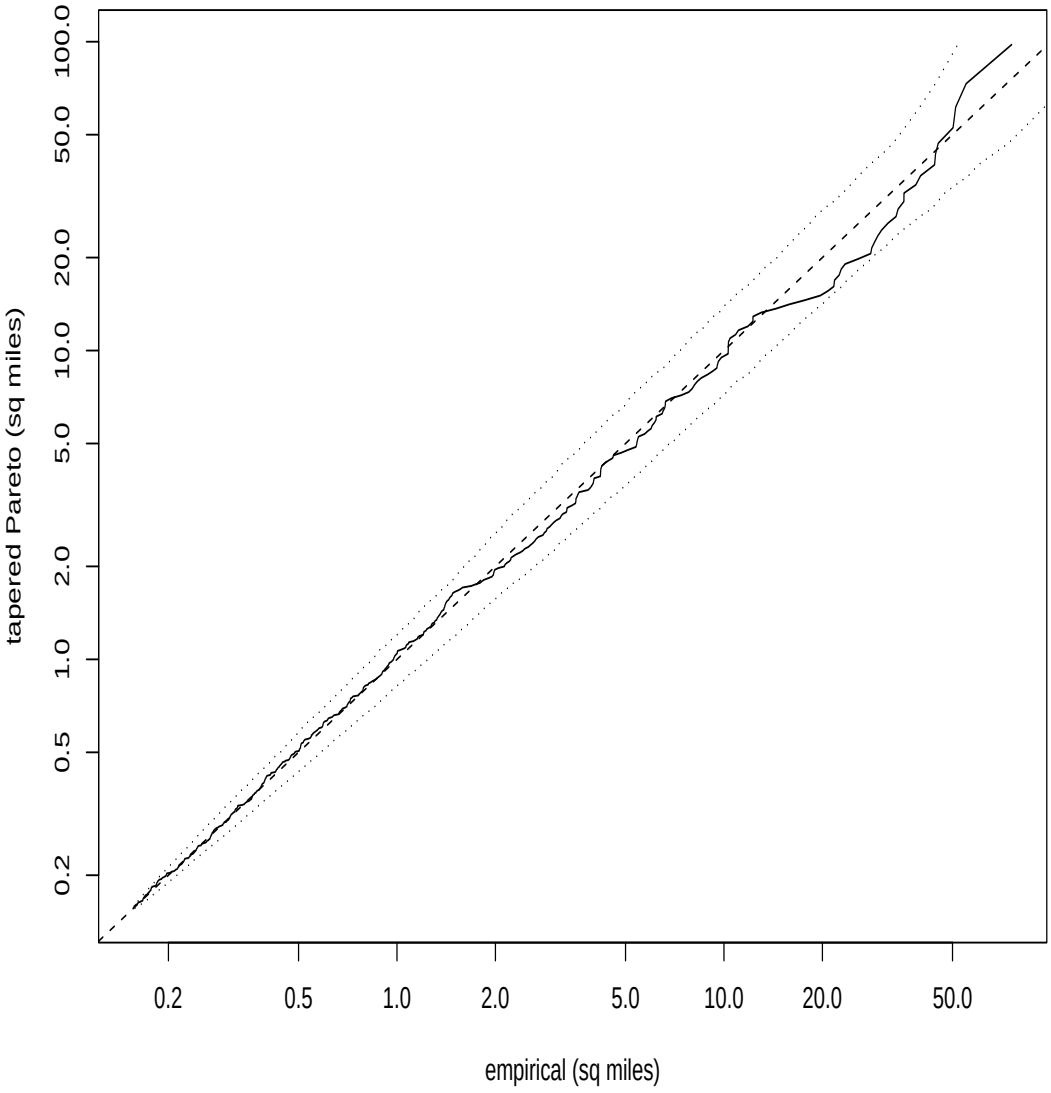
Figure 4: Q-Q plot for the Pareto distribution.

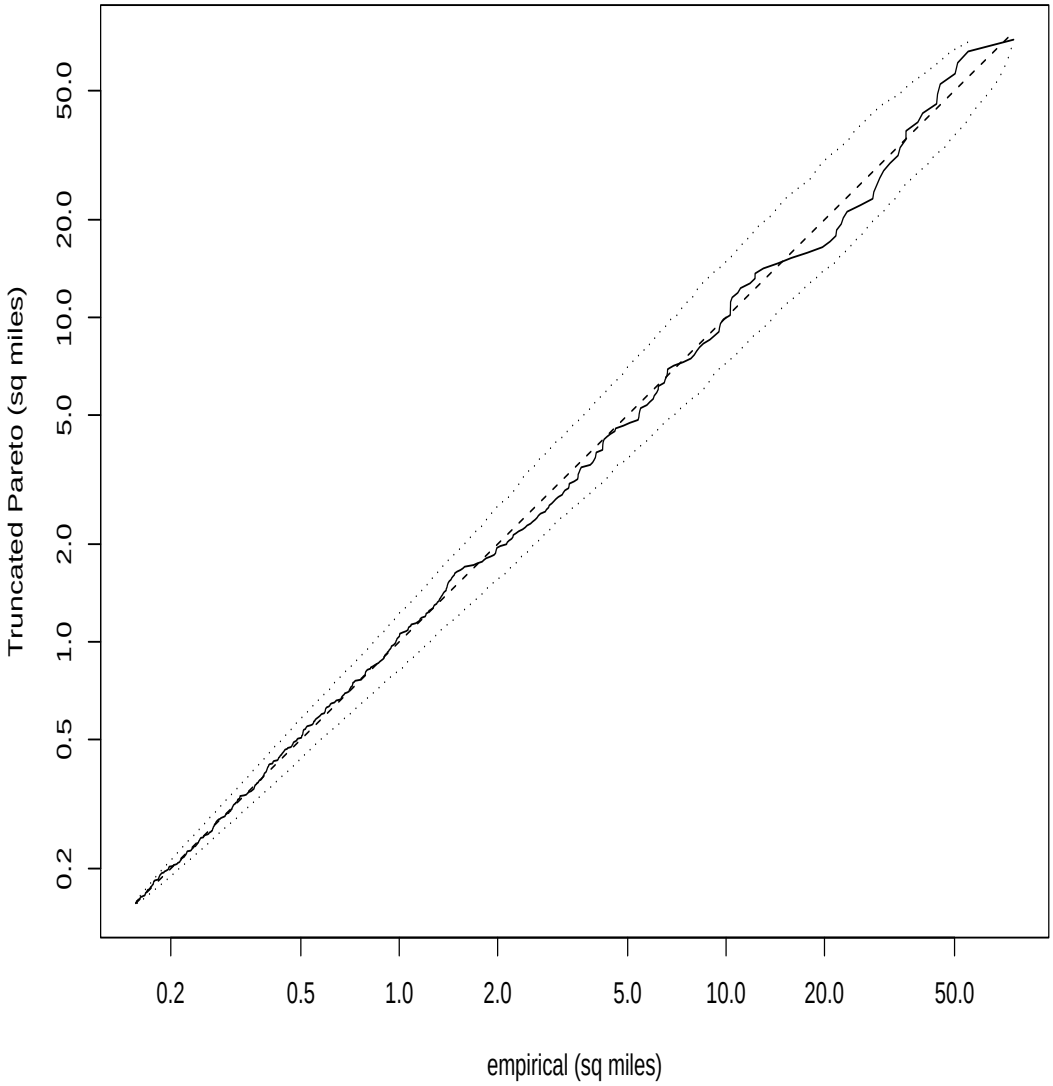
Figure 5: Transformed wildfire sizes based on the tapered Pareto model.

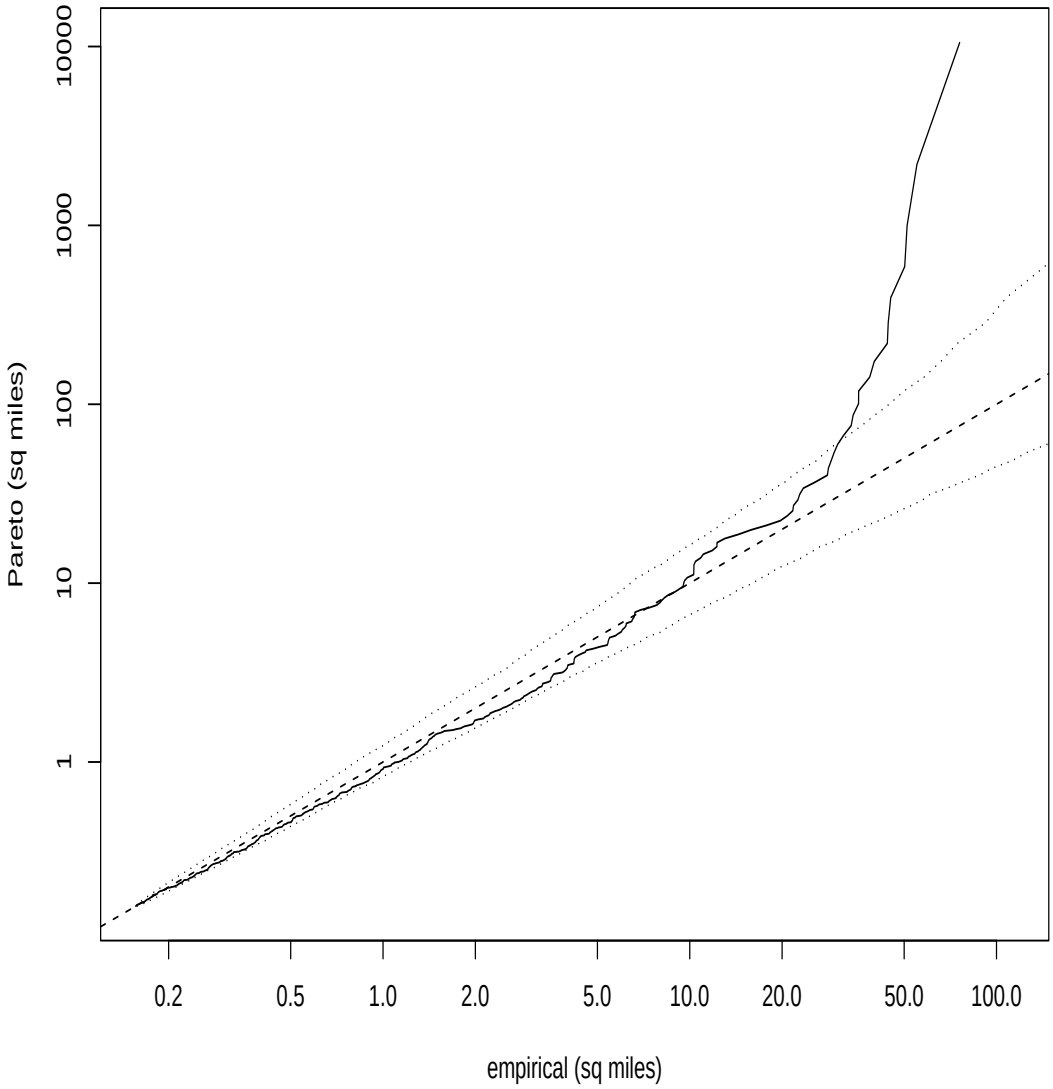
Figure 6: L-function applied to transformed wildfire sizes U_i, transformed according to the tapered Pareto distribution.

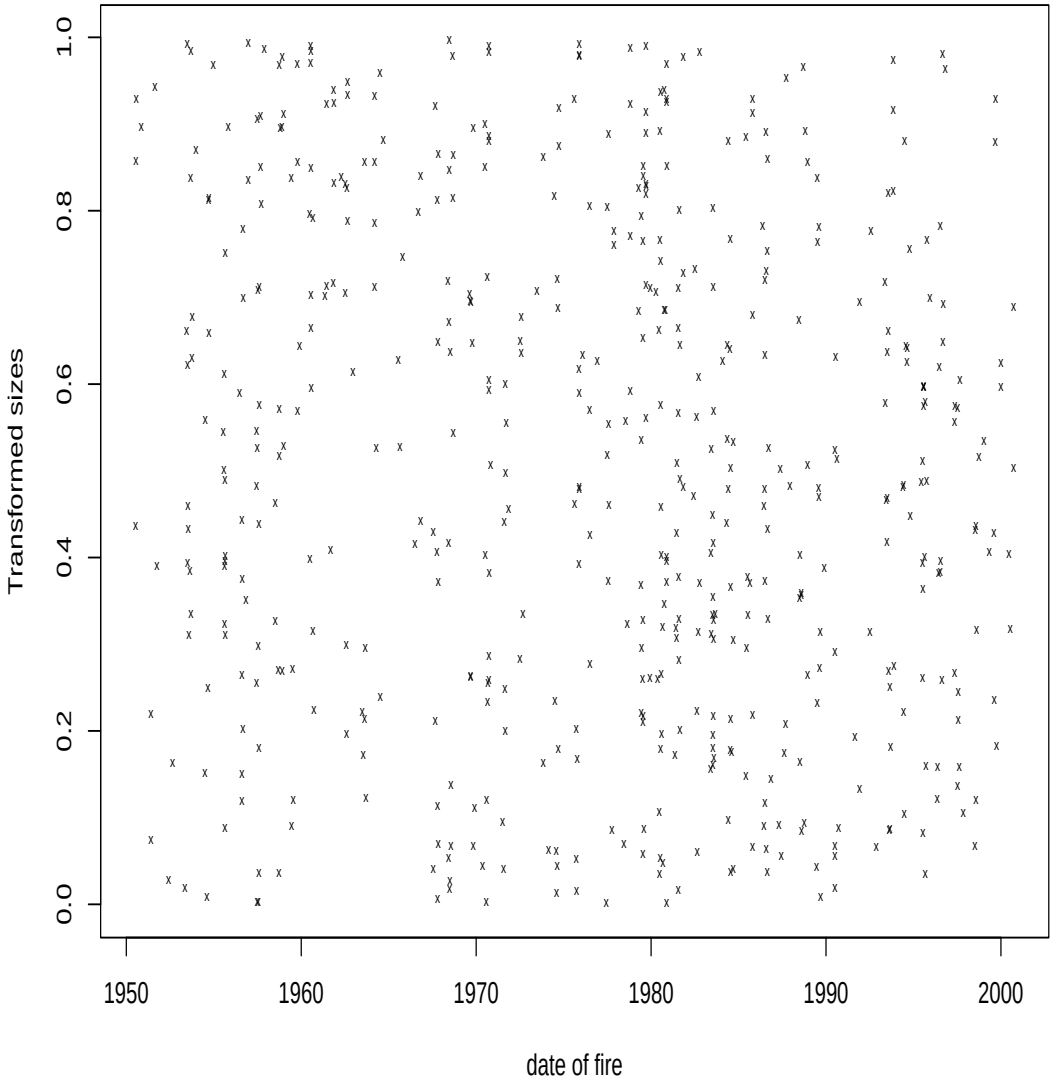
Figure 7: Residuals of L-function of the transformed wildfire sizes U_i.

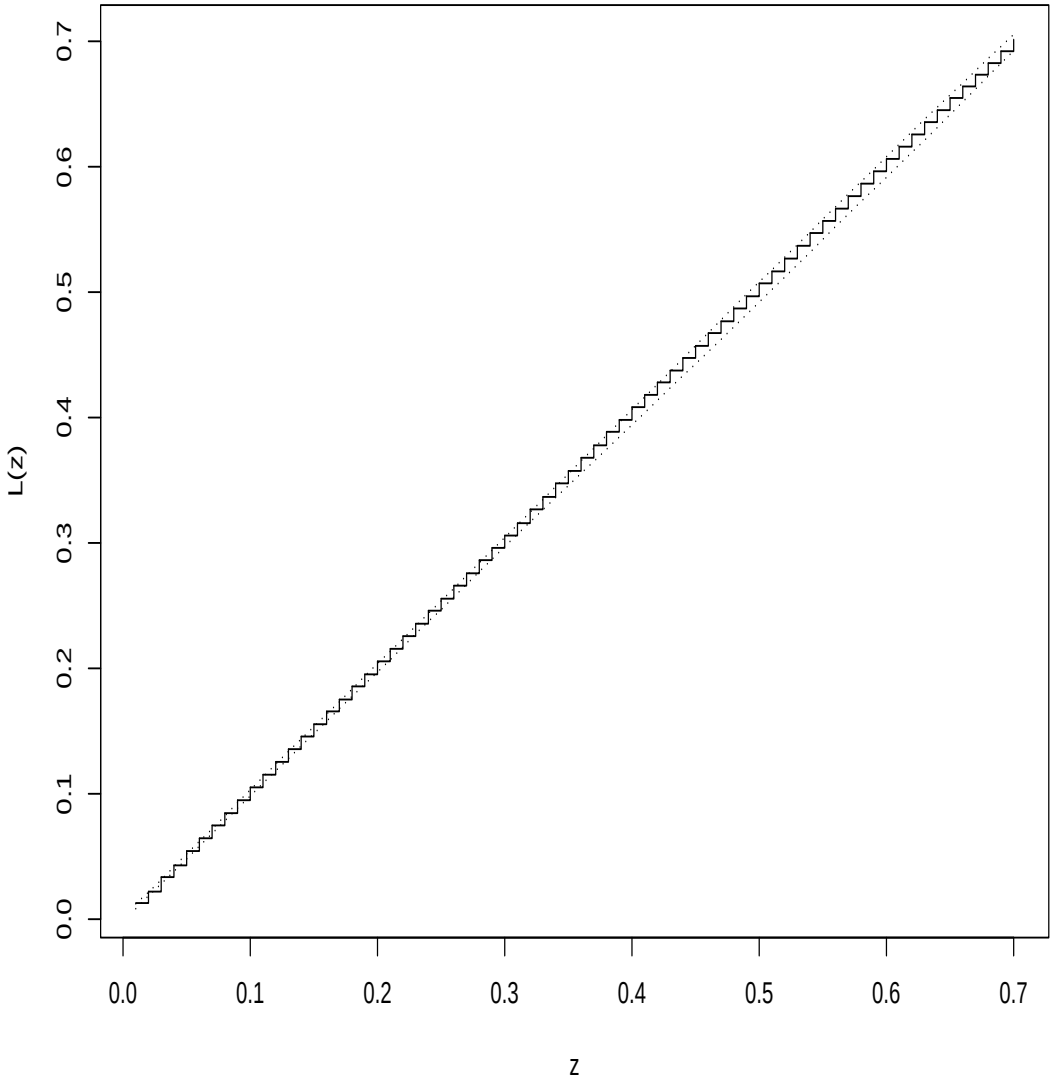

S(x)
0.500
0.050
0.005
0.001
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
x = area (sq miles)
tapered Pareto
Pareto
truncated Pareto
lognormal
half−normal
exponential
Frechet

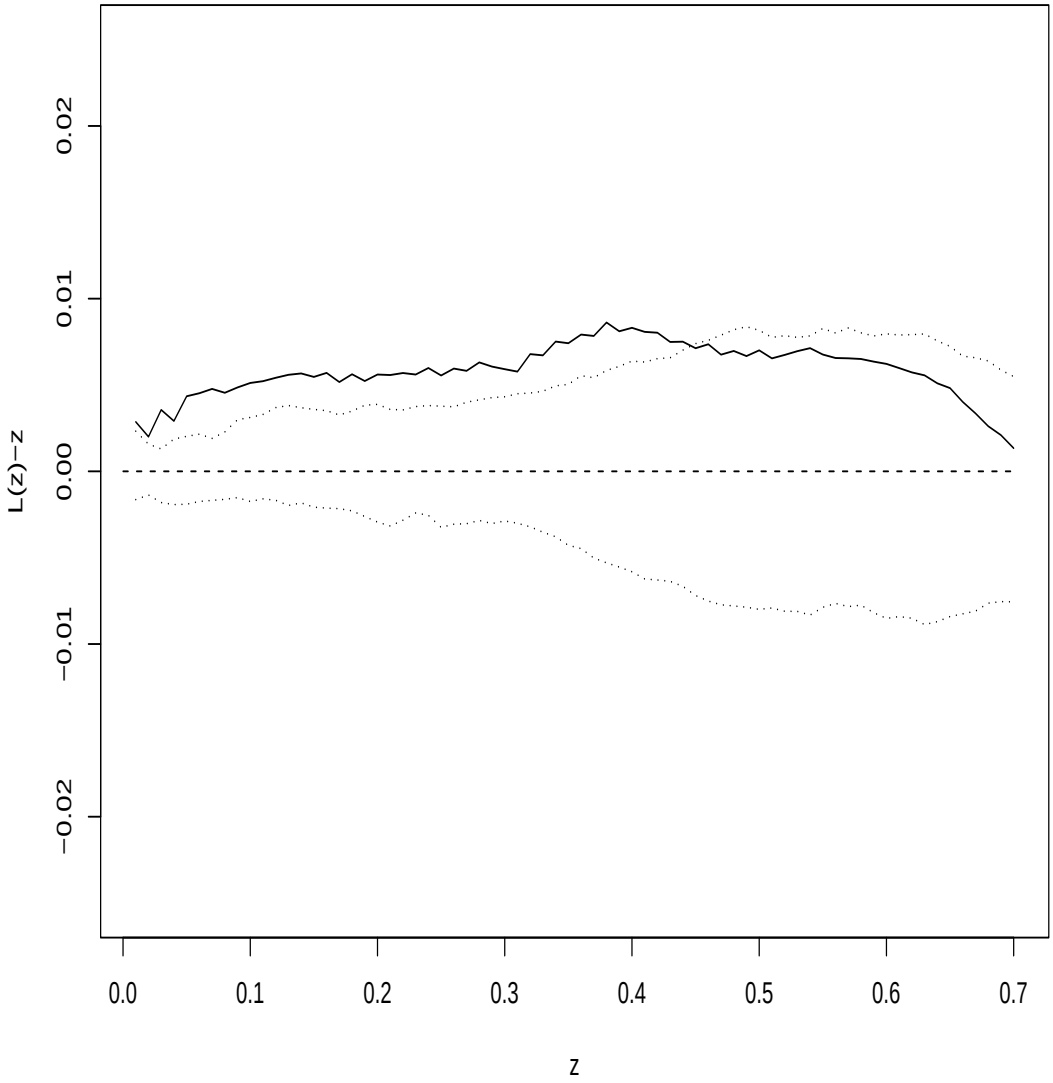

tapered Pareto (sq miles)
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
empirical (sq miles)


Truncated Pareto (sq miles)
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
empirical (sq miles)


Pareto (sq miles)
1
10
100
1000
10000
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
empirical (sq miles)


Transformed sizes
0.0
0.2
0.4
0.6
0.8
1.0
1950
1960
1970
1980
1990
2000
date of fire


L(z)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
z
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


L(z)−z
0.02
0.01
0.00
−0.01
−0.02
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
z