

Reprinted from ECONOMIC DEVELOPMENT AND CULTURAL CHANGE
Vol. 33, No. 1, October 1984
© 1984 by The University of Chicago. All rights reserved.
Printed in U.S.A.

Statistical Determinants of the Population of a Nation's Largest City*

Lee De Cola
University of Ibadan

I. Introduction

In spite of the fact that urbanization on a world scale is the dominant spatial demographic phenomenon of the twentieth century, social scientists manifest a striking lack of unanimity about the nature, causes, and possible future of the urban revolution. What is clear to all observers, however, is that the continuing growth of cities is an extremely uneven phenomenon: while an overall urbanization rate for a nation may reflect the general growth of all urban places, classes of cities may experience a wide range of rates, with individual places showing even more heterogeneous histories.¹

This paper is an attempt merely to understand the tip of the huge and growing urban iceberg. The question I wish to answer is, What are the important statistical determinants of the population of a nation's largest city? The very asking of this question calls for a statement of my normative and scientific reasons for the investigation. On the one hand I shall make no judgment whether a given city may, with respect to some set of criteria, be too big or small. Nevertheless, it will be clear below that some number-one cities are relatively very much larger than others even when their characteristics and national environment are taken into account. On the other hand, there are sound scientific reasons (such as intriguing patterns in the data) for focusing analysis on the upper tail of the national urban size distribution.² Furthermore, because a country's largest city often contains one-third or more of its urban population, there are strong policy motivations for this investigation.

The analysis presented below has four goals. First, the explication of links between certain variables and city population is intended to be a contribution to a theory about what factors influence urban size in general and the size of CITY1 (the largest city in the nation) in particu-

lar.³ The approach adopted here is quite straightforward: to introduce a set of predictor variables into a regression model to see how well they predict LOGPOP1 (the logarithm of the population of CITY1). If urban science is at all well developed, there should be few surprises when this is done.

A second goal of the analysis is to examine the residuals from the model in detail to attempt to discover why a given city is larger or smaller than the model predicts. A medical analogy may be suggestive here. When a patient visits a doctor the first questions the doctor asks relate to symptoms. Often the symptoms are not important in themselves but only as indicators of some underlying process not readily apparent to the patient. Hence the "primacy syndrome" receives more attention than it may deserve in itself because it may point to underlying problems in the national distribution of population.⁴ The analysis reported below will add another diagnostic tool to the study of primacy by generating regression residuals that may be treated as measures of relative urban concentration.

Related to the second goal is a third, to examine the outliers among the residuals for anecdotal insights into the processes generating urban growth. Just as the pattern of residuals throws light on the process of urban concentration, so may the extreme cases show where the processes break down because of "pathological" factors.⁵

A final goal is to direct this analytical light on Africa, a continent that is still relatively unurbanized but is experiencing the world's highest rates of urban growth, and the continent where colonialism continues to play a visible role in urban processes.⁶ Until recently Latin America and Asia have been the favorite laboratories for students of rapid urbanization and large cities, but improved data—and rising concern—demand that the regions of Africa, with their richly varied histories and amazingly diverse current developmental paths, receive much more attention than they have in the past.

Because the word will figure in the analysis that follows, I must begin with a coherent statement of what precisely shall be denoted by "primacy." There seem to be at least three meanings. First, to follow the Latin etymology, CITY1—whatever its size—experiences primacy by being the primate settlement in the national set. Second, there are occasionally cities other than CITY1 (like Jerusalem, Rio de Janeiro, and Peking) which occupy a preeminent place in the urban functional hierarchy and therefore partake of primacy as well. Third, the word has been used in the urban geographical literature more selectively to connote the "excessive" size of CITY1 often observed in the national urban size distribution. The question naturally arises, Excessive with respect to what? Usually the criterion is the linear prediction of a rank-size line or the mean value of some ratio. I intend to expand the set of urban concentration measures with a new and arguably more robust

measure of the relative size of CITY1. I shall attempt, however, to avoid the use of the more value-laden term "primacy."

The organization of the paper is as follows. Section II develops a theoretical treatment of the extreme value problem as applied to urban size and presents an intuitively satisfying path-type model to predict urban population for any city. Section III describes the data used to calibrate the model and gives a brief picture of the international database. Section IV presents the principal findings: regression coefficients, residuals, outliers, and regional analysis; while Section V sets forth a few conclusions, policy implications, and suggestions for further research.

II. Theory and Model

Models predicting urban size have constituted a lively field of research during the twentieth century. Although the literature is large and varied,⁷ much of the work may be grouped under four major headings. First and perhaps oldest are systemic studies that attempt to predict the population of a city from its rank among the cities of a set, the theory being that the set forms a system, each of whose members has a somewhat determinate place in the ranking. Second are hierarchical models, such as those of economic central place theory and political-administrative organization, that predict size from function and service areas. Third and most diverse are stochastic models, among which are random splitting, Markov processes, and the lognormal process of growth. Finally there are more empirical studies of primacy and regional analyses, particularly in Latin America and Asia.

A full discussion of this broad field of research would take me much too far from the task at hand. It seems clear, however, that the most parsimonious model, the one requiring fewest assumptions, is the lognormal, which argues that $POP(i)$, where i is a given city (i need not be the rank), is a random variable whose logarithm is the normally distributed outcome of a stochastic birth and growth process. With this in mind, let us examine the global data.

Let us begin by imagining that we have compiled a list of all the settlements in the world, the problem of what determines a settlement having been solved.⁸ One way to understand the statistical distribution of urban sizes is to rank order this list of settlements by population. A recent United Nations study compiled data on all the cities of the world with an estimated population greater than 100,000 and arrived at the size distribution shown in table 1. If we assume that there are 8,000 cities in the world, then table 1 may be regarded as giving relative cumulative frequencies for the top 21%. A simple linear regression model allows us to estimate the mean and standard deviation of a lognormal distribution that predicts these percentages ($R = .9997$).⁹ The implication of this finding is that at least the largest cities of the

TABLE 1

SIZE DISTRIBUTION OF THE WORLD'S CITIES WITH POPULATION OVER 100,000 IN 1975

Population Interval	Number of Cities	Cumulative Number	Cumulative Percentage Assuming 8,000 Cities
Over 4,000,000	30	30	.4
2,000,000-3,999,999	48	78	1.0
1,000,000-1,999,999	107	185	2.3
500,000-999,999	227	412	5.1
250,000-499,999	441	853	10.7
100,000-249,999	802	1,655	20.7
Total	1655		

SOURCE—United Nations, *Patterns of Urban and Rural Population Growth* (New York: United Nations, 1980), p. 48.

world may be regarded as the tail of a lognormal distribution with a mean of 9.928 and a standard deviation of 1.965 (using natural logarithms).

The consequences of this assumed distribution are rather interesting. The model predicts that the largest city in the world should have a population of about 30,000,000 and the smallest (a very unreliable prediction for this model) a population of about 15, while the median settlement should contain about 20,000 people. But because the model predicts only 500 towns with a population less than 1,000 (6.2% of the settlements), it seems that the distribution applies to that part of the national population which may be called "urban" and not to all possible settlements. It is encouraging therefore to note that the total population residing in the 8,000 cities of the model is about 1.1 billion, not too far from current estimates of the world's urban population.¹⁰

From an empirical standpoint this discussion suggests that, whatever the underlying determinants of urban population, the lognormal model works quite well on a world scale and that the logarithmic transformation of population is justified (in addition to the obvious fact that it reduces the unmanageable skewness of the variable). Another feature of the transformation, which I will make much use of below, is that regression residuals of a logarithmically transformed dependent variable are the logarithms of the ratio of actual to predicted values.

Not much work has been done to model the world city population distribution because there is as yet no consensus on what kind of world system of cities exists—or whether one exists.¹¹ The discussion above is offered merely as a suggestion that, on a world scale, the lognormal model fits the data. But this is no proof of the existence of systemic processes. Except for the very largest cities, which certainly are becoming enmeshed in an ever-tighter network of financial, transportation, and communication links, the world distribution remains more

described than analyzed. Much attention, however, has been directed at urban populations for various partitions of the earth's surface—usually nations, but sometimes larger or smaller units such as continents or states.¹² Here again, the lognormal distribution usually fits the data, but the possibility of systemic relationships among the cities of a geographically connected subset has led to a great deal of work on “rank-size” models in which the size of a city is related to the number of cities larger than it.¹³

The focus of the present research is on the tip of the tail of the urban size distribution within a nation. The simplest model to explain POP1 would be based on the assumption that settlements are distributed randomly over the earth's surface and that nations are random partitionings of this surface.¹⁴ This model suggests that the extreme values of a nation's urban size distribution are simply the maximum values from spatially connected random samples. For example, if we assume 8,000 settlements with a given size distribution, then the expected maximum value for any country should be a function solely of that country's area. So the USSR should (probabilistically) contain the world's largest city because it represents the largest sample, followed by Canada, China, and so on.¹⁵ But although I have found that the zero-order correlation between area and LOGPOP1 (the population of CITY1) is .39, the very largest cities are found in such “unusual” places as Japan, Argentina, and France as well as the more extensive countries, suggesting the obvious fact that there are other important factors determining extreme values.

This problem—that the sizes of the largest cities may be in part the realization of a random process but that their spatial distribution is not random—suggests that we consider what specific factors determine the population of *any* city within a nation, accepting the fact that international differences reflect the influences of national characteristics. Before returning to a discussion of maxima, therefore, let us consider what these factors might be. This discussion is based on the theory that a model that predicts city size for any city should also work for CITY1, although it would have to be modified to take into account prior factors resulting in CITY1's being the largest.

I consider spatial variables first. National land area has already been mentioned as a pure spatial variable, but other factors, such as location within the national space, with respect to natural features, and in relation to other cities, should be included as well. Furthermore, because cities seem to be born, grow, and even die as part of what may be viewed as a self-organizing system—for if it is a system it certainly is not wholly exogenously organized—then such factors as the shape of a country might be significant determinants of location and, by implication, of size.¹⁶

Key socioeconomic variables (including political and technologi-

cal factors) constitute a second set of determinants of urban size. At the national level it would be expected that the rate of national population change is an important determinant of the rate of change, and therefore of the magnitudes, of the population of national cities. Economic development, as well, is known to have important but apparently nonlinear effects on urbanization and concentration, so that it is necessary to include at least one index of the extent to which the national environment is economically developed and technologically integrated.¹⁷ Included in this second group should also be those characteristics of the city itself that represent its economic and political role in the national urban system. At the minimum we should know whether the city is a national or regional capital and its place in whatever transportation and communication network may exist. Naturally, many other national and city indices could be added to a comprehensive model, but these factors will do for a start.

The third set of variables are the demographic factors likely to be of importance in predicting the shape of the national urban size distribution and especially the conformation of its upper tail. To borrow from the simple world model developed above, we seek factors that will be associated with three key parameters of the distribution. If, as seems likely, the national urban size distribution has a lower threshold, then the extent to which the nation is urbanized could stand for this lower limit.¹⁸ Next, the location of the distribution (its mean) will reflect the total national population, on the theory that big countries have big cities. Finally, the skewness of the distribution (which for the log-normal model is sensitive only to the standard deviation of the corresponding normal distribution) may be represented by the tendency of the urban population to be concentrated in the biggest cities.¹⁹ The factors that give a larger positive skew to the distribution not only are related to urbanization but also are reflected particularly in the sizes of the other largest cities. This suggests our examining the extent to which the population of the class of a nation's largest cities may be a determinant of the population of any one of them. If, for example, CITY2 is big then CITY1 will of course be bigger, but how much bigger is a reflection of the extent to which the national population is concentrated in the very largest city. (Another demographic index widely used in predicting the population of a city is its rank in the national distribution. As my attention will focus only on the population of CITY1, this index is only used implicitly here.)

Let us now take the factors predicting city size and organize them into a recursive system with causal implications. The eight concepts derived from the discussion above have 15 potential links. One way of interrelating these concepts is shown as a path-type model in figure 1. Each variable is causally related, directly or indirectly, to every other variable through a network of arrows, with the sign on each arrow

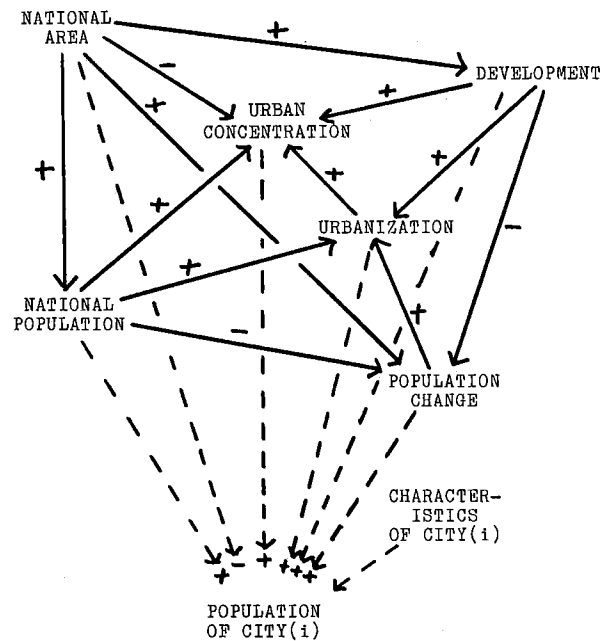


FIG. 1.—Hypothetical causal linkages among concepts

representing expected sign of the partial correlations given all prior variables. Although space does not permit the discussion of the full recursive system, at least three of the arrows deserve comment. The negative link between population and population change reflects simple negative feedback in which the mass of humans depresses further growth given fixed area. The negative link between population change and development reflects a crude assumption that development seems to be a shift from the production of people to the production of things. Finally, area negatively influences concentration by permitting large competing centers.

Three absent arrows also require explanation. No link is shown between national population and development because the relationship is likely to be nonlinear given constant area: in the early stages of population growth additions to national population will contribute to development, but it seems likely that really dense population concentrations, even on the national scale, will depress economic growth. Second, area is not tied to urbanization mainly because the demographic impact of space is expressed through the link to concentration, which is expected to be more sensitive to the availability of other megalopolitan sites than would be the development of cities per se. Third, there is no link from population change to concentration mainly because the present evidence suggests that population growth is quite

unevenly distributed among classes of cities.²⁰ In any case, the indirect link via urbanization seems adequate.

The model is quite general. For example, if $POP(k)$, the population of city of rank k , were at the focus the model would represent a rank-size analysis, correcting for city characteristics. In the present case, however, the diagram links key factors that may be operationally defined in a variety of ways to explain $POP1$.

The coherence of the model—the way it includes and interrelates the key variables—suggests that it should be successful in developing a straightforward way of predicting $POP1$, and at the same time reveal the signs and magnitudes of the linkages between the other indices. This model will inform the analysis below. It illustrates why I argue that the size of a nation's largest city is symptomatically useful in revealing the underlying urbanization and developmental processes.

III. The Data

The multivariate model discussed above was tested using data assembled from several recent sources, issued mainly by the United Nations. Although there is a great deal more detail about each country and the cities in the full database I have been using (such as populations and locations of each country's four largest cities, dates of settlement, and a number of explicitly political variables) the present model focuses on the indices most likely to have some explanatory power in predicting $LOGPOP1$. Descriptions, mean values, and sources for the variables are presented in table 2, while intercorrelations are shown in table 3.

All of the variables are operationalizations of the concepts from the model. $LOGLENGTH$ is, along with $AREA$, a crude measure of the shape of the national space. $LOGENERGY$ is a measure of economic development whose correlation with another measure, the logarithm of per capita GNP, is .92. $PCTURBAN$ is an estimate, according to each country's definition, of the percentage of the national population residing in urban areas. $LOGPOP2$ is used as a measure of concentration, for it represents the maximum value of the rest of the urban size distribution.

The sample of 126 nations includes every country on which the United Nations compiles data and which has at least one city with a population greater than 100,000 (see App. for a list of these cities). Because the spatial distribution of the largest cities is so far from random, the full database includes a complete rank-size distribution only for the first 19 cities in the world. Nevertheless, the countries represent about 97% both of the world's total population and of its inhabited area.

Because of this comprehensiveness, the meaning of the word "sample" is debatable here. In such cross-national research one must

TABLE 2
VARIABLES USED IN THE REGRESSION MODEL

Variable	Mean	Definition	Source
LOGPOPI	2.95	Log(base 10) of population of largest city in country in thousands, 1975 or latest	1
LOGAREA	2.45	Log of area of country in thousands of square kilometers	2
LOGLENGTH	1.01	Log of radius of circumscribing circle in hundreds of kilometers	3
LOGPOP	3.94	Log of national populations in thousands, 1976 or latest	2
PCTCHANGE	2.17	Percent annual change in national population, 1960-70	2
LOGENERGY	2.73	Log of energy consumption per capita in kg coal equivalent, 1975	4
PCTURBAN	40.1	Percent of population in urban places (own country's estimate)	2
LOGPOP2	2.43	Log of population of second largest city in thousands	1
PORT	.548	Dummy: 1 if seaport	3
CAPITAL	.833	Dummy: 1 if capital of country	5
METRO	.579	Dummy: 1 if city data are for metropolitan area	1

SOURCES.—(1) Mainly *U.N. Demographic Yearbook, 1977* (New York: United Nations, 1978); at least two other sources were consulted to confirm relative populations of cities.

(2) UN Conference on Trade and Development, *Handbook of International Trade and Development Statistics* (New York: United Nations, 1979).

(3) *The Times Atlas of the World* (Boston: Houghton Mifflin Co., 1967).

(4) *United Nations Statistical Yearbook, 1976* (New York: United Nations, 1977).

(5) Arthur S. Banks, *Political Handbook of the World* (New York: McGraw-Hill Book Co., 1977).

face the question whether a collection of virtually all the countries of the world constitutes a sample in the statistical sense. On the one hand, the collection is the full population of nations (minus a few small countries and city-states) and therefore questions of statistical significance are not strictly meaningful. I choose, on the other hand, to regard each nation as the outcome of a long experiment that could have turned out quite differently given the contingencies of history. Consequently the usual tests of significance remain valid and useful in that their associated statistics may be regarded as weights that reflect the importance of each variable to the model. As a result of this argument, I shall refer to the 126 cases as a sample both because they are technically not all of the countries of the world, or all possible realizations of the national experiment, and because just as a set may be regarded as a subset of itself, so may a population be regarded as a sample.

Nevertheless, the question of spatial autocorrelation must be acknowledged. This problem, like its temporal counterpart, can lead to underestimation of the variance of the regression coefficients. This is a

TABLE 3
ZERO-ORDER CORRELATIONS AMONG VARIABLES IN THE MODEL

	LOGPOI	LOGAREA	LOGLENGTH	LOGPOP	PCTCHANGE	LOGENERGY	PCTURBAN	LOGPOP2	PORT	CAPITAL
LOGAREA	.39									
LOGLENGTH	.43	.96								
LOGPOP	.81	.54	.54							
PCTCHANGE	-.27	.12	.10	-.20						
LOGENERGY	.45	-.06	-.04	.10	-.51					
PCTURBAN	.41	-.15	-.11	-.01	-.43	.81				
LOGPOP2	.85	.39	.41	.81	-.25	.44	.36			
PORT	.11	-.05	-.01	-.05	-.01	.10	.19	-.00		
CAPITAL	-.25	-.19	-.20	-.30	.09	-.16	-.14	-.42	-.19	
METRO	.13	-.10	-.07	.01	-.04	.05	.16	.08	.19	-.21