ON THE RANK-SIZE DISTRIBUTION FOR HUMAN SETTLEMENTS*

William J. Reed*
Department of Mathematics and Statistics
University of Victoria
PO Box 3045 Victoria B.C.
Canada V8W 3P4.

E-mail: reed@math.uvic.ca

April, 2000. Revised January 2001

JEL classification: R12; C49

*Research supported by NSERC grant OGP 7252. This paper was written while the author was visiting the Department of Mathematics and Statistics, University of Melbourne, whose support and hospitality are gratefully acknowledeged.

Abstract

An explanation for the rank-size distribution for human settlements based on simple stochastic models of settlement formation and growth is presented. Not only does the analysis of the model explain the rank-size phenomenon in the upper tail, it also predicts a reverse rank size phenomenon in the lower tail. Furthermore it yields a parametric form (the double Pareto-lognormal distribution) for the complete distribution of settlement sizes. Settlement-size data for four regions (two in Spain and two in U.S.A.) are used as examples. For these regions the lower tail rank-size property is seen to hold and the double Pareto-lognormal distribution shown to provide an excellent fit, lending support to the model and to the explanation for the rank-size law.

Keywords: Rank-size distribution; Zipf's law; Pareto law; lower-tail rank-size property; Gibrat's law; Yule process; mixture model; double Pareto-lognormal distribution.

1 INTRODUCTION.

It is a remarkable fact that the distribution of city sizes exhibits a high degree of regularity across various countries and periods in history. This observed phenomenon is often referred to as *Zipf's Law*, after Zipf (1949) who observed that the logarithm of population size when plotted against the logarithm of the rank of the city produced points close to a straight line, with negative slope. Nowadays (*e.g.* Brakman *et al.,* 1999, Gabaix, 1999), the term Zipf's Law is often used to refer exclusively to the case of a slope of negative one (rank inversely proportional to size) while for more general negative slope the term *rank-size distribution* is used. It is well known that this phenomenon is closely related to the *Pareto law of incomes* and that a probabilistic way of stating the law is that city sizes above a certain threshold size follow a Pareto (or power law, or fractal) distribution. There have been many attempts to explain this observed regularity. Suh (1987) divides these into two classes: (i) hierarchical models based on microeconomic assumptions (see *e.g.* Allen and Sanglier, 1979, 1981 and the many papers cited in the review of Mulligan, 1984); and (ii) stochastic models which seek to explain the observed distribution as a consequence of simple probabilistic assumptions concerning the formation and growth of cities (see *e.g.* Simon, 1955, Steindl 1965, 1968). However none of the attempts from either side has been wholly successful. At least one author (Sheppard, 1982) has questioned the value of such attempts, claiming that the rank-size law is a 'profoundly over-identified concept'.

Nonetheless as recently as 1999 Brakman *et al.* described the rank-size distribution as 'an empirical regularity in search of a theory.' These authors offered an explanation using a general equilibrium approach in a model which incorporated negative feedbacks due to congestion in a common model of economic geography. Although this model was capable of producing size distributions mimicking the rank-size distribution, the results like those of all previously proposed explanations, are not totally satisfactory in that they can explain the observed distributions only for cities above a certain threshold size. A recent paper by Gabaix (1999) uses a stochastic model employing Gibrat's law of proportional effects (Gibrat, 1931) to describe city growth. The paper, similar in spirit to that of Champernowne (1953) on the Pareto law of incomes, offers an explanation of why the rank-size phenomenon should hold and why the exponent should be unity (*i.e.* why Zipf's law rather than the more general rank-size law should hold). Although empirical studies (*e.g.* Rosen and Resnick, 1980) have indicated the frequent occurrence of exponents different from one, Gabaix claims that this fact can be explained as a consequence of finite sample sizes. If this is the case one would expect that the magnitude of the deviations from unity should be negatively correlated with the number of cities in the study. Unfortuantely this does not appear to be the case. For example Rosen and Resnick examined 44 countries, using the largest 50 cities in each, except for six countries in which there were more than 50 cities of size larger than 100,000. For these six countries (with respectively 59, 91, 138, 149, 151 and 225 cities larger than 100,000), the calculated exponents (ranging between 1.153 and 1.289) were between the 29th. and 39th. largest of the 44 calculated. The U.S.S.R.with 225 such cities had an exponent of 1.278, the 38th. largest of the 44. This does not appear to be compatible with Gabaix's explanation for deviations from unity.

Discussion of city size distributions and the rank-size property has apparently been confined to cities above a certain threshold size. While there have been investigations on the magnitude of the threshold, (*e.g.* Guerin-Pace, 1995), there appears to have been little discussion of the size distribution in its entirety. This article offers an explanation for the observed size distribution of human settlements which holds over the whole range of observed sizes and for which the rank-size (Pareto) property holds in the upper tail. In addition a reverse rank-size property will be predicted and seen to hold empirically in the lower tail. The model is based on very simple probabilistic assumptions about the formation and growth of settlements, which reflect the inherent variability (from settlement to settlement and over time) in these processes.

Thus the explanation for the observed regularity in the size distribution is essentially mathematical in that it is a consequence of stochastic processes with certain characteristics. This does not mean that geographic, economic and other factors are not important in determining the growth and eventual size of any city. Rather it means that when looking at the *distribution* of settlement sizes over a whole region, the effects of the variation in these factors can be modelled effectively by stochastic processes; and the mathematical analysis of these processes leads to a convincing explanation.

Being mathematical, the explanation is not confined to settlement size distributions, but should be relevant for other pheneomena with similar underlying structure. As recognized by others (*e.g.* Steindl, 1968; Gabaix, 1999), the Pareto law of incomes is one such phenomenon, and indeed a very similar model to the one used herein has been proposed to explain (and extend) this law (Reed, 1999).

2 A STOCHASTIC MODEL FOR THE FORMATION AND GROWTH OF SETTLEMENTS.

The foundation and subsequent growth of human settlements depends upon many things including for example geographic, economic and demographic factors. Rather than attempt to model these factors and their interactions as has been attempted in many hierarchical models, in this paper a 'macro' view will be adopted, and the differences between settlements will be regarded as essentially random or unexplained components in a basic underlying model. Thus we will present a stochastic model, which comprises two components, one for the foundation of settlements and the other for their subsequent evolution after foundation. From these component models a probability distribution for the current size of settlements will be derived and its properties discussed, as well as how this theoretical size distribution can be fitted to empirical data. We consider first the evolution of settlements after their foundation.

Individual human settlements grow (and sometimes contract) in different and varying ways. The (proportional) rate of growth in size in a given year will vary from settlement to settlement and for a given settlement will likely vary from year to year (or decade to decade *etc.*), depending on economic, demographic factors, *etc.* At a macro level this variability can be modelled mathematically by assuming that the logarithm of population size for any settlement constitutes a realization of a random walk. This is Gibrat's law of proportional effects (Gibrat, 1931). For analytic convenience in this paper the continuous-time version of this model will be used *i.e.* the size (population) $X(t)$ of a settlement will be assumed to follow Geometric Brownian Motion (GBM) governed by the Itô stochastic differential equation

$$dX = \mu X dt + \sigma X dw. \tag{1}$$

where dw is white noise (*i.e.* the random increment of a Wiener process in time dt). The parameter μ is the mean proportional growth rate over all settlements and all times, and σ is a parameter reflecting the variability in this growth rate. Thus the proportional growth dX/X in time dt for any settlement will comprise a systematic component μdt (reflecting the average growth rate over all settlements, at all times) and a random component σdw (reflecting what happened for the particular settlement at the particular time).

This is the model for settlement growth used by Gabaix (1999), who cites empirical studies justifying Gibrat's law. However unlike Gabaix, in this paper we shall not be looking at equilibrium conditions and thus will not need to assume the existence of a minimum city size acting as a reflecting boundary.

If the initial size of the settlement (at time of foundation) is X_0, then under the GBM model the size $\tilde{X}_T$ of the settlement T time units later will be a lognormal random variable with

$$\ln \tilde{X}_T \sim N(X_0 + (\mu - \sigma^2/2)T, \sigma^2 T) \tag{2}$$

Settlements currently in existence were founded at different times and doubtless had different initial populations. Thus for any given settlement in the country or region under consideration, the variables X_0 and T should be considered as random variables. A simple specification for the distribution of the initial size, X_0, would be that it is of a lognormal form with

$$\ln X_0 \sim N(\mu_0, \sigma_0^2)$$

This has the desirable properties that initial size would always be non-negative, with the variance increasing with the mean. It is possible that the distribution of starting sizes has changed over time (*e.g.* agricultural settlements likely were initially smaller than industrial ones *etc.*). One can easily accommodate this by assuming that X_0 also evolves as a GBM. It makes no essential difference to the development to include this, but for the sake of simplicity of exposition the details are relegated to the Appendix.

The time since foundation will vary from settlement to settlement and its distribution over all settlements in the region will reflect the region's historical development, which of course will have depended on many diverse factors. As with the evolution of settlements, we will ignore all of the details and instead model settlement foundation with a simple stochastic process. The simplest stochastic model that one could assume is that foundations occurred in a *Poisson process* over the last τ time units (*i.e.* they occurred randomly and independently at a constant average rate). This model however is limited in that it does not allow for overall growth in the region. A more realistic model results from assuming that in the time interval $(t, t+dt)$ any existing settlement can form a new satellite settlement with probability λdt. This is a *Yule process* first proposed by Yule (1924) as a model for the creation of new biological species, a process similar in many respects to the foundation of new human settlements. For such a process the expected number of settlements, t time units after the foundation of the first settlement, is $e^{\lambda t}$. In other words the number of settlements is growing, on average, at the proportional rate λ. It should be noted however that the actual evolution of the number of settlements is a random process. For this model one can show that the distribution of the time T since foundation of a settlement currently in existence, is of the form of an exponential distribution truncated at τ (the age of the first settlement), with an atom of probability (reflecting the probability that the given settlement is the oldest) of size $\frac{\lambda\tau e^{-\lambda\tau}}{1-e^{-\lambda\tau}}$ at the point τ. In most cases it is probably reasonable to assume that τ is large, and thus to consider the limiting distribution as $\tau \to \infty$. This is an exponential distribution with density $\lambda e^{-\lambda t}$ for $t > 0$.

Under the Yule process model for the foundation of settlements and the GBM model for their subsequent growth, the distribution of the current size, $\bar{X}$, over all settlements, can be obtained by integrating the density of lognormal distribution of $\tilde{X}_T$ with respect to the exponential distribution of T. This can be done analytically (see Appendix) yielding a probability density for $\bar{X}$ of the form

$$f_{\bar{X}}(x) = \frac{\alpha\beta}{\alpha+\beta}\Big[x^{-\alpha-1}\exp\{\alpha\mu_0 + \alpha^2\sigma_0^2/2\}\Phi\left(\frac{\ln x - \mu_0 - \alpha\sigma_0^2}{\sigma_0}\right) + x^{\beta-1}\exp\{-\beta\mu_0 + \beta^2\sigma_0^2/2\}\Phi^c\left(\frac{\ln x - \mu_0 + \beta\sigma_0^2}{\sigma_0}\right)\Big] \tag{3}$$

on $x > 0$ where Φ is the cumulative distribution function of the standard normal distribution; $\Phi^c = 1 - \Phi$; and α and $-\beta$ $(\alpha, \beta > 0)$ are the roots of a characteristic quadratic equation (See Appendix). This distribution will be called (for reasons made clear in the Appendix) the *double Pareto-lognormal distribution* (dPlN) (Reed, 1999). The possible shapes of the density (both in natural and logarithmic scales) in the cases $\beta > 1$ and $\beta < 1$ are presented in Reed (2000a).

The dPlN distribution has the property that it follows (different) power laws in its two tails *i.e.*

$$f_{\bar{X}}(x) \sim x^{-\alpha-1} \quad (x \to \infty); \qquad f_{\bar{X}}(x) \sim x^{\beta-1} \quad (x \to 0).$$

The first result indicates that for large x, the distribution of size follows a Pareto law (*i.e.* $\ln(\Pr(\bar{X} \geq x))$ is linearly related to $\ln(x)$ with slope

$-\alpha < 0$); and the second result indicates a reverse Pareto law for small x, (*i.e.* $\ln(\Pr(\bar{X} \leq x))$ is linearly related to $\ln(x)$ with slope $\beta > 0$). The upper-tail Pareto law has been widely verified empirically (the rank-size law), but nobody apparently has looked for a lower-tail (reverse) rank-size law. But it does indeed hold. In Reed (2000, b) a logarithmic plot of the (ascending) rank against size for the smallest 5000 settlements for the U.S.A in 1998 is given (as well as a similar plot of (descending) rank against size for the largest 5000 settlements). The degree of linearity in the lower-tail plot is even more striking than that for the upper-tail plot, confirming empirically the presence of the lower-tail rank-size property.

To fit the dPlN distribution to settlement-size data with independent[(1)] observations on n settlements of size $x_1, x_2, \ldots, x_n$ by maximum likelihood (ML) one needs to maximize the log-likelihood

$$l = \sum_{i=1}^{n} \ln(f_{\bar{X}}(x_i)) \tag{4}$$

over the four parameters $\alpha, \beta, \mu, \sigma$. This can be done numerically (*e.g.* using the S-Plus routine nlminb (Anon, 1997)). Plausible starting values for α and β can be found by regressing (on log-log scales) descending rank *vs.* size for large settlements; and ascending rank *vs.* size for small settlements. Using these values starting values for μ and σ can be found by the method of moments. Specifically if $\tilde{\alpha}, \tilde{\beta}$ are the starting values for α, β, starting values for μ, σ can be determined as

$$\tilde{\mu} = \bar{y} - \frac{\beta - \alpha}{\alpha\beta}; \qquad \tilde{\sigma} = \sqrt{s_y^2 + (\frac{\beta - \alpha}{\alpha\beta})^2 - 2\frac{\alpha^3 + \beta^3}{\alpha^2\beta^2(\alpha + \beta)}}$$

where $\bar{y}$ and s_y^2 are the mean and sample variance of the logarithms of observed sizes.

3 EXAMPLES.

The dPlN distribution was fitted to four empirical settlement size distributions - those of two U.S. states[(2)] in 1998 and of two Spanish provinces[(3)] in 1996. These examples were chosen because the datasets include even very small settlements (with fewer than 100 inhabitants – the smallest, in Barcelona province, has just 30 inhabitants). In each country one relatively heavily populated region (California and Barcelona respectively) and one relatively lightly populated region (West Virginia and Cantabria) were selected.

Figs. 1 and 2 show rank-size plots, in the upper and lower tails. Notice how the lower-tail plots exhibit linearity (at least as much as do the upper-tail plots) thereby empirically confirming the lower-tail Paretian behaviour predicted by the model of the previous section.

Further support for the model can be found by fitting the dPlN model to the four datasets using maximum likelihood, by numerically maximizing the log-likelihood (4) as outlined in the previous section. The ML estimates of the four parameters are presented in Table 1.

The fit of the model can be assessed by comparing the observed and fitted distributions as is done in Figs. 3 and 4. In particular Q-Q plots (of observed *vs.* fitted quantiles) provide a good method of revealing lack of fit. Lack of fit is suggested if the plotted points exhibit systematic departures from the

45 degree line. This is not evident here, although the corresponding plots for other suggested distributional forms, such as the lognormal and truncated lognormal, do indicate lack of fit. From Figs. 3 and 4 it is clear that the dPlN distribution provides a good fit to the data in each region. Given the adequacy of the model one can proceed to use asymptotic likelihood ratio (LR) tests to formally test hypotheses concerning parameters. For example a test of Zipf's law (in the upper tail) against the more general rank-size law is obtained by testing $H_0 : \alpha = 1$ *vs.* $H_1 : \alpha \neq 1$. The results for this test indicate that the data are compatible with Zipf's law for the more lightly populated regions (W. Virginia, $P = 0.56$; Cantabria $P = 0.67$) but not so for the more heavily populated ones (California, $P = 0.00001$; Barcelona, $P = .0002$).

Comparisons between any two regions can also be made. For example for two regions the LR test statistic for testing $H_0 : \alpha_1 = \alpha_2;\ \beta_1 = \beta_2;\ \mu_1 = \mu_2;\ \sigma_1 = \sigma_2$ against a general alternative is obtained as twice the difference in the maximized log-likelihood for the model fitted to data for the two regions pooled and its sum for the model fitted to each region separately. For the U.S.A. comparing W. Virginia and California yields a value of 536.4, which on comparison with $\chi^2_{(4)}$ yields a minuscule P-value. Similarly comparing Cantabria and Barcelona yields a LR test statistic of 22.32 and a $P = 0.0002$. Thus (as a glance at the histograms will confirm) there is strong evidence of differences in the size distributions of the two regions in the USA, and similarly of the two regions in Spain. Comparisons between California and

Barcelona, and between W. Virginia and Cantabria also yield very small P-values, and thus significant differences within each pair.

4 OTHER MODELS FOR THE FORMATION OF SETTLEMENTS.

The double Pareto-lognormal distribution for settlement size is based on the assumption that settlements were formed in a Yule process, for which the probability of a new settlement in an infinitesimal time increment is proportional to the current number of settlements. This is a stochastic version of exponential growth (in the number of settlements). As mentioned in Sec. 2 another possible model would be that settlements were formed in a Poissson process, over the last τ years, for which the probability of a new settlement in an infinitesimal time increment is the same at all times. For this model the time T since formation of a randomly chosen settlement is uniformly distributed on $[0, \tau]$. The corresponding distribution for the size $\bar{X}$ of any settlement can be derived. The resulting density (Reed, 1999) looks very similar to that of the dPlN, being unimodal and exhibiting Paretian behaviour in the upper tail if $\mu > 0$ (and in the lower tail if $\mu < 0$). It has four parameters (in addition to τ), which can be estimated by maximum likelihood, using a log-likelihood of the form (4) with the appropriate density. For the datasets for the four regions in Sec. 3, the maximized log-likelihood for this model was in all cases (using a variety of values of τ) somewhat less than that for the dPLN model, indicating less evidence to support this model (Royall, 1997).

However in some cases (*e.g.* Barcelona) the reduction was not large.

It is of course possible to postulate other models for the formation of settlements, and corresponding distributions for the random variables T and $\bar{X}$. However the more complicated the model, the more difficult it becomes to determine closed-form expressions for the distribution of the random variable T, and the corresponding distribution of current size $\bar{X}$. This will not be pursued further, save for noting that the dPlN model would still pertain in the case in which settlements were formed following a Yule process for a certain period (say the time interval $[\tau_1, \tau_2]$) with no new settlements since that time. In this case the mixing distribution for T would be a shifted exponential distribution. The resulting distribution of current size, $\bar{X}$, would still be dPlN with p.d.f. given by (3). The only difference would be that the parameters μ_0, σ_0^2 would not now represent the mean and and variance of the initial size of settlements. Rather they would represent the corresponding parameters for the *current* size of settlements established at the end of the foundation period (*i.e.* at time τ_2). This could be an explanation for the rather large values of the ML estimates of μ_0 obtained in Sec. 3.

5 CONCLUSIONS.

The main result of this paper is to show that there is a simple explanation for the observed phenomenon known as the rank-size law regarding the size distribution of human settlements. In addition to explaining the rank-size law, the analysis predicts the existence of a reverse rank-size law and yields a parametric form for the size disrtribution over its full range. An examination of the empirical size distributions for four regions confirms the predicted lower-tail rank-size property, as well as indicating an exceptionally good fit for the theoretically derived parametric form, thereby lending support to the model and to the explanation for the rank-size law.

The reason behind the rank-size phenomenon is shown to be essentially mathematical. It does not require economic or geographic assumptions relating to natural resources, production, consumption, communications, congestion *etc.* This does not mean that these factors are not important in the evolution of human settlements (indeed few would disagree that these are major factors influencing the growth of a region or city). Rather it means that, from a certain 'macro' point of view, variations in these factors across settlements and over time, can be viewed as following a certain distributions, and that their compound effect on the foundation and evolution of any particular settelement can be viewed as essentially random components in a stochastic model.

This is analogous to the way in which the central limit theorem can be used to explain the widespread occurrence of the normal (Gaussian) distribution in Nature (*e.g.* human heights, tree diameters, fishes length *etc.*). By recognizing that the normal distribution results from the summation (or integration) of large numbers of essentially independent random increments, does not mean that physical and biological factors (climate, nutrient availability, competition *etc.*) are not important in determining tree diameters and fishes lengths. On the contrary they are of primary importance. However it is the variability in these factors over time and between trees, fishes *etc.* which leads to variable growth, and it is the *aggregation* of the effects of these variable factors that leads to the normal distribution. In a similar way it is the aggregation of variable geographic and economic factors which leads to rank-size property for settlement size.

To explain the size of a *particular* city one needs to look at the details of its geography, its economic evolution and many other things (just as to explain the diameter of a particular tree one needs to look at physical and biological factors which have affected its growth). However to explain the *distribution* of sizes of settlements this turns out not to be necessary, because of mathemetical results concerning the interactions of factors which vary over time and from settlement to settlement and which can thereby be regarded as essentially random. The main difficulty in doing this is a modelling one - *i.e.* specifying plausible stochastic models describing the variability in the foundation and evolution of settlements. This has been accomplished by using a Yule process to describe foundations and Gibrat's law for the subsequent evolution of settlement size.

The connection with the central limit theorem goes beyond analogy. For objects following Gibrat's law the proportional rate of growth varies non-systematically with time. This results in the size of the object after a *fixed* time following a lognormal distribution (in the logarithmic scale the central limit theoerem ensures that the sum of many variable components converges to a normal distribution). The reason why settlement sizes do not follow a lognormal distribution, is that they have *not* all been following Gibrat's law *for the same length of time*. Thus the overall distribution of sizes is a mixture of lognormal distributions.(4) The mixing parameter is the time T since foundation. In the paper a plausible model for the foundation of settlements has been used leading to an exponential distribution for T and to the double Pareto-lognormal distribution for settlement size.

Although the specification of the growth process and the foundation process have been quite specific (geometric Brownian motion and Yule process, respectively), it is quite possible that similar results would hold with more loosely specified processes. For example using a uniform mixing distribution, resulting from a Poisson process model for the foundation of settlements, apparently does not greatly affect the qualitative properties of the resulting size distribution. It seems quite plausible that similar results could hold for other mixing distribution, arising from other settlement foundation processes. Geometric Brownian motion can be thought of as a convenient approximation for other 'geometric' (*i.e.* multiplicative) processes, in which the logarithm of size evolves with increments having some common distribution *e.g.* a geometric random walk, or a geometric Poisson jump process. No doubt the growth of human settlements often involves periods of rapid growth, interspersed with with slack periods of little or no growth, or even decline. However the essential 'geometric' property (for which proportional rates are the essential random quantities) is likely to hold, and thus GBM is likely to provide a reasonable approximation.

Another implicit assumption in the models is that of independence - that settlements are established and grow independently of one another. Undoubtedly this is not exactly true. However it can serve as a first approximation and since it leads to a model which, with all of its oversimplifications, leads to a distribution which fits the data well, it is perhaps not necessary to be more elaborate.

Like any good model, the one considered here leaves out more than it includes. However it does apparently capture the essence of the underlying mechanism behind the rank-size phenomenon. Simply put, the main claim of this paper is that the rank-size phenomenon can be explained by the fact that settlements have been growing in a varying, geometric way (*i.e.* with varying growth rates) for different lengths of time, and that when this fact is included it can lead (as Steindl pointed out more than 30 years ago) to a distribution exhibiting the familiar rank-size phenomenon for the largest settlements, as well as fitting empirical size distributions over the full range of sizes and predicting a rank-size phenomenon in the lower tail.

Acknowledgements. I gratefully acknowledge the useful comments of Dr. Harry Clarke (La Trobe University) and Dr. Mary Amiti (University of Melbourne).

Notes.

1. If the observations are not independent, the true log-likelihood will not be of the form (4). However one can still justify parameter estimates obtained by maximizing (4) as those which minimize the Kullback-Leibler information of the data with respect to the model. Barndorff-Nielsen (1977) refers to the procedure as *maximum likeness* estimation.

2. **http://www.census.gov/population/www/estimates/cityplace.html.** A web page of *U.S. Census Bureau.*
3. **http://www.ine.es/htdocs/inre/inre51/pobframe.htm.** A web page of Spain's *Instituto Nacional de Estadística.*
4. It is worth noting that Steindl (1965, 1968) identified the interaction between the evolution and foundation processes as the key to explaining Paretian behaviour in the upper tail of observed size distributions of firms and cities.

REFERENCES.

Allen, P. M. and M. Sanglier. 1979. "A Dynamic Model of Growth in a Central Place System," *Geographical Analysis*, 11, 256-272.

Allen, P. M. and M. Sanglier. 1981. "A Dynamic Model of a Central Place System - II," *Geographical Analysis*, 13, 149-164.

Anon. 1997 *S-PLUS 4 Guide to Statistics,* Seattle, WA: MathSoft, Data Analysis Products Division.

Barndorff-Nielsen, O. 1977. "Exponentially Decreasing Distributions for the Logarithm of Particle Size." *Proceedings of the Royal Society of London, A,* 353, 401-419.

Brakman, Stephen, Harry Garretsen, Charles Van Marrewijk and Marianne van den Berg. 1999. "The Return of Zipf: Towards a Further Understanding of the Rank-Size Distribution," *Journal of Regional Science,* 39, 183-213.

Champernowne, D. G.1953. "A Model of Income Distribution," *Economic Journal*, 63, 318-351.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation," *The Quarterly Journal of Economics,* 114,739-767.

Gibrat, Robert. 1931 *Les inégalités économiques* Paris, France: Librairie du Recueil Sirey.

Guérin-Pace, France. 1995. "Rank-Size Distribution and the Process of Urban Growth," *Urban Studies*, 32, 551-562.

Johnson, Norman. L., Samuel Kotz and Adrienne W. Kemp. 1993.*Univariate Discrete Distributions, Second Edition,* New York, NY: John Wiley and Sons.).

Mulligan, Gordon F. 1984. "Agglomeration and Central Place Theory: a Review of the Literature," *International Regional Science Review*, 9, 1-41.

Reed, William J. 1999, "The Pareto Law of Incomes - an Explanation and an Extension," submitted to *Journal of Business and Economic Statistics.*

Reed, William J. 2000 (a), "The Double Pareto- Lognormal Distribution," submitted to *Journal of Applied Probability.*

Reed, William J. 2000 (b), "The Pareto, Zipf and other Power Laws," submitted to *Economics Letters.*

Rosen, Kenneth T. and Mitchel Resnick, 1980. "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," *Journal of Urban Economics*, 8, 165-186.

Sheppard, Eric. 1982. "City Size Distributions and Spatial Economic Change," *International Regional Science Review*, 7, 127-151.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions," *Biometrika*, 42, 425-440.

Steindl, Josef. 1965. *Random Processes and the Growth of Firms: A Study of the Pareto Law.* London: Griffin; New York: Haffner.

Steindl, Josef. 1968. "Size Distributions in Economics," in *International Encyclopedia of the Social Sciences,* Vol 14, Silks, ed. New York: Macmillan and the Free Press.

Suh, S. H. 1987. "On the Size Distribution of Cities: an Economic Interpretation of the Pareto Coefficient," *Environment and Planning A,* 29, 749-762.

Yule, G. Udny. 1924. "A Mathematical Theory of Evolution Based on the Conclusions of Dr. J. C. Willis, F.R.S.," *Philosophical Transactions B*, 203, 21-87.

Zipf, George K. 1949 *Human Behavior and the Principle of Least Effort* Cambridge, MA: Addison-Wesley.

APPENDIX

Derivation of the distribution of the current size of settlements.

Under the assumptions of the model of Sec. 2, the size, $\tilde{X}_T$, of a settlement founded a *fixed* time T years ago, with an initial size of X_0, has a lognormal distribution, with

$$\tilde{Y}_T = \ln(\tilde{X}_T) \sim N\left(X_0 + (\mu - \frac{\sigma^2}{2})T,\ \sigma^2 T\right),$$

(equation(2)). If the initial size X_0 follows a lognormal distribution with parameters μ_0 and σ_0^2, then size X_T is lognormally distributed with

$$Y_T = \ln(X_T) \sim N\left(\mu_0 + (\mu - \frac{\sigma^2}{2})T,\ \sigma_0^2 + \sigma^2 T\right) \tag{5}$$

The moment generating function (m.g.f.) of Y_T is

$$M_{Y_T}(\theta) = E\left(e^{\theta Y_T}\right) = \exp\left(\mu_0\theta + \sigma_0^2\theta^2/2 + \left[(\mu - \frac{\sigma^2}{2})\theta + \sigma^2\theta^2/2\right] T\right)$$

The distribution of the logarithm of the current size of a *randomly selected* settlement, $\bar{Y}$, say, can be obtained by integrating the density of Y_T with respect to the distribution of T. Alternatively its m.g.f. can be found, using conditional expectations, as

$$M_{\bar{Y}}(\theta) = \mathrm{E}\left(\exp(\theta\bar{Y})\right) = \mathrm{E}_T\left(\mathrm{E}_{Y|T}\left(\exp(\theta Y)\right)\right) = \mathrm{E}_T\left(M_{Y_T}(\theta)\right)$$

which from the above can be written

$$M_{\bar{Y}}(\theta) = \exp\left(\mu_0\theta + \sigma_0^2\theta^2/2\right)\ M_T\left((\mu - \frac{\sigma^2}{2})\theta + \sigma^2\theta^2/2\right).$$

If T is exponentially distributed with parameter λ, then the m.g.f. of T is $M_T(\theta) = \frac{\lambda}{(\lambda-\theta)}$ from which it follows that

$$\begin{aligned} M_{\bar{Y}}(\theta) &= \frac{\lambda \exp\left(\mu_0\theta + \sigma_0^2\theta^2/2\right)}{\lambda - (\mu - \frac{\sigma^2}{2})\theta - \frac{\sigma^2}{2}\theta^2} \\ &= \exp\left(\mu_0\theta + \sigma_0^2\theta^2/2\right) \frac{\alpha\beta}{(\alpha-\theta)(\beta+\theta)} \end{aligned}$$

where α and $-\beta$ $(\alpha, \beta > 0)$ are the two roots of the characteristic (quadratic) equation

$$\frac{\sigma^2}{2}z^2 + (\mu - \frac{\sigma^2}{2})z - \lambda = 0. \tag{6}$$

Now it is easily confirmed that $\frac{\alpha\beta}{(\alpha-\theta)(\beta+\theta)}$ is the m.g.f. of the *double exponential distribution* (or asymmetric Laplace distribution) with density function

$$f(x) = \begin{cases} \frac{\alpha\beta}{\alpha+\beta}e^{\beta x} & \text{if } x < 0 \\ \frac{\alpha\beta}{\alpha+\beta}e^{-\alpha x} & \text{if } x \geq 0 \end{cases}$$

Also since $\exp\left(\mu_0\theta + \sigma_0^2\theta^2/2\right)$ is the m.g.f of an $N(\mu_0, \sigma_0^2)$ random variable, it follows that the distribution of $\bar{Y}$ can be represented as that of the sum of independent normal and double exponential random variables. From this it follows that the distribution of $\bar{X} = e^{\bar{Y}}$ can be represented as that of the product of independent random variables, U and V say, one (U) with a lognormal distribution and the other (V) with a *double Pareto* (dP) distribution with p.d.f

$$f(v) = \begin{cases} \frac{\alpha\beta}{\alpha+\beta}\, v^{\beta-1}, & \text{for } v \leq 1 \\ \frac{\alpha\beta}{\alpha+\beta}\, v^{-\alpha-1}, & \text{for } v > 1 \end{cases} \tag{7}$$

which is the distribution of the exponentaial of a random variable with the above double-exponential distribution. This representation of the distribution of $\bar{X}$ as the product of lognormal and double Pareto components is the reason for the name *double Pareto-lognormal* (dPlN) distribution (see Reed, 2000 for more on properties of the dPlN distribution).

The p.d.f. of $\bar{X}$ can be obtained from the p.d.f. of $\bar{Y} = \ln(\bar{X})$ which in turn can be found by convolving a double exponential density with a normal density. The details are tedious and are omitted. The result is

$$\begin{aligned} f_{\bar{Y}}(y) = \tfrac{\alpha\beta}{\alpha+\beta} \Big[& e^{-\alpha(y-\mu_0)+\alpha^2\sigma_0^2/2}\Phi\left(\tfrac{y-\mu_0-\alpha\sigma_0^2}{\sigma_0}\right) + \\ & e^{\beta(y-\mu_0)+\beta^2\sigma_0^2/2}\Phi^c\left(\tfrac{y-\mu_0+\beta\sigma_0^2}{\sigma_0}\right)\Big] \end{aligned} \tag{8}$$

from which the p.d.f (3) of $\bar{X}$ in Sec. 2 follows. The observed sizes of n settlements can be thought of as the realizations of n independent, identically distributed random variables all with the above distribution. The log-likelihood for such observations is thus (4).

Consider now the situation in which the distribution of the initial size of settlements evolves in time. If at some base reference time it followed a lognormal distribution with mean and variance parameters A_0 and B_0^2 and it subsequently evolved in time following the GBM

$$dX_0 = a_0\ X_0 dt + b_0\ X_0\ dW_0,$$

then t time units after the reference time it would follow a lognormal distribution with mean and variance parameters $A_0 + (a_0 - b_0^2/2)t$ and $B_0^2 + b_0^2 t$. If the base reference time is, say, τ time units before the present time, the distribution of the starting size of a settlement founded T time units ago will be lognormal with mean and variance parameters $A_0 + (a_0 - b_0^2/2)(\tau - T)$ and $B_0^2 + b_0^2(\tau - T)$. Replacing μ_0 and σ_0^2 above by these quantities, leads to X_T having a lognormal distribution with parameters $A_0 + (a_0 - b_0^2/2)\tau + (\mu - a_0 - \frac{\sigma^2 - b_0^2}{2})T$ and $B_0^2 + b_0^2\tau + (\sigma^2 - b_0^2)T$.

This is of the same form as (5), with both the mean and variance parameters being linear functions of T (with μ replaced by $\mu - a_0$; σ^2 replaced by $\sigma^2 - b_0^2$; μ_0 replaced by $A_0 + (a_0 - b_0^2/2)\tau$; and σ_0^2 replaced by $B_0^2 + b_0^2\tau$). With these replacements the derivation of the double Pareto-lognormal distribution for the current size of a randomly selected settlement follows as before.

TABLE 1: Maximum likelihood estimates of the four parameters (α, β, μ_0 and σ_0) for the double Pareto-lognormal distribution fitted to the empirical size distributions for the four regions discussed in Sec. 3.

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}_0$	$\hat{\sigma}_0$
W. Virginia '98	1.09	4.71	6.26	0.883
California '98	1.87	0.991	10.43	0.861
Cantabria '96	1.08	1.46	7.20	0.453
Barcelona '96	1.28	3.03	7.19	1.67

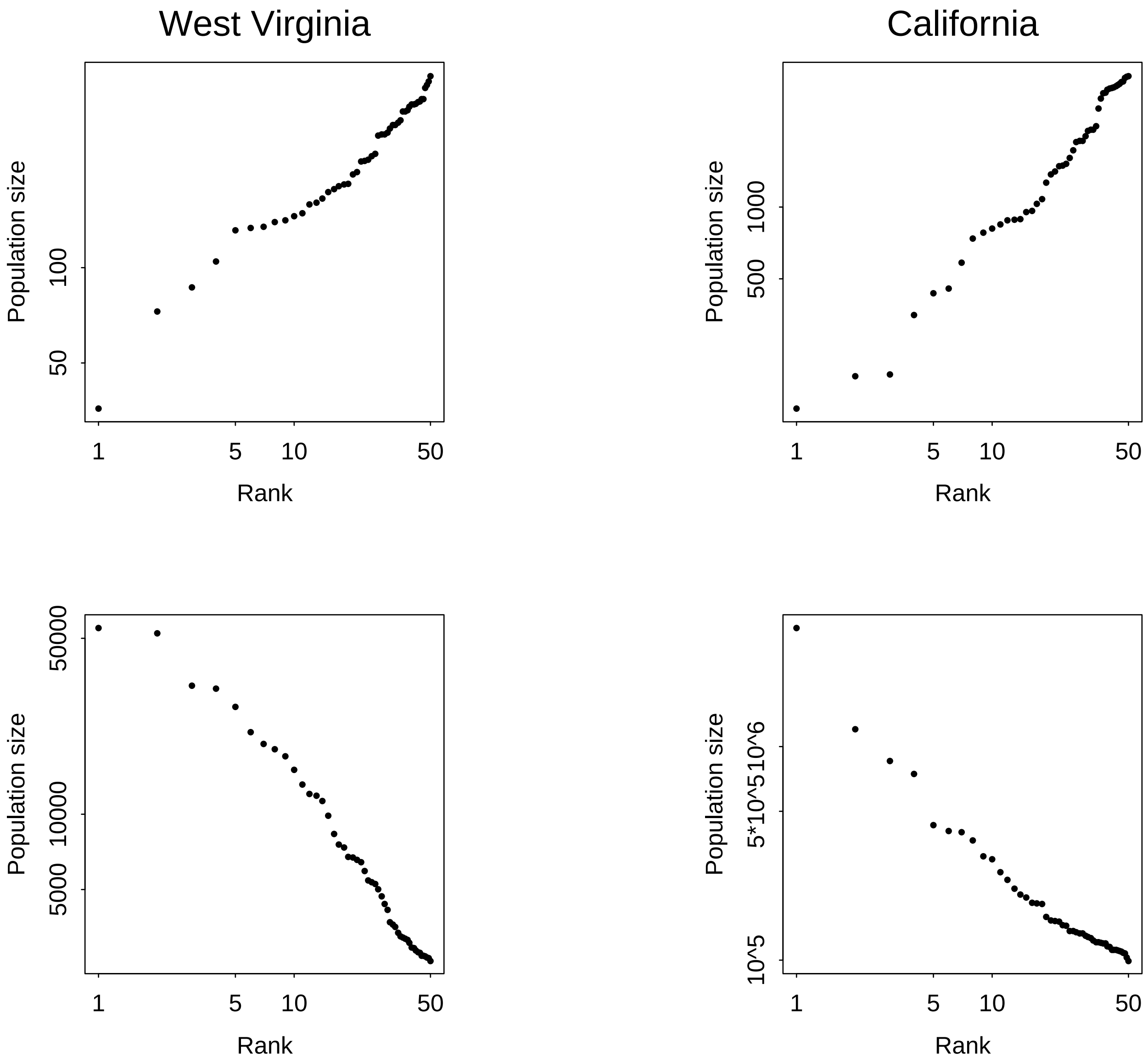
Figure captions.

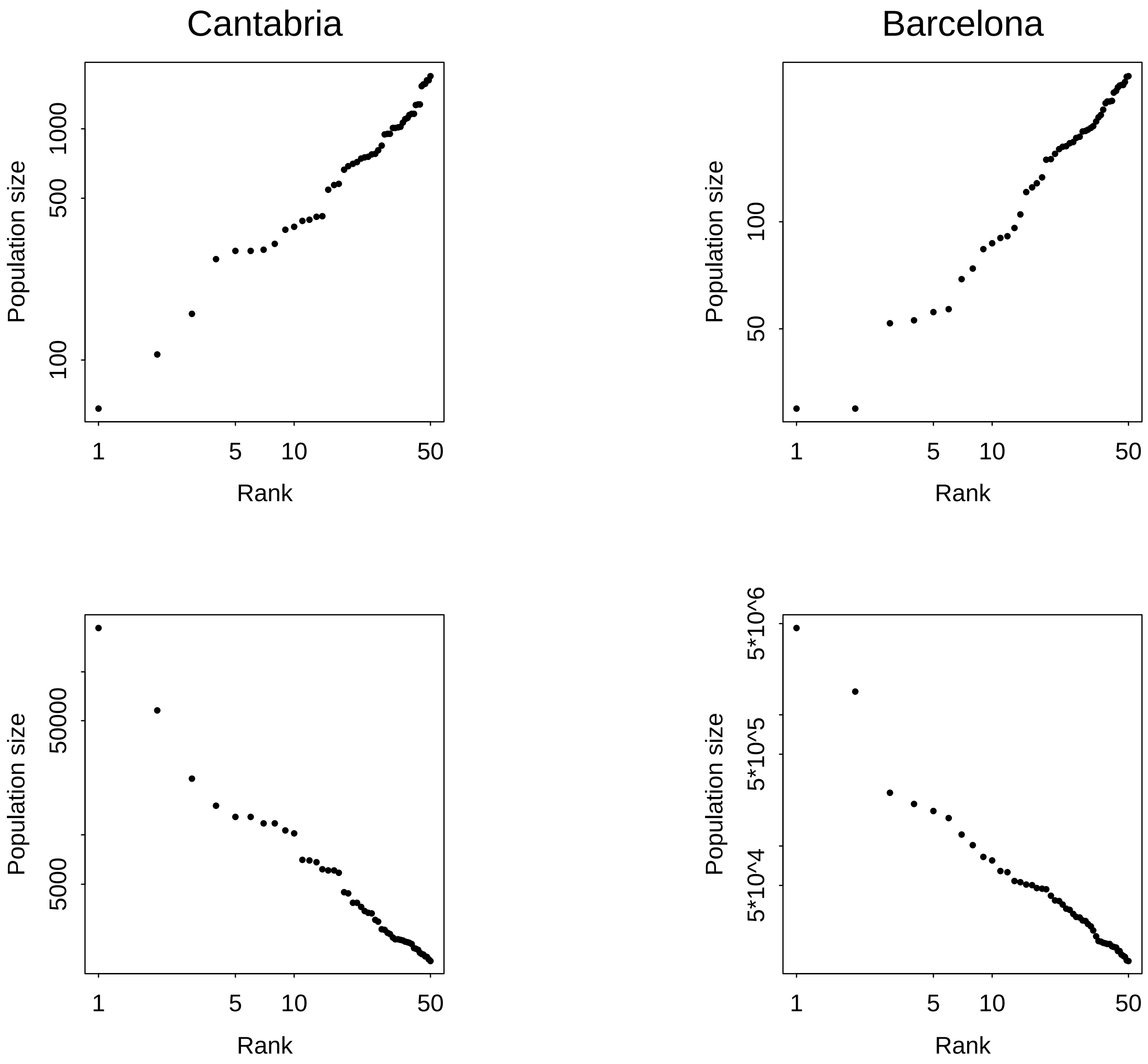
Fig. 1 Rank-size plots for two American states. The top row shows lower-tail rank-size plots (population-size vs. ascending rank in logarithmic scales) for the smallest 50 settlements in respectively West Virginia and California. The bottom row shows the more familiar upper-tail rank-size plots (population size vs. descending rank in logarithmic scales) for the largest 50 settlements in each of the two states. The linear fit is at least as good in the lower tail as in the upper tail.

Fig. 2 Rank-size plots for two Spanish provinces. The top row shows lower-tail rank-size plots (population-size vs. ascending rank in logarithmic scales) for the smallest 50 settlements in respectively Barcelona and Cantabria provinces. The bottom row shows the more familiar upper-tail rank-size plots (population size vs. descending rank in logarithmic scales) for the largest 50 settlements in each of the two provinces. The linear fit is at least as good in the lower tail as in the upper tail.

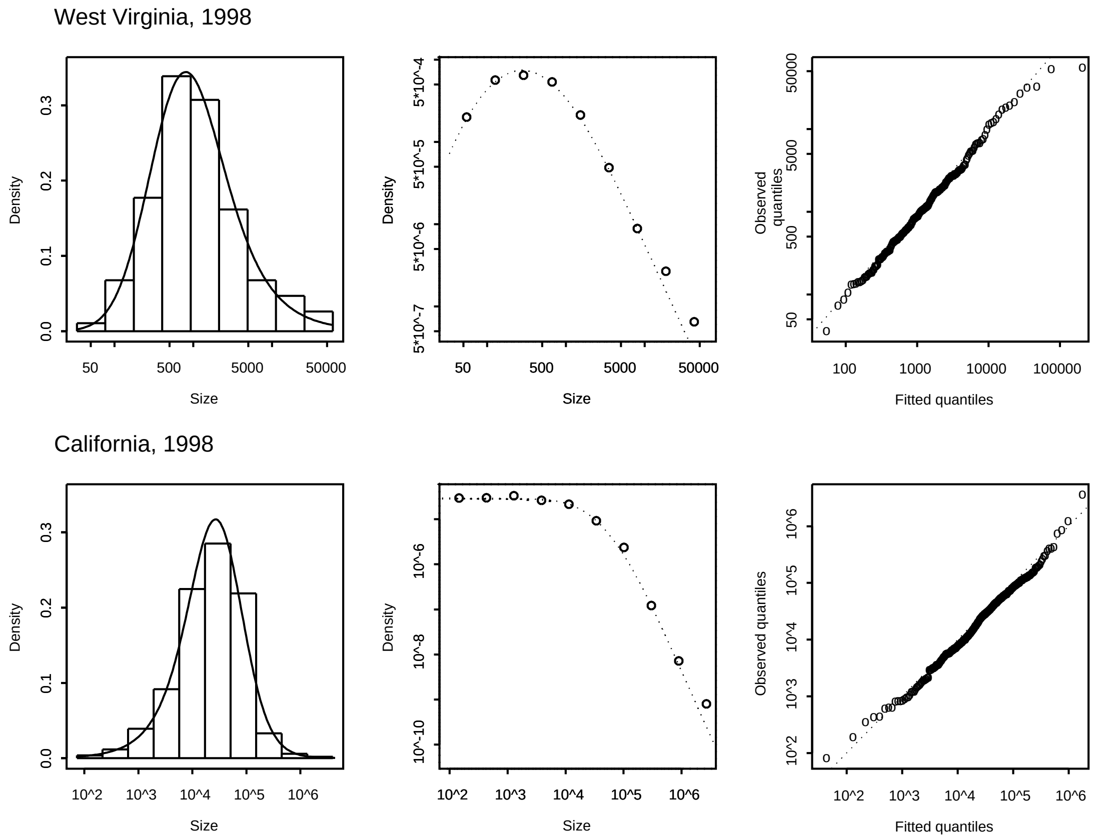
Fig. 3 The double Pareto-lognormal model fitted to W. Virginia data (top row) and to California data (bottom row). The three panels in each row show (from left to right): the fitted density superimposed on a histogram of size (logarithmic scale); the fitted density (dotted line) plotted against size (both on logarithmic scales) superimposed on empirical density; Q-Q plot of observed quantiles against fitted quantiles, both on logarithmic scales, with a 45^{O} line (dotted).

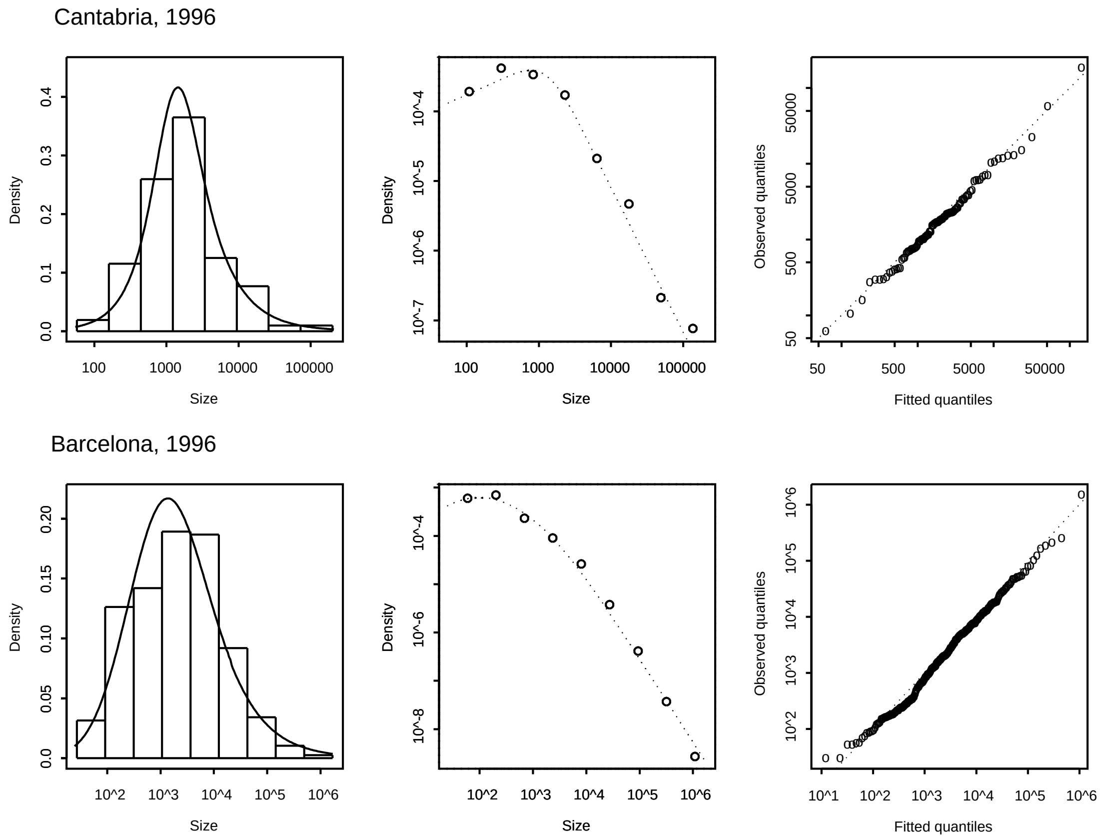
Fig. 4 The double Pareto-lognormal model model fitted to Cantabria data (top row) and to Barcelona province data (bottom row). The three panels in each row show (from left to right): the fitted density superimposed on a histogram of size (logarithmic scale); the fitted density (dotted line) plotted against size (both on logarithmic scales) superimposed on empirical density; Q-Q plot of observed quantiles against fitted quantiles, both on logarithmic scales, with a 45^{O} line (dotted).


West Virginia
California
Population size
100
50
1000
500
50000
10000
5000
5*10^5
10^6
10^5
1
5
10
50
Rank



Cantabria
Barcelona
Population size
1000
500
100
1
5
10
50
Rank
Population size
100
50
1
5
10
50
Rank
Population size
50000
5000
1
5
10
50
Rank
Population size
5*10^6
5*10^5
5*10^4
1
5
10
50
Rank


West Virginia, 1998
Density
0.0
0.1
0.2
0.3
50
500
5000
50000
Size
Density
5*10^-7
5*10^-6
5*10^-5
5*10^-4
50
500
5000
50000
Size
Observed quantiles
50
500
5000
50000
100
1000
10000
100000
Fitted quantiles
California, 1998
Density
0.0
0.1
0.2
0.3
10^2
10^3
10^4
10^5
10^6
Size
Density
10^-10
10^-8
10^-6
10^^2
10^3
10^4
10^5
10^6
Size
Observed quantiles
10^2
10^3
10^4
10^5
10^6
10^2
10^3
10^4
10^5
10^6
Fitted quantiles


Cantabria, 1996
Density
0.0
0.1
0.2
0.3
0.4
100
1000
10000
100000
Size
Density
10^-7
10^-6
10^-5
10^-4
100
1000
10000
100000
Size
Observed quantiles
50
500
5000
50000
50
500
5000
50000
Fitted quantiles
Barcelona, 1996
Density
0.0
0.05
0.10
0.15
0.20
10^2
10^3
10^4
10^5
10^6
Size
Density
10^-8
10^-6
10^-4
10^2
10^3
10^4
10^5
10^6
Size
Observed quantiles
10^2
10^3
10^4
10^5
10^6
10^1
10^2
10^3
10^4
10^5
10^6
Fitted quantiles