Title
The distribution of Voronoi cells generated by Southern California earthquake epicenters

Authors
Schoenberg, Frederic P
Barr, Christopher
Jungju Seo

Publication Date
2007-01-29

The distribution of Voronoi cells generated by Southern California earthquake epicenters

Frederic Paik Schoenberg[1], Christopher Barr[1], and Jungju Seo[1]

Running Title: Tessellation of earthquake epicenters.

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

phone: 310-794-5193

fax: 310-206-5658

email: frederic@stat.ucla.edu

Postal address: UCLA Dept. of Statistics

8142 Math-Science Building

Los Angeles, CA 90095–1554, USA.

Abstract

The cells of Voronoi diagrams generated by epicentral locations of Southern California earthquakes are inspected. The tapered Pareto distribution is shown to fit quite well to the distribution of cell areas and perimeters. This same distribution, which has been used to model the distribution of seismic moments, is also a close approximation to the empirical distributions of times and distances between successive earthquakes for the same catalog of Southern California events. Verification is performed using a variety of different windows and sub-sampling procedures in order to confirm that the results are not an artifact of the particular parameters of the selected earthquake catalog.

1 Introduction

Given a point pattern consisting of points $p_1, p_2, ..., p_n$ lying in some metric space $\mathcal{S}$, a Voronoi diagram is a division of $\mathcal{S}$ into n distinct cells $C_1, C_2, ..., C_n$, such that cell C_i consists of all locations in $\mathcal{S}$ that are closer to point p_i than to any other point of the point pattern. That is,

$$C_i = \{\mathbf{x} : ||x - p_i|| = \min_j ||x - p_j||\}.$$

The collection of all such cells is called a Voronoi tessellation. Voronoi tessellations have proven to be useful in a wide variety of disciplines including biology, astronomy, forestry, geology, and ecology (see Okabe et al. 2000). In seismology, Voronoi tessellations and their variants have been used in the description of seismic plates (Fohlmeister 1994), in the weighting of instrumental recordings for source-parameter inversion or isoseismal construction for a given earthquake (Pettenati and Sirovich 1993; Sirovich et al. 2002), and in the character-

ization of the spatial variation in parameters in epidemic-type aftershock sequence (ETAS) models (Ogata et al. 2003).

The primary focus of the present paper is on the properties of the cells in the Voronoi tessellation generated by the epicentral locations of earthquakes in Southern California. Our main finding is that the distributions of areas and perimeters of these cells are approximated very closely by the tapered Pareto model. This same distribution has been used to describe the distribution of seismic moments (Jackson and Kagan 1999; Vere-Jones et al. 2000; Kagan and Schoenberg 2001) and is closely related to the Pareto distribution which has often been used in modeling the distance in space or time between an earthquake and its aftershocks (Ogata 1998).

The discovery that, as with seismic moments, Voronoi cell areas and perimeters seem to have a distribution that is better approximated by the tapered Pareto law rather than the pure Pareto suggests a comparison of the fit of the two models to the distributions of temporal and spatial distances between earthquakes as well. We find that in fact the tapered Pareto distribution fits remarkably well to the distribution of inter-event times and inter-event distances for Southern California earthquakes, and the fit is far superior to that of the pure Pareto distribution.

The paper continues as follows. Section 2 describes the Southern California earthquake dataset considered here. Section 3 reviews certain models used to approximate the distribution of cell areas and cell perimeters, as well as inter-event times and distances, and the fit of these models is presented in Section 4. The robustness of these findings is considered in Section 5 and a discussion is given in Section 6.

2 Data

The dataset explored here consists of all 7,567 shallow (less than 70km in depth) local and regional earthquakes of local magnitude at least 3.0 recorded by the Southern California Earthquake Center (SCEC) between 01/01/1984 and 04/01/2007, with latitudes and longitudes ranging from 32^o to 37^o and from -112^o to -114^o, respectively. The spatial boundary considered is identical to that in Veen and Schoenberg (2005). The data are thought to be relatively complete down to a local magnitude of about 3.0 (Ouillon and Sornette 2005).

In addition to the possibility of missing events, there may be errors in the estimates of epicentral locations, moment magnitudes, and origin times, and these errors are thought to be especially substantial for the smallest events and those happening in a short space-time window after a prior event (Kagan 2004). As a result, our focus is primarily on the estimation of the upper 90% portions of the various distributions considered. That is, in analyzing the distribution of cell perimeters, for instance, we consider the tessellation generated by all 7,567 epicenters and consider the fit of models for only the upper 90% of these cell perimeters, since the smallest 10% of the cells have perimeters that are likely estimated with very substantial noise.

3 Methods

Voronoi tessellations of planar point process data are readily constructed using the *deldir* library (Turner, 2002) available within R (R Development Core Team, 2006). Cells intersecting with the boundary of the region considered are excluded from our analysis, since their areas and perimeters are so heavily dependent on the choice of boundary.

In addition to the distribution of areas and perimeters of Voronoi cells, the distributions of the inter-event times and inter-event distances, defined here as the time and spatial distance, respectively, between any two earthquakes occurring sequentially in time, are also inspected. Rather than categorize earthquakes as mainshocks or aftershocks, we simply define the inter-event times and distances for such a local catalog as the times and distances, respectively, between any two successive earthquakes in the catalog.

A variety of models are fitted to the distribution of Voronoi cell areas, cell perimeters, inter-event times, and inter-event distances. One such model is the Pareto distribution, whose cumulative distribution function is given by:

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

The parameter β in the Pareto model is the slope of the decrease in survivor function $1 - F(x)$ with x, when plotted on log-scale. The lower truncation point a, sometimes called the *completeness threshold*, represents a lower limit typically based on the sensitivity of the records in question.

Many phenomena have relatively heavy-tailed distributions but not quite as heavy-tailed as the Pareto, and such observations may be modeled using a tapered version of the Pareto distribution. The tapered Pareto distribution was originally suggested by Vilfredo Pareto himself (Pareto 1897), and has been used to describe the distribution of the sizes of earthquakes (Jackson and Kagan 1999; Vere-Jones et al. 2000) and wildfires (Schoenberg et al. 2003). The tapered Pareto has cumulative distribution function:

$$F_{tap}(x) = 1 - (a/x)^{\beta} exp(\frac{a-x}{\theta}), a \leq x \leq \infty \tag{2}$$

Here θ is a threshold after which frequency begins to decay especially rapidly. Additional

information concerning the density, characteristic function, moments, and other properties of the tapered Pareto can be found in Kagan and Schoenberg (2001).

For comparison, some other commonly-used models are fitted to the empirical distributions discussed in this paper, including the log-normal and exponential distributions, for which the cumulative distribution functions are $F(x) = \Phi\left(\frac{\log(x)-\mu}{\sigma}\right)$ and $F(x) = 1 - \exp(-\lambda x)$, respectively, where Φ denotes he standard normal cumulative distribution function. Details on the moments, estimates and other properties of the log normal and exponential distributions are given in Johnson et al. (1995).

Parameters are estimated by maximum likelihood using the Nelder-Mead optimization algorithm with various starting values; in all cases considered here, the resulting parameter estimates did not depend highly on the choice of starting value. The diagonal of the inverse of the Hessian of the log-likelihood function is used to provide standard errors for the parameter estimates, and Monte Carlo simulations are used to confirm these results.

Goodness-of-fit of the resulting models is evaluated by examining plots of survivor functions and Q-Q plots. In addition, the Akaike Information Criterion (AIC) is useful for comparing the relative fits of models to a given dataset. The AIC, defined as twice the negative log-likelihood plus twice the number of fitted parameters, rewards a model for fitting well and thus have a higher log-likelihood, while including a penalty based on the number of estimated parameters in order to avoid over-fitting (Akaike 1977).

4 Results

Figure 1 shows the Voronoi tessellation generated by the epicenters of recorded earthquakes of magnitude at least 3.5 occurring in the selected region from 1/1/1984 to 6/7/2007. The vast majority of the cells are very small, with areas ranging from 0.1 to 1.0 km^2, while some are quite large, with areas of several hundred km^2. This is a feature of heavy-tailed distributions such as the Pareto distribution and its variants.

Figure 2 displays the empirical survivor functions for Voronoi cell area, Voronoi cell perimeter, inter-event distance, and inter-event time, on a logarithmic scale. Overlaid within each panel are the fitted survivor functions for the tapered Pareto, Pareto, log-normal, and exponential distributions, with parameters fitted by maximum likelihood. The fitted parameter estimates, along with their accompanying standard errors, are reported in Table 1. From Figure 2 it is seen that the tapered Pareto distribution fits quite well to all four empirical distributions. The pure Pareto distribution fits poorly in every case, grossly overestimating the frequency of very large values of each variable. By contrast, the exponential distribution overestimates the frequency of very small values and, particularly in the case of cell area and perimeter, grossly underestimates the frequency of very large values. The log-normal distribution has a shape comparable to that of the tapered Pareto, but in each case considered here the tapered Pareto distribution appears to fit more closely to the data.

The rows labeled "Complete" in Table 2, which show relative values of the AIC for each of the four models and each of the four variables considered, confirms the superior fit of the tapered Pareto distribution relative to these various alternatives. Lower values of AIC indicate superior fit, and in each case the AIC for the tapered Pareto was the lowest of the

four models.

Q-Q plots for the various distribution are compared to the empirical distribution of inter-event times in Figure 3. The quantiles of the empirical distribution all fall within or very nearly within the 95% bounds for the tapered Pareto distribution. For the Pareto, log-normal and exponential distributions, large discrepancies with the data are readily evident. The results for cell area, cell perimeter, and inter-event distance are similar to those in Figure 3.

Figure 4 displays tapered Pareto Q-Q plots for cell area, cell perimeter, inter-event time, and inter-event distance. The tapered Pareto distribution provides a close approximation to the empirical distribution for all four variables. The agreement between the data and the tapered Pareto distribution for the case of the perimeters of Voronoi cells is especially close. In the case of inter-event distances, the best-fitting tapered Pareto distribution somewhat overpredicts the frequency of very large values. The tapered Pareto distribution also appears to slightly underpredict the frequency of the largest cell areas. Most other discrepancies between the model and the data are rather negligible.

5 Robustness

One may question the extent to which the results above depend on the particularities of the space-time window considered here. To investigate the robustness of these results with respect to time, the dataset described in Section 2 was divided into three portions: 01/01/1984 to 12/31/1991; 01/01/1992 to 12/31/1999; and 01/01/2000 to 04/01/2007. The tapered Pareto distribution is fit to each of the four variables considered (cell area, cell perimeter,

inter-event distance, and inter-event time) within each of the three temporal sub-divisions. In each case, the tapered Pareto again provides very close fit to the data. Figure 5 shows the empirical and fitted (tapered Pareto) survivor functions for cell area; those for cell perimeter, inter-event time, and inter-event distance were similar.

Further results on the sensitivity to choice of space-time window are presented in Table 2, which compares median difference in AIC for models fit to earthquakes from 50 different spatial subsets of the SCEC dataset described in Section 2. Each subset consists of all earthquakes within a randomly-chosen square sub-region, whose center is selected randomly from a uniform distribution on the spatial domain of the entire dataset. The edge lengths of the sub-regions are chosen uniformly between 20 and 150 kilometers, subject to the constraint that each subset contain between 300 and 500 events, so that results of the subsets would be comparable in terms of AIC, which depends critically on sample size. For each such subset, inter-event times and inter-event distances, defined as the temporal or spatial distance to the next event *within the subset only*, are computed, and the Voronoi tessellation of this subset of events is constructed. The fit of various models to the distribution of cell areas, cell perimeters, inter-event distances, and inter-event times is summarized in Table 2, which reports the median difference in AIC between the tapered Pareto model and the other models, for all 50 subsets. Another 50 iterations were performed using only earthquakes of magnitude 3.5 and greater, and the results are also reported in Table 2. The positive differences in AIC indicate that the median AIC for the tapered Pareto distribution is smaller than the median AIC for each of the other models, and in fact the differences between the tapered Pareto and pure Pareto distribution are quite substantial.

Note that the empirical distribution of cell areas changes whenever more earthquakes are

observed. Figure 6 shows how the maximum likelihood estimates of the parameters θ and β in the tapered Pareto distribution vary as the size of the time window grows to include an increasing number of events. In each case, the parameters are fitted to data spanning from 01/01/1984 to T, where T varies from 01/01/1985 to 04/01/2007. As T increases, the number of earthquakes within the same spatial region increases, so the mean area of the Voronoi cells must decrease. The estimates of θ decrease rapidly until approximately 2000 earthquakes are included in the catalog, after which the estimated values of $\hat{\theta}$ appear to decrease far more gradually. Estimates of the parameter β increase as the size of the dataset increases and seem to stabilize in the range of $0.25 - 0.28$ once the catalog contains at least 4000 events.

6 Discussion

The degree to which the distribution of Voronoi cell areas and perimeters for Southern California earthquakes is approximated by the tapered Pareto distribution is somewhat surprising. The results suggest that the organization and spatial clustering of these shallow earthquakes is sub-critical, as the decay in the frequency of the largest cells is much faster than one would expect from the pure Pareto law. The results suggest that the large-scale clustering patterns of these earthquakes are fundamentally distinct from those at small scales, and that there appears to be a gradual transition from the very tight power-law clustering of earthquakes in space-time over small times and distances to a regime that is better described by the exponential distribution in the tails. These results do not appear to be an artifact of the small sample size, since the parameters of the tapered Pareto seem to have approached

Table 1: Parameter estimates (and SEs) for the dataset described in Section 2.

Variable	Tapered Pareto	Pareto	Log Normal	Exponential
Spatial Distance	$\hat{\beta} = 0.0881$ (0.00259) $\theta = 202$ (3.65)	$\hat{\beta} = 0.265$ (0.00321)	$\hat{\mu} = 3.90$ (0.0211) $\sigma = 1.74$ (0.0149)	$\hat{\lambda} = 0.00735$ (8.74e-05)
Time Distance	$\hat{\beta} = 0.0947$ (0.00197) $\theta = 2.44$ (0.0495)	$\hat{\beta} = 0.193$ (0.00234)	$\hat{\mu} = -1.63$ (0.0301) $\sigma = 2.48$ (0.0213)	$\hat{\lambda} = 0.803$ (9.74e-03)
Cell Area	$\hat{\beta} = 0.273$ (0.00398) $\theta = 302$ (14.2)	$\hat{\beta} = 0.321$ (0.00393)	$\hat{\mu} = -1.40$ (0.0260) $\sigma = 2.12$ (0.0184)	$\hat{\lambda} = 0.0223$ (2.72e-04)
Cell Perimeter	$\hat{\beta} = 0.388$ (0.00963) $\theta = 37.4$ (1.32)	$\hat{\beta} = 0.639$ (0.00783)	$\hat{\mu} = -2.17$ (0.0130) $\sigma = 1.07$ (0.00922)	$\hat{\lambda} = 0.0604$ (7.39e-04)

convergence for the dataset considered here, nor do they appear to depend critically on the choice of observation window. However, further study is needed to determine whether similar results are obtained with other catalogs and in other seismic zones.

The pure Pareto distribution is commonly used to describe a wide variety of seismological phenomena, including the distributions of seismic moments, earthquake inter-event times and inter-event distances. The pure power-law form is used not only for the seismic moment distribution but also the branching behavior in standard models used for earth-

Table 2: Median difference in AIC between tapered Pareto and alternative models, for sub-samples from the dataset decribed in Section 2, and for lower magnitude thresholds of 3.0 and 3.5. AIC differences are reported in terms of the AIC for the given distribution minus the AIC for the tapered Pareto distribution, so that larger values indicate worse fit. Medians over 50 randomly selected sub-samples are reported.

Variable	Data Type	Magnitude	Pareto	Log Normal	Exponential
Spatial Distance	Sample	3.0+	347	105	84
	Sample	3.5+	531	78	45
	Complete	3.0+	7028	2156	2658
Time Difference	Sample	3.0+	220	62	1495
	Sample	3.5+	122	62	1733
	Complete	3.0+	6243	1945	9042
Cell Area	Sample	3.0+	61	111	755
	Sample	3.5+	85	56	356
	Complete	3.0+	975	1436	17903
Cell Perimeter	Sample	3.0+	71	123	249
	Sample	3.5+	104	83	252
	Complete	3.0+	1337	1811	3828

quake hazard forecasting, such as the ETAS models of Ogata (1998) and Ogata et al. (2003) or the branching models of Kagan and Knopoff (1987), Kagan (2004) and others. Recent discussions have focused on a "unified scaling law" involving the Pareto distribution in characterizing the distribution of inter-event times as a function of the parameters governing the size of the spatial-temporal observation window (Bak et al., 2002). Further, recent studies have suggested that the pure Pareto distribution be used to describe the distributions of a broad range of natural hazards, including not only earthquakes but also wildfires, asteroid impacts, eruptions of volcanos, and landslides (Malamud et al. 2005).

However, the current results raise the question of whether the description of earthquake catalogs as well as the forecasting of future seismicity may be improved by using the tapered Pareto distribution in place of the pure Pareto. Indeed, the tapered Pareto distribution not only offers improvement in approximating the distribution of seismic moments (Jackson and Kagan 1999; Vere-Jones et al. 2000; Kagan and Schoenberg, 2001), but also is shown here to offer superior fit to the distribution of earthquake inter-event times and inter-event distances, as well as the distributions of Voronoi cell area and perimeter. Further, unlike the pure Pareto, the tapered Pareto distribution has finite mean and variance may thus more naturally agree with physical notions such as the finiteness of seismic moment flux and of deformational energy (Sornette and Sornette 1999).

References

[1] Akaike, H. (1977). On entropy maximization principle. in *Applications of Statistics*, P.R. Krishnaiah, ed., North-Holland: Amsterdam, 27–41.

[2] Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(6), 1–42.

[3] Bak, P., Christensen, K., Danon, L., and Scanlon, T. (2002). Unified scaling law for earthquakes. *Physical Review Letters* 88(17), 178501-1 – 178501-4.

[4] Fohlmeister, J.F. (1994). Hotspots, mantle convection and plate tectonics – a synthetic calculation. *Pure and Applied Geophysics*, 143(4), 673–695.

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

[6] Johnson, N., Kotz S, Balakrishman, N. (1995). Continuous Univariate Distributions. *2nd ed.* Wiley, New York.

[7] Kagan, Y. Y., and Knopoff, L. (1987). Random stress and earthquake statistics: time dependence. *Geophysics Journal of the Royal Astronomical Society* 88, 723–731.

[8] Kagan, Y. Y. (2004). Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America*, 94(4), 1207–1228.

[9] Kagan, Y. Y. and Schoenberg, F. P. (2001). Estimation of the upper cutoff parameter for the tapered Pareto distribution. *Journal of Applied Probability*, 38A, Festershrift for David Vere-Jones, D. Daley, editor, 158-175.

[10] Malamud, B.D., Millington, J.D.A., and Perry, G.L.W. (2005). Characterizing wildfire regimes in the United States. *Proceedings of the National Academy of Science of the United States of America* **102**(13), 4694-4699.

[11] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2), 379–402.

[12] Ogata, Y., Katsura, K., and Tanemura, M. (2003). Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis. *Journal of the Royal Statistical Society, Series C*, 52(4): 499-509.

[13] Okabe, A., Boots, B., Sugihara, K. and Chiu, S. (2000). Spatial Tessellations, *2nd ed.* Wiley, Chichester.

[14] Ouillon, G., and Sornette, D. (2005). Magnitude-dependent Omori law: theory and empirical study. *Journal of Geophysical Research***110**, B04306.

[15] Pareto, V. (1893). Cours d'Economie Politique, Tome Second, Lausanne, F. Rouge, quoted by Pareto, V. (1964), Oevres Completes, publ. by de Giovanni Busino, Geneve, Droz, v. II.

[16] Pettenati, F., and Sirovich, L. (2003). Tests of source-parameter inversion of the US Geological Survey intensities of the Whittier Narrows 1987 earthquake. *Bulletin of the Seismological Society of America*, 93(1): 47–60.

[17] R Development Core Team. (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[18] Schoenberg, F. P., Peng. R., and Woods, J. (2003) On the distribution of wildfire sizes. *Environmetrics*, 14(6), 583-592.

[19] Sirovich, L., Cavallini, F., Pettenati, F., and Bobbio, M. (2002). Natural-neighbor isoseismals. *Bulletin of the Seismological Society of America*, 92(5): 1933-1940.

[20] Sornette, D., and Sornette, A. (1999). General theory of the modified Gutenberg-Richter law for large seismic moments. *Bulletin of the Seismological Society of America* 89, 1121-1130.

[21] Stoyan, D., Kendall, W., and Mecke, J. (1995) Stochastic Geometry and its Applications, *2nd ed.* Wiley, Chichester.

[22] Turner, R. (2002). deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tesselation. R package version 0.0-5. http://www.math.unb.ca/ rolf/

[23] Veen, A., and Schoenberg, F. P. (2005). Assessing spatial point process models for California earthquakes using weighted K-functions: analysis of California earthquakes. in *Case Studies in Spatial Point Process Models*, Baddeley, A., Gregori, P., Mateu, J., Stoica, R., and Stoyan, D. (eds.), Springer, NY.

[24] Vere-Jones, D., Robinson, R., and Yang, W.Z. (2000). Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysics Journal International* 144(3), 517–531.

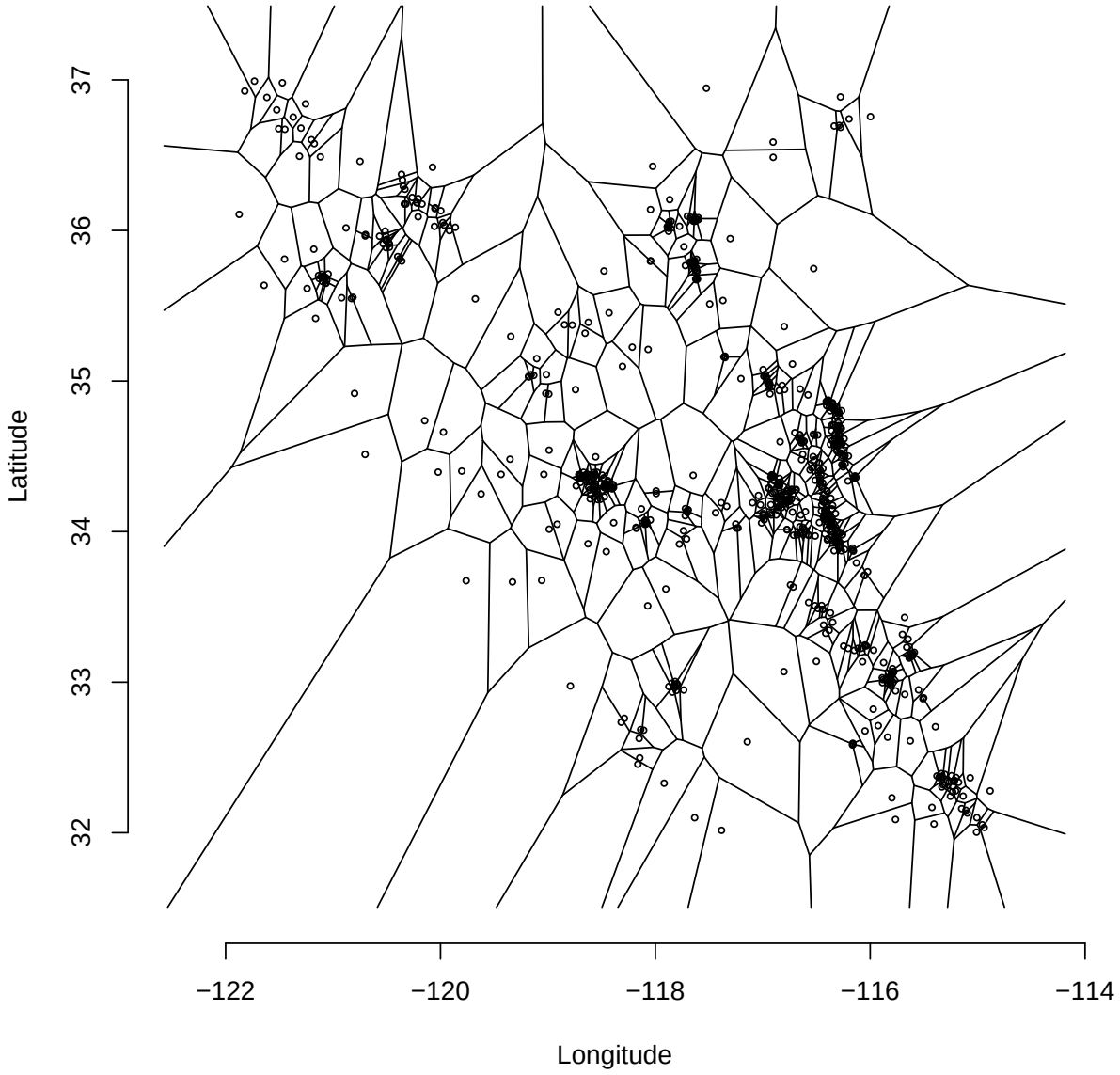


Figure 1: Voronoi diagram for epicenters of local and regional shallow Southern California Earthquake Center (SCEC) events of local magnitude 3.5 and greater, from 01/01/1984 to 04/01/2007.

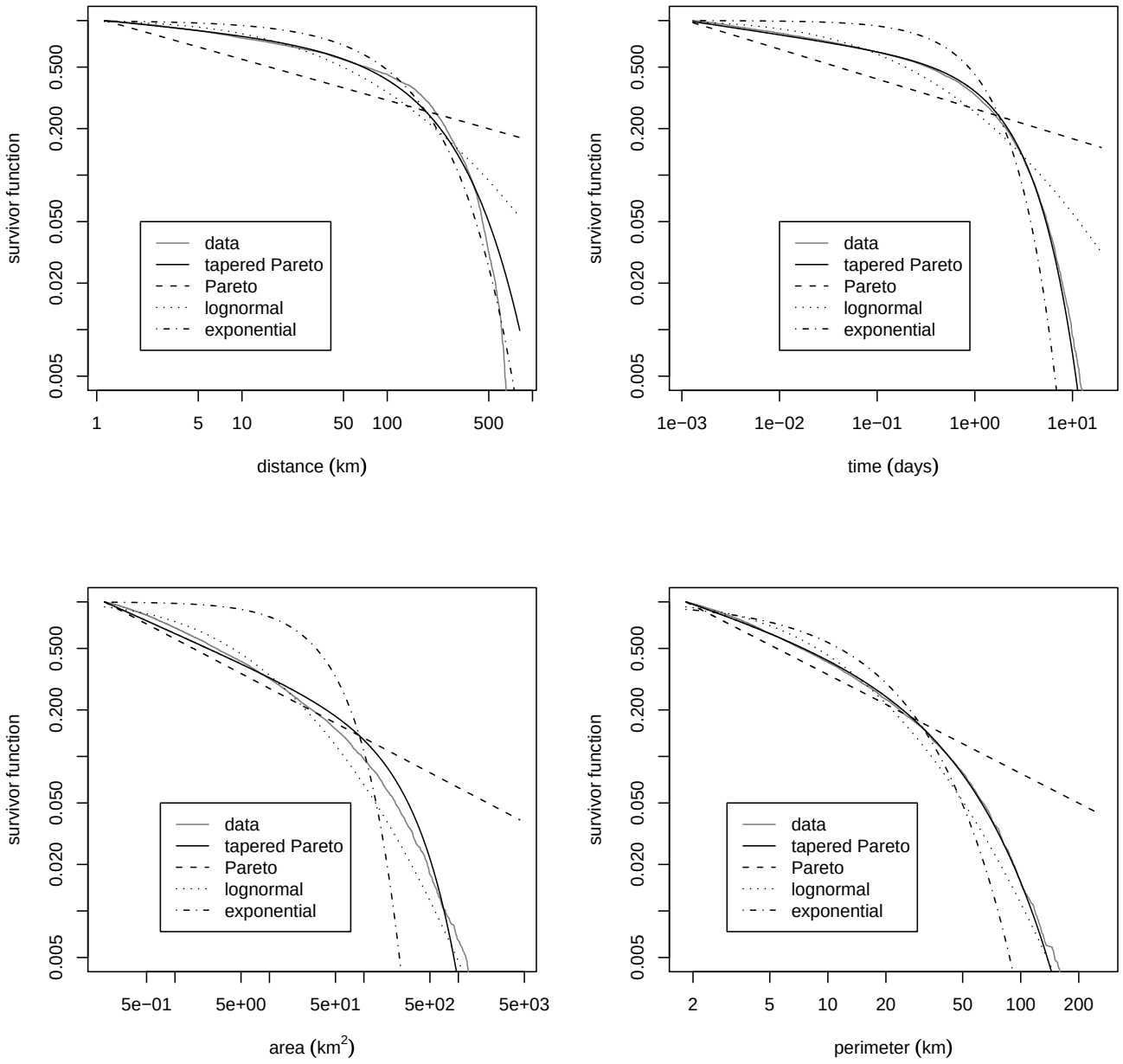


Figure 2: Estimated and empirical survivor functions for inter-event distance, inter-event time, cell area, and cell perimeter.

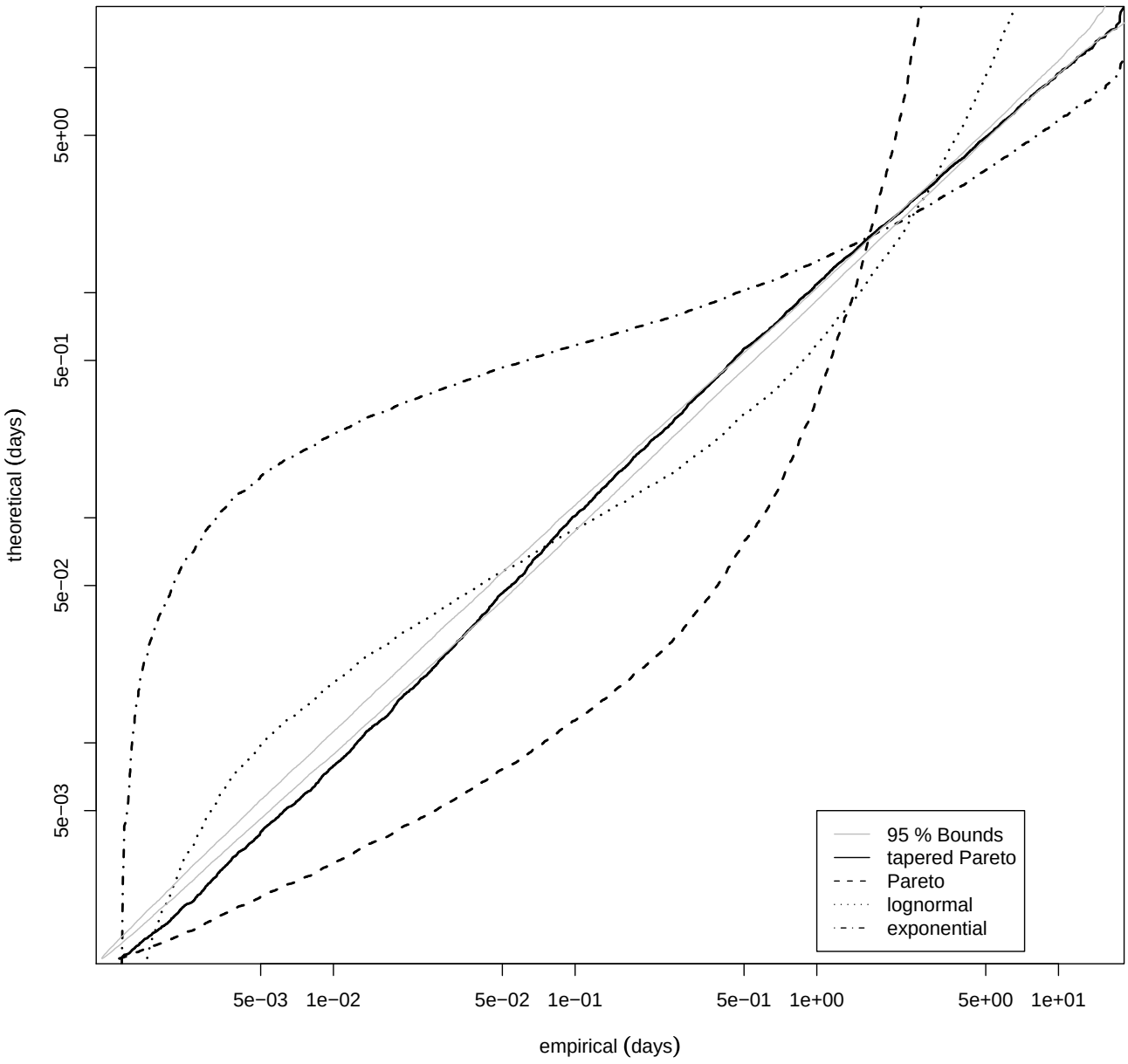


Figure 3: QQ plot for inter-event times of SCEC earthquakes.

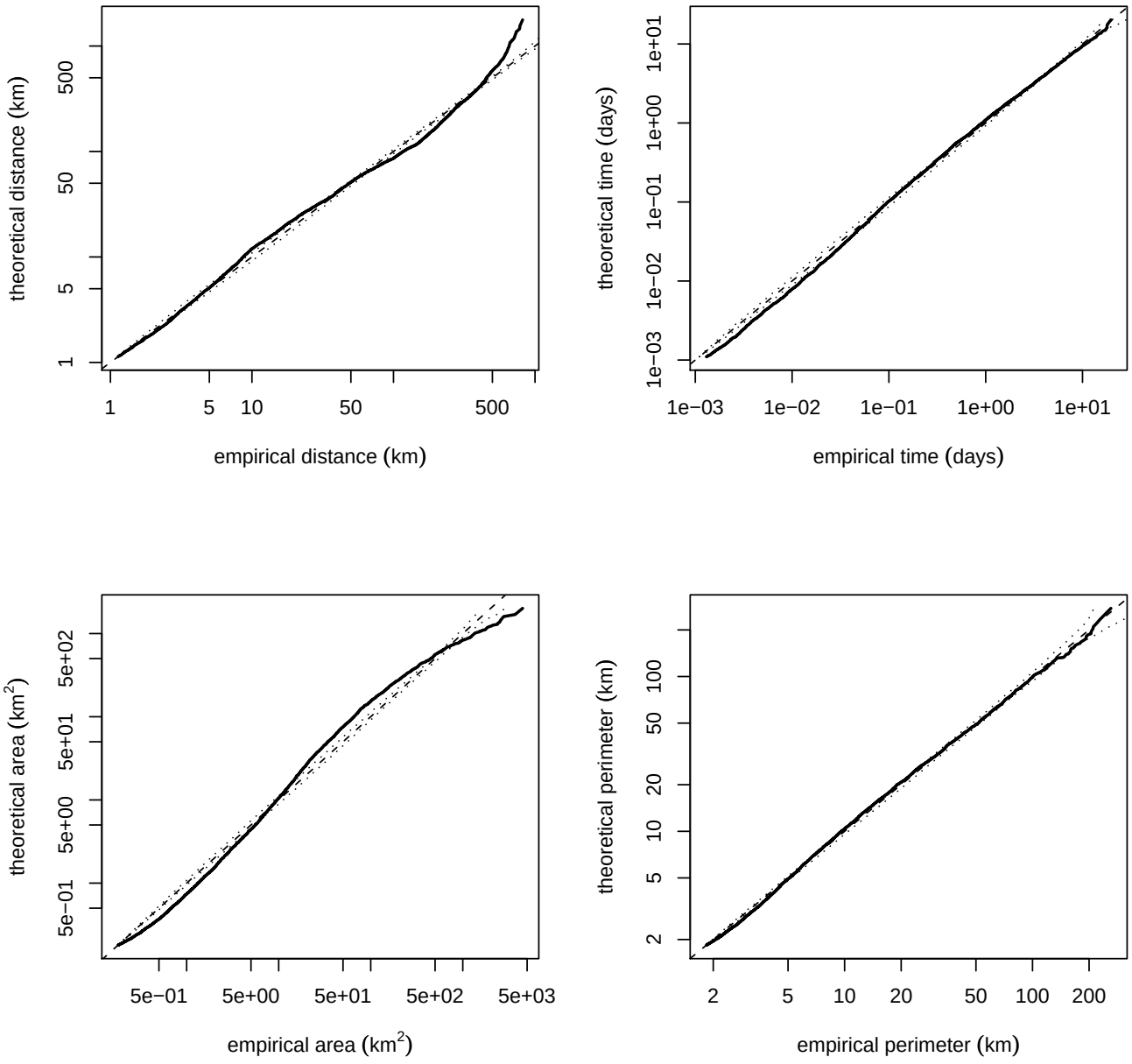


Figure 4: QQ plots for inter-event distance, inter-event time, cell area, and cell perimeter. Dashes indicate 45^o lines and dots indicate 95%-confidence bounds.

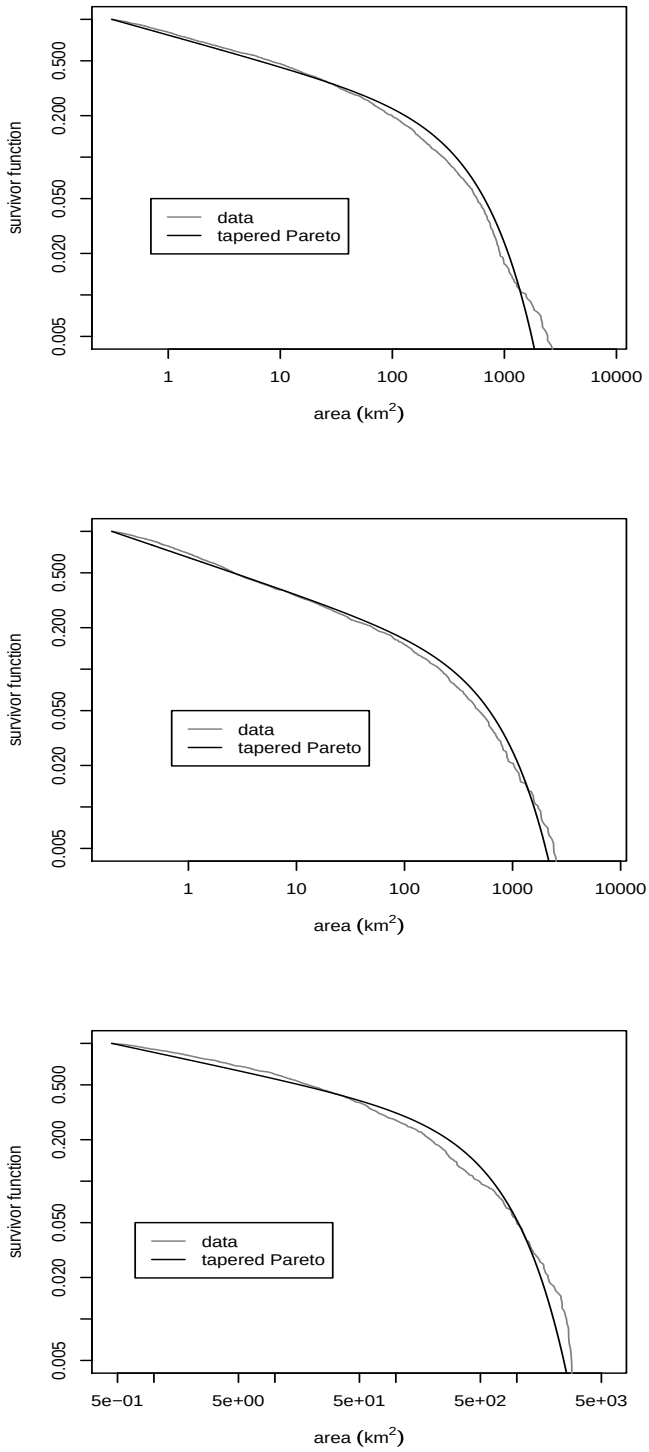
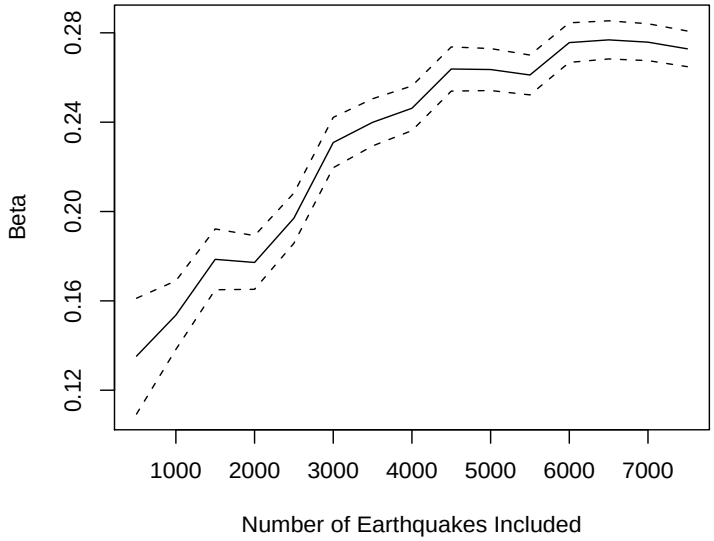


Figure 5: Estimated and empirical survivor functions for cell areas in the Voronoi tessellation of SCEC earthquake epicenters; 1984 - 1992 (top), 1993 - 2000 (middle), and 2001 - 2007 (bottom).



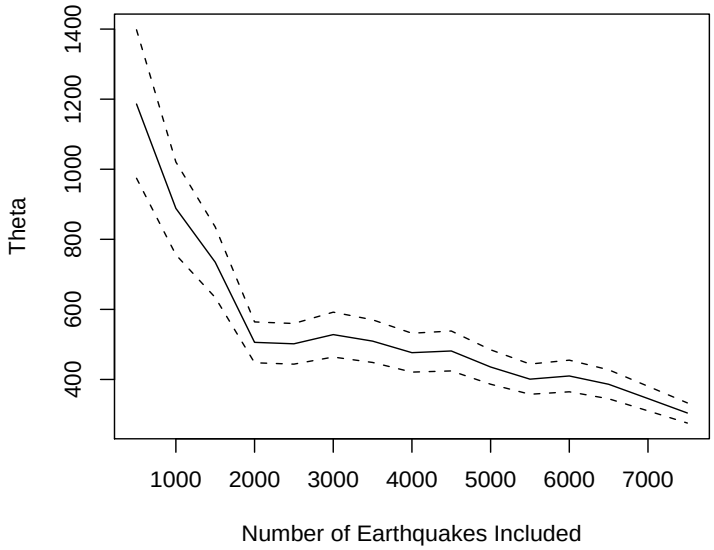


Figure 6: Maximum likelihood estimates of parameters in the tapered Pareto distribution fit to the cell areas of the Voronoi tessellation of subsets of earthquakes, as a function of number of earthquakes in the subset of the dataset.