Environment and Planning A, 1985, volume 17, pages 1637 - 1652

Lognormal estimates of macroregional city-size distributions, 1950 - 1970

L De Cola
Department of Geology and Geography and Regional Research Institute, West Virginia University, Morgantown, WV USA 26506
Received 30 January 1985; in revised form 21 June 1985

Abstract. A three-parameter lognormal model is used to estimate the city-size distribution of the world and of eight UN-defined macroregions. The model is found to fit the data better than the Pareto function, and to provide a powerful means of comparing distributions among regions. Although system concentration (measured by the standard deviation index) is relatively stable in Europe and in the world at large, it is decreasing in North America, Africa, and East Asia, and increasing in Latin America and South Asia. Cities in the 250000 - 500000 size class are somewhat more numerous than predicted, suggesting the possibility of some kind of 'optimum'. The theory of extreme values is used to predict the most populous city of a region and to compare predictions with actual maxima, demonstrating that the largest cities in the world are well within systematic possibilities.

1 Introduction

The theory of fractal geometry as applied to spatial phenomena (Mandelbrot, 1983, chapters 29, 38) and the law of proportionate effect as applied to size distributions (Vining, 1984) are among processes which give rise to sets, the sizes of whose elements show an inverse, roughly hyperbolic association between frequency and magnitude (see Table 1 below). A large and complex collection of models has been developed to explain and describe this association as it applies to the populations of cities within regions. One possible taxonomy of these models may be found in another paper (De Cola, 1985a); Mills (1972) and Richardson (1973) give an introductory discussion and Carroll (1982) a comprehensive survey of empirical work. Dominant among the stochastic models that describe the nature of this relationship in a dynamic way are the lognormal and Pareto processes, compared by Quandt (1964) and later by Parr and Suzuki (1973) both for US data, and by Okabe (1979). It is the contention of this paper that, at least for the post World War 2 size distribution of large cities in macroregions (continents or subcontinents), the three-parameter lognormal model is more successful (figure 1) and, perhaps more important, better reflects the major descriptive dimensions of city-size distributions.

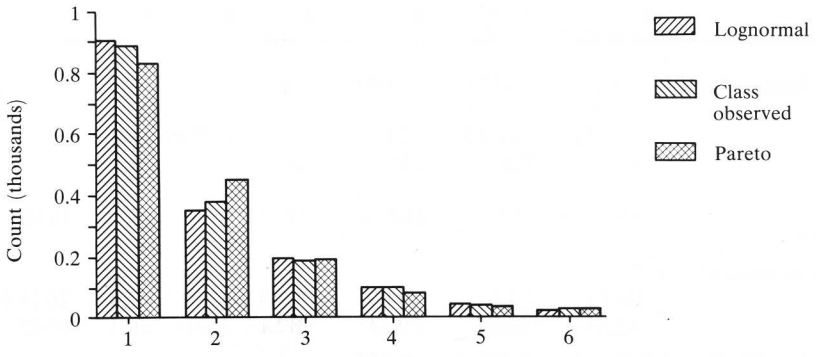

Figure 1. Predicted and observed class counts, for world cities 1970.

In the paper a theoretical development of the problem of estimating parameters from a lognormal distribution is introduced. This theory is then used in an empirical description of 1950–1970 world and macroregional city-size distributions (UN, 1980, pages 48–52) in terms of three parametric indices of extent (number of cities), scale (distributional location), and heterogeneity (distributional dispersion). Time series of these indices are next examined in the light of current theories of urban system development, and predicted maximum settlement size is also used to compare systems. The paper concludes with a summary of findings and a recommendation that regional scientists adopt the three-parameter lognormal model as the standard specification of settlement-size distributions.

2 Theory

We begin with an arbitrarily defined region containing N settlements, each with population x. Let the settlements be ordered from the smallest to the largest and indexed by i, their position in the order: $x_1, x_2, ..., x_N$. As in the present case of table 1, let the data consist of K ascending quantiles or (as they will be called) class limits: $0 < x^1 < x^2 < ... < x^K$ which are the boundaries of half-open intervals such that n^k is the cumulative class count of settlements more populous than x^k so that $N \geqslant n^1 \geqslant n^2 \geqslant ... \geqslant n^K$. The empirical distribution of x is therefore

$$P(x \leqslant x^k) = 1 - \frac{n^k}{n^1}, \qquad \text{for } x > x^1, \tag{1}$$

where x^1 is the lower bound of the data that results from our not having a complete enumeration of the city-size distribution.

Fitting a two-parameter Pareto function to these data (Quandt, 1966) calls for finding estimates $\hat{A}$ and $\hat{b}$ of the parameters of the equation

$$\frac{n^k}{n^1} = A(x^k)^{-b}. \tag{2}$$

Such an approach usually yields satisfying results by the R^2 criterion, although a plot of residuals against x often shows a distinct inverted-U shape (Vining, 1976). Table 1 presents class count predictions for the Pareto estimation, and figure 1 compares these counts with the actual and lognormal predictions.

The major competing paradigm—the lognormal model—is somewhat more difficult to estimate because, except for the simplest fully enumerated regions, we do not

Table 1. 1970 world city-size distribution and estimates (source: UN, 1980, page 48)

	Class (k)						Total
	1	2	3	4	5	6	
Distribution							
Population lower limit (millions) (x^k)	0.1	0.25	0.5	1	2	4	
ln (x^k)	11.51	12.43	13.12	13.82	14.51	15.20	
Cumulative class count (n^k)	1615	726	345	159	63	24	
Observed class count ($n^k - n^{k+1}$)	889	381	186	96	39	24	1615
Estimates of class counts ($\hat{n}^k - \hat{n}^{k+1}$)							
Lognormal[a]	904.8	353.1	194.7	97.4	42.5	22.4	1614.9
Pareto[b]	828.1	451.2	192.5	82.1	35.0	26.1	1615

[a] Based on $\hat{\mu} = 11.677$, $\hat{\sigma} = 1.590$; $\chi^2 = 3.293$, $p = 0.655$.
[b] Based on $\hat{A} = \exp 14.56$, $\hat{b} = 1.229$; $\chi^2 = 18.59$, $p = 0.0024$.