The Evolution of City Size Distributions *

Xavier Gabaix
MIT and NBER

Yannis M. Ioannides
Tufts University

August 7, 2003

Abstract

We review the accumulated knowledge on city size distributions and determinants of urban growth. This topic is of interest because of a number of key stylized facts, including notably Zipf's law for cities (which states that the number of cities of size greater than S is proportional to $1/S$) and the importance of urban primacy. We first review the empirical evidence on the upper tail of city size distribution. We offer a novel discussion of the important econometric issues in the characterization of the distribution. We then discuss the theories that have been advanced to explain the approximate constancy of the distribution across very different economic and social systems, emphasizing both bare-bone statistical theories and more developed economic theories. We discuss the more recent work on the determinants of urban growth and, in particular, growth regressions, economic explanations of city size distributions other than Gibrat's law, consequences of major shocks (quasi natural experiments), and the dynamics of U.S. urban evolution.

Keywords: City size distribution, Gibrat's law, Hill estimator, persistence of city size distributions, power laws, random growth, urban growth, urban hierarchy, urban primacy, Zipf regression, Zipf's law.

JEL classification codes: C2, N9, R1.

Written for the *Handbook of Urban and Regional Economics, Volume IV: Cities and Geography*, J. Vernon Henderson and Jacques Francois Thisse, editors, North-Holland Publishing Company, Amsterdam.

*xgabaix@mit.edu, yannis.ioannides@tufts.edu. We thank Vernon Henderson, Tom Holmes, Henry Overman (our discussant), and Jacques Thisse for very useful comments. Gabaix gratefully acknowledges the hospitality of the Russell Sage Foundation in academic year 2002–3. Ioannides acknowledges generous support by the National Science Foundation and by the John D. and Catherine T. MacArthur, through the Research Network on Social Interactions and Economic Disparities.

Contents

1 Introduction

The evolution of city size distributions has attracted sustained by interest of researchers over a long period of time. The existence of very large cities, the very wide dispersion in city sizes, the remarkable stability of the hierarchy between cities over decades or even centuries, and the role of urbanization in economic development are all particularly interesting qualitative features of urban structure worldwide. Another surprising regularity, Zipf's law for cities, has itself attracted considerable interest by researchers. Therefore, it is tempting to see the urban evolution of different economies through the persistence of certain patterns in city size distribution worldwide. It is of special interest for a theory to predict Zipf's law and other empirically important features.

The chapter reviews the theoretical underpinnings and accumulated knowledge on the evolution of city sizes. It puts considerable emphasis on Zipf's law for cities. After it identifies the empirical record on Zipf's law, the chapter turns next to technical issues, which are associated with the sort of econometric evidence on which empirical Zipf's law currently rests, and to more ambitious empirical investigations of Zipf's law with city data. The chapter reviews the implications for city size distributions of major analytical developments in urban economics and related areas in economics over the last forty years or so and contrasts them with the empirical literature. The chapter also examines the general empirical evidence on the evolution of city size distributions in the U.S. as well as internationally.

The predictions of economic theory may be classified as emanating from roughly two major traditions. These traditions coexist and are not construed as being mutually exclusive; they are merely meant to serve as categories for organizing the literature. One is associated with urban economics, that is, in particular, systems of cities theories. A second is associated with economic geography or analytical geography, more generally. The urban structure reflects such important economic forces as increasing returns possibly at various levels (which produce centripetal forces), congestion (which produces centrifugal forces), trade (intracity, intercity and international) and non-market interactions, all of which play important roles in both of these traditions. Both of these traditions in the literature yield predictions about size distributions that are in some sense aspatial and emphasize in varying degrees differences across cities in terms of specialization. The chapter emphasizes theories and applications that examine the evolution of the city size distribution in a given economy as an outcome of forces that lead to

appearance of new and to decline of existing cities. This is also stressed by the most recent research on urban structure, as we see in more detail further below.

2 Zipf's Law for The Upper Tail of the City Size Distribution

We fix ideas and set notation by starting from the statistical regularity known as Zipf's law for cities. As early as Auerbach (1913), it was proposed that the city size distribution could be closely approximated by a power law distribution.[1]

2.1 Zipf's Law: Definitions

Let S_i denote the normalized size of city i, that is, the population of city i divided by the total urban population[2]. City sizes are said to satisfy *Zipf's law* if, for large sizes S, we have

$$P\left(\text{Size} > S\right) = \frac{a}{S^{\zeta}}, \tag{1}$$

where a is a positive constant and $\zeta = 1$. That is, the size of a city times the percentage of cities with larger size equals a constant.

An approximate way of stating Zipf's law, the so-called rank size rule, is a deterministic rule that follows from the definition: the second largest city is half the size of the largest, the third largest city a third the size of the largest, etc. That is, if we rank cities from largest (rank 1) to smallest (rank n), and denote their sizes $S_{(1)} \geq ... \geq S_{(n)}$, respectively, the rank i for a city of size $S_{(i)}$ is proportional to the proportion of cities greater than i. Therefore, by rewriting Equ. (1) we have: $S_{(i)} \simeq k/i$ for some constant k.

It is important to bear in mind that even if Zipf's law holds perfectly, the rank-size rule would hold only approximately.[3] Still, it is very useful to visualize Zipf's law.

[1] It is important to point out that there is no universally accepted definition of a *city* for statistical purposes. In the U.S. context, research has been conducted with both city proper data and data for metropolitan statistical areas. Differences in data availability worldwide may make international comparisons tricky. Rosen and Resnick (1980) show that the Pareto exponent of city size distributions tend to be closer to 1 when agglomerations are more carefully constructed, i.e. are closer to "true" agglomerations rather than administratively defined "cities". We return to this below.

[2] Talking about steady state distributions requires a normalization of this type.

[3] See Gabaix (1999, Proposition 4) for a precise statement of the rank-size rule, and the goodness of fit one can expects from it. The rank-size rule is a good approximation

Insert Figure 1 about here

To do this, we take a country (for instance the United States), and order its cities by population: New York has rank 1, Los Angeles has rank 2, etc. We then draw a graph, known as Zipf's plot: on the y-axis, we place the log of the rank (N.Y. has log rank $\ell n1$, L.A. log rank $\ell n2$); on the x-axis, the log of the population of the corresponding city (which will be called the "size" of the city). We take, like Paul Krugman [1996a, p.40], the 135 American metropolitan areas listed in the Statistical Abstract of the United States for 1991.[4] The result is something very close to a straight line. This is rather surprising, because there is no tautology causing the data to automatically generate a straight line. Furthermore, fitting a linear regression yields:

$$\ln \text{Rank} = 10.53 - 1.005 \ln \text{Size}, \tag{2}$$
$$(.010)$$

where the standard error is in parentheses, and the R^2 is 0.986. The slope of the regression line is very close to -1, and is measured with very high precision. Returning to levels yields a very close approximation to the rank-size rule. As we argue further below, power laws like the one reported in Equ. (2) fit empirical city size distributions quite well. Still, it is important to approach the issue rigorously in terms of econometric arguments.

Our approach in reviewing the literature on the evolution of city size distributions emphasizes conditions, theoretical or empirical, under which one may replicate with accuracy empirical regularities in city size distributions worldwide. We are interested in economic theories with behavioral foundations that predict such empirical regularities as Zipf's law, and others as well, but do not insist that the evidence on city size distributions be used to discriminate among those theories.

---

is for cities of high rank, but not for the largest cities. For instance, the rank-size rules says that ratio of the largest city to second largest city is 2. But Zipf's law implies that this ratio is widely variable, indeed has a smallest 95% confidence interval equal to $[1, 20]$. This comes from the Renyi theorem described in section 2.2.2, which says that $P\left(S_{(1)}/S_{(2)} > x\right) = 1/x$ for $x > 1$. So $[1, 1/.05]$ is the smallest 95% confidence interval for $S_{(1)}/S_{(2)}$.

[4] The Statistical Abstract of the U.S. lists all the agglomerations with size above 250,000 inhabitants. The exponent ζ is sensitive to the choice of the cutoff size above which one selects the cities. For a lower cutoff, the exponent ζ is typically lower. We come back to this issue, and a possible explanation, in section 2.4. The statistical literature (Embrechts et al. 1997) offers ways to discipline the selection of the cutoff, but those optimum cutoff techniques have not, to our knowledge, been used in the context of the city size distribution.

Some definitions are in order. A *power law* is a distribution function of the type $P(\text{Size} > S) \sim a/S^{\zeta}$ for large S. The positive number ζ is called the *power law exponent*. The literature sometimes uses the terms *Pareto law* (resp. *Pareto exponent*) instead of power law (resp. power law exponent). *Zipf's law* is the statement that $\zeta = 1$.[5] *Gibrat's law* states that the growth rate of an economic entity (firm, mutual fund, city, etc.) of size S has a distribution function with mean and variance that are independent of S.[6] Those conditions will be sometimes referred to respectively as Gibrat's law for means and Gibrat's law for variances.

2.2 Statistical Methods to Measure Power Law Exponents

We discuss next why a power law exponent ζ is notoriously difficult to estimate with city size and rank data. Embrechts *et al.* (1997) provides a very useful review of the different methods. We will present the two most commonly used, the Zipf regression and the Hill estimator. Both present pitfalls important to bear in mind.

2.2.1 The Zipf Regression Method and its Pitfalls

With n cities of ordered sizes $S_{(1)} \geq \ldots \geq S_{(n)}$, the Zipf regression fits an ordinary least squares (OLS) regression of the log rank i on the log size $S_{(i)}$ of the type (2):

$$\ell ni = A - \zeta_n \ell n S_{(i)}. \tag{3}$$

This procedure is the most commonly used in the empirical literature. One can show for large n, the coefficient ζ_n tends with probability 1 to the true ζ.

The advantage of this procedure is that it yields a visual goodness of fit with the power law. For large samples, such as with financial data, it is reasonably accurate. However, it has pitfalls in small samples. We provide next a Monte Carlo analysis of this phenomenon.

[5] This definition implies that the variance of S is infinite for $\zeta < 2$, and the mean is infinite for $\zeta < 1$. This is, strictly speaking, impossible, as the distribution of S is bounded above (by the total urban population in the case of absolute sizes, or 1 in the case of normalized sizes). So a more rigorous definition should be that the density is $p(S) = a'/S^{\zeta+1}$ for S over a range $[S_1, S_2]$ over which the power law applies, and $p(S)$ can be arbitrary elsewhere. Empirically, this range $[S_1, S_2]$ typically include the top 100 or so cities.

[6] It is sometimes used in the literature to mean that the distribution of growth rates of firms of size S is independent of S, not just the first and second moments.

We fix n, the number of cities, and draw n i.i.d. city sizes S_i from an exact power law with coefficient 1.[7] So Zipf's law holds perfectly in our Monte Carlo simulations. Take for instance a sample $n = 100$. We get a mean exponent $E[\zeta_{100}] = 0.94$, so that *the OLS procedure on average underestimates the value of* ζ, here by an amount 0.06. One can interpret the origin of the bias in the following way: the expected value of the ratio between $S_{(2)}$ and $S_{(1)}$ is 0.5, but the smallest 95% confidence interval for $S_{(1)}/S_{(2)}$ is $[1, 20]$ (see footnote 3). So typically, the value of $S_{(1)}$ will be above the value predicted by the linear regression with slope -1. In other words, the size of the largest city will look "too big". The best OLS fit will correct this by making the slope less steep, so that the best fit value of ζ_n will be less than the true value of ζ.

OLS regression reports an average standard error $\sigma^{\text{nominal}}(\zeta_{100}) = 0.013$, but the true standard error is $var(\zeta_{100})^{1/2} = 0.13$. Hence a 95% confidence interval for ζ_n is $[0.68, 1.20]$, when a naive view of OLS would one lead to expect $[.974, 1.026]$. This shows that the nominal standard errors reported in the OLS regression considerably *underestimates* the true standard error on the estimated coefficient. As a result, *taking the OLS estimates of the standard errors at face value will lead one to reject Zipf's law much too often.* For references, we report the results of Monte Carlo simulations in Table 1 below and the associated Zipf's law estimations for $n = 20, 50, 100, 200, 500$. The reason for those low nominal standard errors is that the ranking procedure creates positive correlations between the residuals, whereas the OLS standard error assumes that the errors are independent. So the total amount of error is understated by OLS. Actually, one can show that the true standard error is:

$$var(\zeta_n)^{1/2} \sim \zeta (2/n)^{1/2} \tag{4}$$

for large n.[8]

Value of the number n of cities in the sample	20	50	100	200	500
Mean ζ_n	0.90	0.92	0.94	0.96	0.98
Mean nominal OLS standard error on ζ_n	0.048	0.023	0.013	0.0078	0.0037
True standard error on ζ_n	0.28	0.18	0.13	0.098	0.063
Approximate true standard error on ζ_n : $\sqrt{2/n}$	0.31	0.20	0.14	0.100	0.063
True 95% confidence interval for ζ_n	$[0.37, 1.43]$	$[0.57, 1.27]$	$[0.68, 1.20]$	$[0.77, 1.15]$	$[0.85, 1.10]$

[7] Concretely, we draw n i.i.d. variables u_i uniformly distributed in $[0, 1]$, and construct the sizes as $S_i = 1/u_i$ and rank them.

[8] See Gabaix and Ioannides (2003) for the derivation.

Table 1. Statistics on the OLS coefficient ζ_n from regression (3), assuming that Zipf's law holds perfectly ($\zeta = 1$). The values come from 20,000 Monte Carlo simulations for each value of n. Under a general null of power law distribution with exponent ζ, the value for the statistics on ζ_n are those of the table multiplied by ζ. $\sqrt{2/n}$ is the asymptotic approximation of the true standard error on ζ_n, as discussed in the text.

Value of the number n of cities in the sample	20	50	100	200	500
Mean α_n	1.14	1.08	1.05	1.03	1.02
Mean nominal OLS standard error on α_n	0.065	0.029	0.016	0.0086	0.0039
True standard error on α_n	0.33	0.20	0.14	0.099	0.063
Approximate true standard error on α_n : $\sqrt{2/n}$	0.31	0.20	0.14	0.100	0.063
True 95% confidence interval for α_n	$[0.51, 1.76]$	$[0.69, 1.47]$	$[0.78, 1.33]$	$[0.84, 1.23]$	$[0.89, 1.14]$

Table 2. Statistics on the OLS coefficient α_n from regression (5), assuming that Zipf's law holds perfectly ($\alpha = 1/\zeta = 1$). The values come from 20,000 Monte Carlo simulations for each value of n. Under a general null of power law distribution with exponent ζ, the value for the statistics on α_n are those of the table multiplied by $1/\zeta$. $\sqrt{2/n}$ is the asymptotic approximation of the true standard error on α_n, as discussed in the text.

In Table 2, we replicate the study with the other OLS-based approach used in the literature, that is, regressions of log size on log rank:

$$\ell nS_{(i)} = A' - \alpha_n \ell ni. \tag{5}$$

For very large n, under the null of a power law with exponent ζ, α_n tends to $\alpha = 1/\zeta$. The results are similar to those of Table 1. The estimate of α_n is now biased upward. The origins of the upward bias are presumably the same as those of the downward bias of ζ_n, as $\alpha = 1/\zeta$. The true standard error on α_n are slightly higher than those on ζ_n. So if one chooses an OLS procedure, regression (3) is preferable to regression (5).

We conclude by discussing a pitfall associated with an augmented Zipf regression. That is, the literature reports regressions of log rank, ℓni, against log size, $\ell nS_{(i)}$, and its square:

$$\ell ni = a + b\ell nS_{(i)} + c\left(\ell nS_{(i)}\right)^2. \tag{6}$$

A coefficient c statistically different from 0 is interpreted as a departure from Zipf's law. It may be a statistical artifact, however. To show this, we perform Monte Carlo simulations like above with n cities drawn from Zipf's law. We run (6) and count the frequency at which the t-statistic on c is greater than 1.96 in absolute value, which would naively lead one to detect

a deviation from Zipf's law. For n =20, 50, 100, 200 and 500, one finds a statistically significant coefficient c respectively 65%, 78%, 85%, 90% and 93% of the time. Hence *in the OLS regression in equation (6) , one will often finds a statistically significant coefficient c, even if Zipf's law holds perfectly.* This fact has bearing on whether it is appropriate to reject Zipf's law on the strength of econometric evidence of a statistically significant quadratic term in OLS regressions of rank against the logarithm of size. We return to this in the discussion below of Black and Henderson (2002).

We conjecture that the reason why in regression (6) the coefficient c is typically found significant is the same as the reason why OLS has too low a nominal standard error on the ζ term in regression (3). That is, the positive correlations between residuals that are introduced by ranking cause the true amount of noise in the regression to be understated. All nominal standard errors are too low and, in particular, the coefficient c appears to be as significant from 0 too often.

To conclude, if one wants to rely on OLS to estimate ζ, the safest thing to do is to use a Monte-Carlo simulation with the sample size n in order to get the expected value of the bias and the true standard error of the estimator.One can also get the value of the bias by interpolation from our Tables 1 and 2, and using equation (4) take the value $\zeta\sqrt{2/n}$ as an estimate of the standard error[9].

2.2.2 The Hill (Maximum Likelihood) Estimator

An alternative procedure is the Hill estimator of ζ, the Pareto exponent in Equation (1) [Hill (1975)]. Under the null of perfect power law, it is the maximum likelihood estimator. For a sample of n cities with sizes $S_{(1)} \geq ... \geq S_{(n)}$, this estimator is:

$$\widehat{\zeta} = \frac{n-1}{\sum_{i=1}^{n-1} \ell n S_{(i)} - \ell n S_{(n)}}. \tag{7}$$

It inherits the efficiency properties of a maximum likelihood estimator.[10] An estimate of the standard error of $1/\widehat{\zeta}$ is constructed the following way. One calculates the "local slopes"

$$\tau_i = i\left(\ell n S_{(i)} - \ell n S_{(i+1)}\right),$$

[9] More tables are provided in Gabaix and Ioannides (2003).

[10] As we discuss below, Dobkins and Ioannides (2000) actually report estimates of the Zipf exponent obtained by means of this estimator.

for $i = 1, \ldots, n-1$. The Rényi representation theorem on ordered statistics (see, e.g., Reiss (1989), 36–37) shows that the τ_i are i.i.d. exponential variables with $P(\tau_i > \tau) = e^{-\zeta\tau}$ for $\tau \geq 0$. $\widehat{\zeta}^{-1}$ is just the empirical mean of the slopes τ_i, $\widehat{\zeta}^{-1} = \left(\sum_{i=1}^{n-1} \tau_i\right) / (n-1)$. A consistent standard error uses the standard deviation of the slopes:

$$\sigma_n\left(1/\widehat{\zeta}\right) = \left(\frac{\sum_{i=1}^{n-1}\left(\tau_i - 1/\widehat{\zeta}\right)^2}{n-2}\right)^{1/2} (n-1)^{-1/2}.$$

If $1/\widehat{\zeta} \gg \sigma_n\left(1/\widehat{\zeta}\right)$, the delta method gives the standard error on $\widehat{\zeta}$:

$$\sigma_n\left(\widehat{\zeta}\right) = \widehat{\zeta}^2 \left(\frac{\sum_{i=1}^{n-1}\left(\tau_i - 1/\widehat{\zeta}\right)^2}{n-2}\right)^{1/2} (n-1)^{-1/2}. \tag{8}$$

The properties of the Hill estimator in finite samples can be very worrisome. Embrechts *et al.* (1997, 330–345) discuss these in great detail. The core reason for the bad non-asymptotic properties of the Hill estimator is that the true distribution may have the expansion, for large S :

$$G(S) = P(\text{Size} > S) = \frac{a}{S^{\zeta}} + \frac{b}{S^{\zeta+\gamma}} + o\left(\frac{b}{S^{\zeta+\gamma}}\right), \tag{9}$$

and $\gamma > 0$. The terms $b/S^{\zeta+\gamma}$ introduces a bias that can be very high in small samples:[11]

$$E\left[\widehat{\zeta}\right] = \zeta + \frac{b\gamma}{a} E\left[\frac{1}{S_{(i)}^{\gamma}}\right].$$

[11] To convey the intuition for the result, we give the following heuristic derivation. Call $y_i = -\ell n G(S_i)$, where $G(x)$ is the true countercumulative distribution function written in (9). Then y_i is a standard exponential variable, and the Renyi theorem implies that $u_i = i\left(y_{(i)} - y_{(i+1)}\right)$ are i.i.d. standard exponentials. But

$$u_i \simeq i\frac{G'\left(S_{(i)}\right)}{G\left(S_{(i)}\right)}\left(S_{(i)} - S_{(i+1)}\right) = \frac{S_{(i)}G'\left(S_{(i)}\right)}{G\left(S_{(i)}\right)} i\frac{S_{(i)} - S_{(i+1)}}{S_{(i)}} \simeq \left(\zeta + \frac{b\gamma}{aS_{(i)}^{\gamma}}\right)\tau_i$$

so that

$$\widehat{\zeta} = E[\tau_i]^{-1} \simeq \zeta + \frac{b\gamma}{a} E\left[1/S_{(i)}^{\gamma}\right].$$

Hence the nominal standard error (8) of the Hill estimators can also considerably underestimate the true estimation error, as it overlooks the bias term.

A number of estimators have been proposed to address these issues, but many years of research have not yielded any simple, consensus solution to this problem. The state of the art may be the sophisticated nonlinear procedures advocated by Beirlant *et al.* (1999), Embrechts *et al.* (1997), and Feuerverger and Hall (1999). Those procedures often directly estimate the parameters ζ, b/a and γ in the expansion (9). This is still a domain of active research.

It would be interesting to have a thorough econometric study of this issue in order to assess how important the bias problem is.[12] With those caveat in mind, we propose one more methodological remark before proceeding to a review of the empirical results.

2.3 A Methodological Note: "Estimate, Don't Test"

Before evaluating the empirical evidence, it is useful to keep in mind an injunction of Leamer and Levinsohn (1995). They argue that in the context of empirical research in international trade, too much energy is spent to see if a theory fits exactly. Rather, researchers should aim at broad, though necessarily non-absolute, regularities. In other words, "estimate, don't test". *The main question of empirical work should be how well a theory fits, rather than whether or not it fits perfectly* (i.e., within the standard errors). With an infinitely large data set, one can reject any non-tautological theory. Consistently with this suggestion, some of the debate on Zipf's law should be cast in terms of how well, or poorly, it fits, rather than whether it can be rejected or not. For example, if the empirical research establishes that the data are typically well described by a power law with exponent $\zeta \in [0.8, 1.2]$, than this is a useful result: It prompts to seek theoretical explanations of why this should be true. Likewise, if further research establishes a degree of confidence for Gibrat's law, then theory should fit that, within the degree of confidence that the data offer.[13]

[12] This bias problem can be very important in financial data [Beirlant *et al.* (1999)], as indeed theories of the origins of power law behavior in financial data [Gabaix *et al.* (2003)] welcome the possibility of a bias term $b/S^{\zeta+\gamma}$.

[13] We wish to thank Henry Overman for suggesting this discussion.

2.4 Empirical Results on Cities

Before we proceed with reviewing empirical results, we wish to underscore an important data issue. That is, it matters whether one deals with urban agglomerations (i.e. metropolitan areas) or with city-proper data. Conceptually, the proper entity is the urban agglomeration as an urban economy, but often international data just give the city proper data. One would expect that the exponent ζ should be larger for city proper than the urban agglomeration data, in that urban agglomerations are not bound by legal definitions of cities-proper and therefore likely to have a longer upper tail. This point was made first by Rosen and Resnick (1980) and has been revisited recently by Brakman *et al.* (1999, 2001). The latter report comparisons, *ibid.,* pp. 206–208, 220–221, using international data.[14] With these differences notwithstanding and unless otherwise indicated, the terms urban and metropolitan are used as synonyms throughout the chapter.

Support for Zipf's law comes from numerous country studies and comparative international evidence. Rosen and Resnick (1980), Brakman *et al.* (2001) and Soo (2003) are the most complete empirical international comparative studies. These are typically conducted along the lines of Equ. (3). Rosen and Resnick examine city size distributions for 44 countries in 1970. The average Zipf's exponent is 1.13 with a standard deviation of 0.19, with almost all countries falling between 0.8 and 1.5. Brakman *et al.* (1999, 2001 pp. 206–208, 220–221) show that city-proper data are associated with higher Zipf exponents (mean=1.13, S.D.=0.19, $N = 42$) than urban agglomeration data (mean=1.05, S.D.=0.21, $N = 22$). Soo (2003) updates these results without altering the basic findings. He finds a Zipf coefficient of 1.105, for cities, but 0.854 for urban agglomerations.[15]

The estimated dispersion in the Zipf exponent is large. Some interpret this as mixed evidence for Zipf's law. We recall, however, that Table 1 above shows that large dispersion of exponents is to be expected under Zipf's law. Looking at the average of exponent estimates, however, we see that if the average value ζ is not exactly equal to 1, it is typically in the range $[0.85, 1.15]$. We conclude that power laws describe well the empirical regularity, with a Zipf exponent typically around 1. Furthermore, *predicting a value in a range say* $[0.8, 1.2]$ *may be included in the list of criteria used*

[14] The data are available at the United Nations web site http://unstats.un.org/unsd/citydata.

[15] Soo's non-parametric examination of the estimated Zipf coefficient across countries produces a distribution that is quite close to normal, with the variations being explained better by political economy variables than by economic geography variables.

to judge the success of urban theories.

Dobkins and Ioannides (2000) report OLS estimates of ζ, that are obtained along the lines of (3) with repeated cross sections of U.S. Census data for metro areas. Their estimates decline from 1.044, in 1900, to 0.949, in 1990. They also report maximum likelihood estimates for power law distributions, along the lines of (7) with the same data, which decline from .953, in 1900, to .553, in 1990. When they use the upper one-half of the sample only, a practice that conforms to some other estimations of Zipf's law (such as Fujita, Krugman and Venables (1999), Ch. 12), the estimate of ζ declines from 1.212, in 1900, with 56 metro areas in the entire sample, to .993, in 1990, with 167 metro areas in the sample. Gabaix (1999b) reports an estimate equal to 1.005, using the 135 largest metro areas in 1991 as reported in the *Statistical Abstract of the United States.*

Despite remarkable fits obtained for Zipf's law with U.S. city size data, problems remain. Nonparametric results by Dobkins and Ioannides (2000) and a finding of a significant quadratic term in a log rank regression (according to Equ. (6)) reported by Black and Henderson (2002), continue to raise genuine doubts about the validity of Zipf's law as a description of the entire distribution of city size for the US. We return to this issue further below when we review two very relevant recent papers. One is Duranton (2002), who compares simulation results of an interesting new model that utilizes quality ladders with the empirical distributions for U.S. and for France and explains departures from Zipf's law at both ends of the distribution. Two is Rossi-Hansberg and Wright (2003), who develop a system-of-cities inspired model that implies Zipf's law in special cases and also explains departures from Zipf's law at both ends of the distribution.

Black and Henderson (2002) examine the performance of Zipf's law with the twentieth century US city size distribution data. Their criticism of Zipf's law rests on a regression of the logarithm of city rank against the logarithm of size with metro area data. (Their data differ little from the Dobkins-Ioannides data.) Their results show that the Zipf coefficient declines from .861 in 1900 to .842 when all cities are used, and increases from 1.01 in 1900 to 1.18 in 1990, when only the top one-third of the size distribution is used. Their estimate of the coefficient c of the quadratic term in Equ. (6) is statistically significant. It would be useful to revisit those issues with the pitfalls described in section 2.2.1 in mind.

The approaches to estimation of Zipf's law that we discussed above are based on working with the steady-state size distribution of cities and therefore require some notion of stability of the underlying stochastic process. Difficulties with consistent definitions of cities over time, as when metropol-

itan area definitions in the U.S. change over time, make it hard to rely entirely on panel data. However, Black and Henderson (2002) and Dobkins and Ioannides (2000, 2001) do work with panel data. Ioannides and Overman (2003), on the other hand, constitutes the first attempt to use the Gibrat's law to test the validity of Zipf's law. We discuss their work further below.

We wish to draw the reader's attention to sources of information that have not been fully explored. Historians have produced fascinating series of urban populations that are reported in Bairoch (1988), Bairoch *et al.* (1988), Van der Woude *et al.* (1990) and De Vries (1984). The casual impression of the authors is that in some decades, large cities grow faster than small cities, but in other decades, small cities grow faster. This would suggest that Gibrat's law for means holds only as a long run average. But to our knowledge, no one has systematically used those data. They clearly deserve attention[16].

Insert Figure 2 about here

Finally, we wish to note that Zipf's law has been shown to hold for the bulk of the firm size distribution. Axtell (2001) and Okuyama *et al.* (1999) present evidence for the U.S. and Japan respectively. Our Figure 2 reproduces Axtell (2001). If the countercumulative density of the distribution is $G(x) = a/x^{\zeta}$, the density, its derivative, is $g(x) = a\zeta/x^{1+\zeta}$, so that a plot of log density vs log size will show an affine curve with slope $-(1+\zeta)$: $\ln g(x) = -(1+\zeta)\ln x + \text{constant}$. Axtell (2001) finds $\zeta = 1.059$ (S.E. 0.054) for the 5 million firms in the U.S. Census in 1997. Hence one can safely say that, except for very small and very large firms, U.S. firms follow Zipf's law. This is interesting because many of the conceptual issues that arise for cities arise also for firms. Most worked out theories of the firm would predict that many details should matter for the distribution. Fixed costs, and increasing or decreasing marginal costs, the type of competition, the cost and benefit of integration, should influence the size distribution of firms. This view of the world this way begs the question of why those details should be have the proper values that generate Zipf's law. However, random growth model offer a simple way to understand Zipf's law.

[16] There is also an interesting connection between Zipf's law and Christaller's Hierarchy principle. This principle states that if an industry is present in a certain city, it tends to be present in larger cities as well. Mori, Nishikimi and Smith (2003) show that this implies that is a negative correlation between the average size of the cities that host an industry, and the number of those cities. They call this the Number-Average Size rule. They provide empirical evidence this new, very interesting stylized fact.

Also, though random growth seems to suggest that, in the long run, firms and cities behave like constant return to scale economies, one does need a feature that is not constant return to scale to generate firms and cities in the first place – for instance a fixed cost, or an initial advantage. Perhaps this similarity of firms and cities will help guide some new theorizing. In any case, this strong support of Zipf's law for firms should increase one's posterior about the probability of Zipf's law for cities.

3 Random Growth and Zipf's Law

A first formal attempt to obtain power laws, and therefore, Zipf's law in particular, is Simon (1955). Simon assumes that urban population grows by discrete increments or "lumps." A new lump becomes a new city, with some probability; or, it goes to augment an existing city, with a probability that is proportional to the recipient city's population. Simon obtains a power law distribution as a limit of this process, but the model yields Zipf's law only as a special case. Dobkins and Ioannides (2001) confirm broad features of Simon's model, that is, that the probability of new cities appearing in the immediate vicinity of old cities and thus leading to large urban agglomerations, is increasing in the size of the existing city. Simon's model encounters some serious problems. In the limit where it can generate Zipf's law, it does not converge well, and requires that the number of cities grow indefinitely, in fact as fast as the urban population. Gabaix (1999b) and Krugman (1996) detail these problems.

3.1 From Gibrat's Law to Zipf's Law

We discuss next a variant of random growth theories that builds on Gibrat (1931).[17] The conclusion is that *if different cities grow randomly with the same expected growth rate and the same variance (Gibrat's Law for means and variances of growth rate), then the limit distribution of city sizes converges to Zipf's law.* We follow here the treatment of Gabaix (1999b), who also discusses the consequences of deviation from Gibrat's law.

Specifically, the distribution of city sizes will converge to $G(S)$, given by equation (1), if Gibrat's Law holds for city growth processes, that is, if city

[17] The first economic model with a power law may be Champernowne (1953). The classic mathematical treatment is Kesten (1973). Those random growth processes have enjoyed a renewed popularity in physics. Interesting analyses are include Levy and Solomon (1996), Marsili and Zhang (1997), Manrubia and Zanette (1997), Malcai, Biham and Solomon (1999), and Sornette (2001).

growth rates are identically distributed, independent of city size, and with a mean equal to the mean growth rate of the total urban population. It is straightforward to verify this claim. Let γ_t^i be the total growth of city i: $S_{t+1}^i = \gamma_{t+1}^i S_t^i$. If the growth rates γ_t^i are independently and identically distributed random variables with density function $f(\gamma)$, and given that the average normalized size[18] must stay constant and equal to 1, $\int_0^\infty \gamma f(\gamma) d\gamma = 1$, then the equation of motion of the distribution of growth rates expressed in terms of the countercumulative distribution function of S_t^i, $G_t(S)$, is

$$G_{t+1}(S) = \int_0^\infty G_t\left(\frac{S}{\gamma}\right) f(\gamma)\, d\gamma.$$

Its steady state distribution G, if it exists, satisfies

$$G(S) = \int_0^\infty G\left(\frac{S}{\gamma}\right) f(\gamma)\, d\gamma.$$

It is straightforward to verify that $G(S) = a/S$, where a is a constant, satisfies this equation. Gabaix (1999b) examines in further detail the precise conditions that generate Zipf's law.

3.2 Deviations from Gibrat's Law

3.2.1 Deviations that Affect the Distribution

Recognizing the possibility that Gibrat's Law might not hold exactly, Gabaix (1999b) also examines the case where cities grow randomly with expected growth rates and standard deviations that depend on their sizes. That is, the size of city i at time t varies according to:

$$\frac{dS_t}{S_t} = \mu(S_t)dt + \sigma(S_t)dB_t, \tag{10}$$

where $\mu(S)$ and $\sigma^2(S)$ denote, respectively, the instantaneous mean and variance of the growth rate of a size S city, and B_t is a standard Brownian motion. In this case, the limit distribution of city sizes will converge to a law with a *local* Zipf exponent, $\zeta(S) = -\frac{S}{p(S)}\frac{dp(S)}{dS} - 1$, where $p(S)$ denotes the

[18] One has $E[\gamma] = 1$ if all cities follow Gibrat's law. The more general condition for $E[\gamma] = 1$ is that cities in the relevant range have a growth rate that is independent of size, and that this growth rate is equal to the growth rate of the total urban population. Gabaix, Ramalho and Reuter (2003) elaborate this point in a more general context that allows birth and deaths.

stationary distribution of S. Working with the forward Kolmogorov equation associated with equation (10) yields:

$$\frac{\partial}{\partial t}p(S,t) = -\frac{\partial}{\partial S}\left(\mu(S)\ S\ p(S,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial S^2}\left(\sigma^2(S)\ S^2\ p(S,t)\right). \qquad (11)$$

The local Zipf exponent that is associated with the limit distribution is given by $\frac{\partial}{\partial t}p(S,t) = 0$, can be derived and is given by:

$$\zeta(S) = 1 - 2\frac{\mu(S)}{\sigma^2(S)} + \frac{S}{\sigma^2(S)}\frac{\partial\sigma^2(S)}{\partial S}, \qquad (12)$$

where $\mu(S)$ is relative to the overall mean for all city sizes.

Gabaix's theoretical contribution offers an opportunity for direct tests of the origin of Zipf's law in the form of Gibrat's Law for city growth rates. The empirical approach of Ioannides and Overman (2003) allows for a city's growth rate to depend on city size and to vary according to a law like equation (10) above. To do this, they non-parametrically estimate the mean and variance of city growth rates conditional on size. This allows them to test the validity of Gibrat's Law. It appears to be confirmed. We report the graphs in Figure 3. They then use equation (12) to directly estimate the local Zipf exponents. As we saw earlier, direct estimation of $\zeta(S)$ has turned out to be difficult to implement with standard parametric econometric procedures. However, non-parametric estimation lends itself readily to such a task. It is for this reason that Ioannides and Overman (2003) is arguably the strongest empirical confirmation to date of the validity of Zipf's law with U.S. data for metropolitan areas.

Insert Figure 3 about here

Their findings also help explain two interesting features of the size distribution of U.S. cities. First, as outlined above, estimates of the Zipf exponent for U.S. cities decline overtime. Gabaix (1999b) suggests that a possible explanation for this declining Zipf exponent is that towards the end of the period, more small cities enter, and that these small cities have a lower local Zipf exponent. The Ioannides and Overman estimations show that this suggestion is probably correct. Second, comparison of nonparametric estimates of the log rank – log size relationship to a standard parametric estimate suggests that the slope of the countercumulative function should increase absolutely and then decrease again at the upper end of the range of values, as Black and Henderson (2002) and Dobkins and Ioannides (2000)

document. The Ioannides and Overman finding of a local Zipf exponent that hovers between .8 and .9 for most of the range of values of city sizes and then rises and finally falls is consistent with this pattern. *They conclude that, at least for the upper tail of the distribution, the Gibrat assumption is indeed verified.* More work is called for to look at this issue.

We can offer a simple explanation for this flattening of the Zipf curve (lower exponent ζ) for small cities, which in effect means few small cities. It is conceivable that smaller cities have a higher variance than large cities. Variance would decrease with size for small cities, and then asymptote to a "variance floor" for large cities. This could be due to the fact that large cities still have a very undiversified industry base, as the examples of New York and Los Angeles would suggest. Using Equation (12) in the baseline case where all cities have the same growth rate, which forces $\mu(S) = 0$ for the normalized sizes, gives: $\zeta(S) = 1 + \partial \ln \sigma^2(S)/\partial \ell nS$, with $\partial \ln \sigma^2(S)/\partial \ell nS < 0$ in the domain where volatility decreases with size. So potentially, this might explain why the ζ coefficient is lower for smaller sizes.

3.2.2 Deviations from Gibrat's Law that do not Affect the Distribution

In this section we will see that the basic Gibrat process may be weakened considerably. First, the urban growth may accommodate a wide range of growth processes, as long as they contain a unit root with respect to the logarithm of city size: in particular, growth processes can have some mean-reverting component. Second, Zipf's law is compatible with the predictability present in the data (see section 6.1) as long as the determinants themselves are not ultimately correlated with size, and mean revert at long horizons.

To examine those facts analytically, we use discrete time notation and write:

$$\ell nS_{it} - \ell nS_{i,t-1} = \mu(\chi_{it}, t) + \varepsilon_{it}, \tag{13}$$

where χ_{it} is a possibly time-varying vector of characteristics of city i; $\mu(\chi_{it}, t)$ is the expectation of city i's growth rate as a function of economic conditions at time t; and ε_{it} is white noise. In the simplest Gibrat model, ε_{it} is independently and identically distributed over time and $\mu(\chi_{it}, t)$ is constant. We examine in turn the consequence of relaxing those assumptions.

Mean reversion versus unit root in the evolution process

First, we continue to assume a constant $\mu(\chi_{is}, t) = \mu$, but we examine the consequence of relaxation of the assumption of an i.i.d. ε_{it}. We suppose

a stochastic structure of the form

$$\varepsilon_{it} = b_{it} + \eta_{it} - \eta_{i,t-1},$$

where b_{it} is i.i.d., and η_{it} follows a stationary process. This gives

$$\ell nS_{it} - \ell nS_{i,0} = \mu t + \sum_{s=1}^{t} b_{is} + \eta_{it} - \eta_{i0}. \tag{14}$$

The $\sum_{s=1}^{t} b_{is}$ term in the above equation gives a unit root in the growth rate process, which is what ensures convergence to Zipf's law. The η_{it} term can have any stationarity (as long as the tails of $e^{\eta_{it}}$ are less fat than the Zipf distribution).[19] This means that, *for Zipf's law to hold, the city evolution process (14) can contain a mean reversion component, as long as it contains a non-zero unit root component.* Hence, in growth regressions the presence of a mean-reversion term is a priori compatible with Zipf's law — the crucial ingredient being the presence of a unit root term. Hence one can imagine that the next generation of city evolution empirics could draw from the sophisticated econometric literature on unit roots developed in the past two decades and surveyed by Stock (1994).

Economic predictability

We now examine the consequences of a non-constant $\mu(\chi_{it}, t)$ in (13). This is motivated by the empirical literature on urban growth (see section 6.1), which obtains a predicted value for the growth rate $\mu(\chi_{it}, t)$ as a function of a vector of characteristic χ_{it} of city i. In terms of the above section, this translates into:

$$\eta_{it} - \eta_{i,0} = \sum_{s=1}^{t} \left[\mu(\chi_{it}, t) - \overline{\mu}\right],$$

where $\overline{\mu}$ is the average growth rate. In view of the previous paragraph, Zipf's law requires that η_{it} be stationary. Let us unpack the economic meaning of

[19] We offer a heuristic derivation of this fact. Say that the process is $S_t = B_t H_t$, where $dB_t/B_t = b_t = \sigma dz_t$ is a Brownian motion with zero drift as in the simplest Gibrat process, and $H_t = e^{\eta_t}$ is an independent stationary process that follows a diffusion. S_t is reflected in the lower tail. One can write the forward Kolmogorov equation, and see that $p(B, H) = aB^{-2} f(H)$ is a solution of this equation if a is a constant and $f(H)$ is the steady state distribution of H. It is highly plausible, though we did not attempt to prove it, that this is the unique solution for large values of B. If H has power laws less fat than 1, i.e. if $E[H] < \infty$, then $P(S > x) = aE[H]/x$ and Zipf's law holds.

this condition. η_{it} is stationary if: (1.i) for a fixed χ, $\mu(\chi, t) - \overline{\mu}$ is "sufficiently" mean-reverting; or if: (1.ii), for a given city i, the χ_{it}'s are "sufficiently" mean-reverting. Case (1.i) says that the dependence on t indicates that some permanent characteristic can have impacts that are good in some time periods, bad in others. For instance the importance of temperature depends on the availability of heating systems or air conditioning. Proximity to iron ore deposits is a growth factor in some decades, and a decline factor in others. Case (1.ii) means that "good" characteristics are temporary. For instance, having better fiscal policies, or a more educated population, might be temporary, as policies and capacities change. If either (1.i) or (1.ii) hold, one sees how the growth regressions mentioned in section 6.1 can hold. If $\sum_{s=1}^{t} [\mu(\chi_{it}, t) - \overline{\mu}]$ is not stationary, then we have case (2): cities with the "right" characteristics will dominate, and the city size distribution will diverge. This divergence could be very slow. For example, suppose that city number 50 is endowed with a permanent advantage that make it grow at a rate higher than the rest of the urban population by a rate of 1% rate per year. It will need, in order of magnitude, $T = \ln(50)/0.01 \simeq 400$ years to overtake city number 1. Hence one needs an extremely persistent advantage to ensure this divergence of the distribution. It is somewhat unlikely that such advantages can persist without decaying or being imitated, with the help of directed technological or political change. This, and the evidence on Zipf's law, suggests that (2) is not possible, and rather that we must be in cases (1.i), (1.ii), or both. It would be extremely interesting for the empirical literature on urban growth to determine this, and to examine more precisely the mechanism by which (1.i) or (1.ii) happen, as one can conjecture they do.

3.3 Economic Models that Deliver Gibrat's Law

One could argue that a major challenge for urban theory is to deliver models that generate Gibrat's law, at least approximately. The dominant model of urban structure, that is the system of cities approach [Henderson (1974; 1988)] and the new economic geography [Fujita *et al.* (1999)] in their pure forms both fail the task of predicting a Zipf's law, and in fact not even a power law.[20]

Gabaix (1999b) offers a simple model of amenity shocks to cities, which cause intercity migration that in turn produce population shocks that are proportional to existing populations. When such amenity shocks are in-

[20] For the latter, see several prominent reviews of Fujita *et al.* (1999), such as Anas (2001); Davis (2002); Neary (2001).

dependent and identically distributed, the conditions of Gibrat's Law are satisfied. Gabaix (1999a) examines how extensions of such a model can be compatible with unbounded positive or negative externalities.

In a recent paper, Córdoba (2003) examines systematically the conditions for Zipf's law and concludes that "Gibrat's Law is not *an* explanation [...] but it is *the* explanation.". In other words, Gibrat's law is a necessary condition for Zipf's law. In Córdoba's model at equilibrium, cities are specialized and produce one good. Cities arise because of Marshallian externalities and there are no transport costs. Córdoba shows that for Zipf's law to arise, one first needs to have a balanced growth path. Remarkably, this is possible only if: (i) consumers have Cobb-Douglas preferences; or, (ii) Marshallian external effects have equal elasticities. If either tastes or productivity have power law distribution, one gets a power law distribution of city sizes. This power law distribution of tastes or productivity can itself come from a random growth process. The result is extended to the case of diversified cities with production of non-tradeables. One can expect that the analysis in Córdoba (2003) will motivate even more research on economic models compatible with Gibrat's law.

Rossi-Hansberg and Wright (2003) use ideas from the system-of-cities theory of Henderson (1974) and its urban growth application of Black and Henderson (1999) to develop a model where the urban structure eliminates local increasing returns to scale to yield constant returns to scale in the aggregate. This is accomplished by a model where local production takes place with a Cobb-Douglas production function and constant returns to scale in capital and labor services. Labor services are produced using raw labor and human capital, again with a Cobb-Douglas production function. Cities specialize completely in the production of different products. Total factor productivity affecting local production is produced in the style of endogenous growth [Lucas (1988), Romer (1990)] from total human capital and total labor in the city under Cobb-Douglas production function, is affected multiplicatively by a exogenous shock, and is external to each firm. Their specifications lead to a critical feature of the model, in that the optimal city size, that is the size that maximizes output net of commuting costs, implies that total commuting costs in each city are a constant fraction of total city output. This implies in turn that optimal city size is proportional to the square of the average product of labor. The model admits a balanced growth path along which growth is positive even if population growth is zero. Furthermore, along a balanced growth path, the growth rate of each city type may be written in terms of three components: one is proportional to the growth rate of human capital per person in each city type; a second

is proportional to the rate of growth of the total factor productivity shock in the industry; and a third is proportional to the excess of the contemporaneous total factor productivity over a weighted sum of past realizations of total factor productivities. So, faster growth of human capital leads to larger cities, while faster population growth leads to smaller cities.

Proposition 4, in Rossi-Hansberg and Wright, characterizes emergence of Zipf's law in exactly two restrictive cases. One case obtains if capital is not used in production and the growth rate of the total factor productivity shock is time-independent. In this case, productivity shocks are permanent and produce permanent increases in the level of the marginal product of labor making its growth rate scale-independent. A second case obtains if industry production is according to an AK model, where there is no human capital and production is linear in physical capital, all capital depreciates after production, there is no population growth and productivity shocks are temporary. In this case, productivity shocks have a permanent effect on the marginal product of labor through the accumulation of human capital. If neither of the above conditions are satisfied, Rossi-Hansberg and Wright show that the growth rate of cities exhibits reversion to the mean and that the standard deviation of city sizes increases with the standard deviation of industry shocks. That is, if a city is large, defined as having experienced a history of productivity shocks above average, it can be expected to grow slower than average in the future, and the opposite would be true for small cities. Therefore, there would be relatively few small cities and large cities are not large enough. Consequently, the departure of the log rank–log size from the straight line associated with Zipf's Law is as found in the U.S. data. This is, of course, great progress in the long-standing effort to provide plausible microfoundations for Zipf's law for cities by delivering good news for all sides. Zipf's law can be the outcome, albeit in a special case of a very important class of models, that is, those inspired by the system-of-cities approach augmented by adopting features of the endogenous growth theory.

3.4 Power Laws at Both End of the City Size Distribution: Random Growth with Exponential Compounding

Reed (2001; 2002) and Reed and Hughes (2002) advocate an interesting variant of the random growth process. This is obtained by a compounding geometric Brownian motion for city growth rates with the exponential distribution as follows. A Gibrat assumption of geometric Brownian motion but with constant instantaneous mean and variance, and given an initial state S_0, letting the process run for a fixed time T yields a size S_T that is

lognormally distributed. However, if time T is exponentially distributed — for instance, if the cities "die" at a Poisson rate δ – then a power law distribution is obtained at the upper tail, which is expected, but also a power law distribution in the lower tail. Thus, the outcome is a double Pareto, with different Pareto law exponents above and below the threshold, which is given by the initial state S_0. Reed (2002) offers some evidence that this is empirically relevant, as the bottom tail of the distribution has a distribution of the type $P(S < x) \sim x^{\gamma}$ for $\gamma > 0$. The hypothesis of a non-zero death rate is likely to be indeed relevant in the lower tail of the distribution. We expect future research to be stimulated by these contributions.

4 Economic Explanations for Zipf's Law Other than Gibrat's Law

In principle, the distribution of city sizes may satisfy Zipf's law even if city growth rates do not satisfy Gibrat's Law. One such possibility, suggested by Krugman (1996b), is that the presence of Zipf's law in features of physical geography that are relevant for the properties of the urban system that adapts to them, may cause city sizes to obey it as well. Other theories may predict stable distributions for city sizes as outcomes of deterministic or random growth processes which may also satisfy Zipf's law. For example, Henderson-style systems of cities theories are not incompatible with Zipf's law for cities, in that the actual cause of Zipf's law for city sizes may be found among the underlying determinants of city sizes [Henderson (1988)].

4.1 Zipf's Law for Cities Coming from a Power Law of Natural Advantages

Krugman (1996b) suggests that Zipf's law for cities might come from a Zipf's law of natural advantages. Indeed, he presents some evidence that the size of rivers follows Zipf's law. This might give rise to a power law of cities. A simple model to helps think about those issues. Call A a index of natural advantages of a city — for instance,its proximity to the coast or the size of the river near it. Consider that the output of a city i with amenities A_i is $F(A_i, K_i, S_i)$, with F exhibiting constant returns to scale with respect to all of its arguments: K_i is the amount of capital and S_i the amount of labor in city i. Consider a model without randomness. Equalization of marginal products across cities gives $F_K(A_i, K_i, S_i) = r$ and $F_L(A_i, K_i, S_i) = w$, where r and w are the rental prices of capital

and labor respectively. The constant returns to scale assumption gives $F_K(A_i, K_i, S_i) = F_K(1, S_i/A_i, S_i/A_i)$, so that the solutions are of the type $K_i = kS_i$ and $L_i = lS_i$. The population of city i is proportional to the natural advantages of the city.

If the distribution of natural advantages across cities is power law with exponent ζ_A, (i.e. there is a b such that $P(A_i > A) = bA^{-\zeta_A}$ for A large) we get:

$$P(S_i > S) = P(A_i > S/l) = b = b(S/l)^{-\zeta_A} \sim S^{-\zeta_A}$$

so that the population distribution is power law with exponent ζ_A: $\zeta_S = \zeta_A$. Hence, if we have evidence that $\zeta_A = 1$, we would have an explanation for Zipf's law.

Obviously, more research is needed to assess this hypothesis. One of the difficulties is that the link between say the river flow f and the corresponding economic amenity A that would enter in the productivity function is not obvious. Should we have $A = f$, or $A = f^\beta$ for some $\beta \neq 1$? This matters, as one can show that if f is power law distribution with exponent ζ_f, relation $A = f^\beta$ will yield $\zeta_S = \zeta_A = \zeta_f/\beta$. Hence one has to give a reason why $\beta = 1$.

As an explanation for the persistence of city size distributions, the transmission of power laws of physical geography to city sizes needs to be refined further. For example, we know from Fujita and Mori (1997) that discontinuities in the physical landscape, such as natural ports and waterways, have important consequences for the location of cities. Still, it is clearly not the case that their actual dimensions may affect the size of cities in all instances. For example, the size of a navigable river is pertinent, but coastal location does not lend itself to such measurement. Furthermore, such theories are problematic as theories of growth when a particular physical amenity is held fixed. Clearly, this issue needs to be addressed further by the literature.

4.2 Zipf's Law for Cities in Models of Self Organization and Endogenous City Formation

We review next a number of recent papers that develop models that combine several theoretical ideas. All of these papers use simulations to test their theoretical predictions.

Axtell and Florida (2001) offer a hybrid theoretical model of an urban system that predicts Zipf's law at its steady state. They attempt to "reconcile the tension between centripetal and centrifugal forces that we believe determines city sizes at the micro level, and the as-if-constant returns dynamics that seem to apply at the macro level" [*ibid.*].

Axtell and Florida propose a model of firm formation which leads to city formation by the location decisions of firms. Individual agents are myopic and interact in team production. Total team output is increasing and convex in team effort, and agents receive compensation equal to the equal shares of output. Individuals' choice of income versus leisure imply that a Nash equilibrium in effort levels exist but is Pareto-dominated by higher effort levels which are not individually rational. There exists a maximum stable size for firms beyond which groups are dynamically unstable. This implies that for firms beyond a certain size random perturbations lead to unstable adjustments. Each agent's location is originally random. Agents are allowed to move among firms or to start their own firms. When an agent starts a new firm, she selects a new location from among a finite number of locations, with a small probability, and it stays put, with a large probability. The authors claim that firms' growth rates are Laplace-distributed, their variance decreases with firm size according to a power law, wages are increasing in firm size, constant returns to scale prevail at the aggregate and city sizes obey Zipf's law.

Duranton (2002) is one of the most interesting economic models of city growth and aims at matching the observed distribution of city sizes. It actually does so very well: it offers a fairly good fit (with several free parameters), with approximate power laws for both the upper and lower tails [*c.f.* Reed (2002)]. Several ingredients of Duranton's model are familiar to students of new economic geography. It uses the quality-ladder model of growth developed by Grossman and Helpman (1991) in an urban framework. Cities grow, or decline, as they win, or lose, industries following new innovations. So small innovation-driven technological shocks are the main engine behind the growth and decline of cities. The paper shows that observed regularities about the city size distribution are compatible with the basic building blocks of urban economics, like the existence of agglomeration economies, crowding costs, etc. In particular, these building blocks are crucial for the theory's good simulation performance.

Duranton's model has the virtue of offering a plausible explanation of the mobility of cities through the size distribution, and of generating a non-trivial such distribution from economic decisions of firms. The model does not match Gibrat's law: the mean and variance of growth rates decreases with size.[21] This is due to the fact that it does not model that larger

[21] At the time of the completion of this chapter, the quantitative predictions of the model for the mean and variance of growth rates as a function of size were not made explicit, so that it was unclear how close or far they are from empirical processes.

cities have very diversified industrial base, which is an intuitive reason why Gibrat's law for variance may hold.

The model does match both the U.S. and French city size distributions when key parameters are calibrated based on appropriately different fundamentals.[22] Duranton's simulations show that the 10th and the 90th percentiles, that are predicted by the model, bound the U.S. distribution above the size of 220,000 inhabitants and the French distribution in its entirety. Still, in spite of his success in matching both the U.S. and French data, Duranton underscores that the real test should be whether proposed theories work well in explaining sources of urban growth and decline.

While Krugman (1996a) argues that the basic features of the urban system ought to be studied in models of self-organization, it was only until relatively recently that these newer theories were actually utilized to study empirically testable aspects of the urban system. The two most important contributions along such an approach is Brakman, Garretsen, Van Marrewijk and van den Berg (1999) and Brakman, Garretsen, and Van Marrewijk (2001). The new economic geography models of cities that they develop provide, in particular, for congestion costs via the specification of labor requirements for the production of the intermediate goods produced in different cities. Their simulations yield outcomes that resemble Zipf's Law. However a Zipf coefficient near 1 is obtained only for certain parameter values which they associate with what they refer to as "industrialization", that is, large decrease in transportation costs and increasing importance of footloose industry with increasing returns to scale. Their pre- and post-industrialization scenaria are associated with Zipf coefficients exceeding 1.

5 Dynamics of the Evolution of City Size Distributions

Eaton and Eckstein (1997) is, arguably, the most noteworthy recent study that focused on the persistence of the city size distribution and one of the most important contributions to the recent urban growth literature. The paper starts with a comparison of the dynamic evolution of the city size distribution between France and Japan. These countries have maintained national borders (that is when colonial possessions are ignored) that have remained unaltered during recent history and have urban systems with the number of cities remaining roughly constant. Eaton and Eckstein emphasize

[22] This is important because the US data imply a roughly concave Zipf's curve and the French data a roughly convex one.

the observed persistence of the distribution over time, which they refer to as *parallel* growth of French and Japanese cities during 1876–1990 and 1925–1985, respectively. They confirm this finding by means of several alternative empirical techniques, such as Lorenz curves, Zipf regressions (logarithm of rank against logarithm of size) and non-parametric transition matrices of evolving size distributions.

They propose a theory that explains these facts and combines features of Henderson (1974) and of Lucas (1988). Their model allows for urban congestion but not intercity transportation. Persistence in the relative city size distribution is ensured by assuming that returns to learning in each city is proportional to the weighted average of human capital stocks in all cities, where the interaction coefficients are constrained to be consistent with steady-state growth. If city populations are growing at the same rate, than so are wages and consumption. Eaton and Eckstein also allow for intercity migration and examine conditions for utility costs of migration such that relative populations remove incentives for individuals to migrate. These conditions take the form of lower and upper bounds on the relative populations of two successive ranks. These bounds converge to the same quantity, if the effective rate of time discounting is equal to 0, for each city depend upon the ratio of human capitals in the respective cities and economy-wide parameters. However, while their result *does explain* the existence of *invariant* city size distributions, it does *not explain* why this distribution should obey a Zipf law, or even a power law. Nonetheless, the model is sufficiently flexible to let them set parameters that fit the data quite well. This result of parallel growth is also associated with parallel growth in total factor productivities across cities.

As the above discussion makes clear, the results of Eaton and Eckstein (1997) depend critically on conditions that bound intercity migration. Some of the earlier literature on city size distributions, such as Suarez-Villa (1988) and Tabuchi (1986), also emphasize the relationship between ad hoc laws governing intercity population flows and the stability of the city size distributions. This is, of course, not surprising. It thus appears that additional progress would be made if general models of intercity migration and trade would be built.

In the context of Eaton and Eckstein's approach, the reader naturally wonders what would happen to city size distributions in an economy marked by *expansion of its land mass and emergence of new cities.* Dobkins and Ioannides (2000) were the first to address this question recently with respect to the urban system. We next turn to questions of spatial evolution by posing them in the context of recent research on the spatial distribution of

economic activity in the U.S.

5.1 Spatial Concentration of Economic Activity in the U.S.

Before we go into details of this literature, it would be interesting to provide a broader historical perspective on the spatial concentration of economic activity in the U.S.. Recent research has also examined the spatial distribution of population at different levels of aggregation. Beeson, DeJong and Troesken (1999) and Beeson and DeJong (2002) examine regional patterns of population growth at the state and county level from 1790 to 1990. They find that state-level populations show convergence while county level populations show divergence. While initial tendencies towards convergence lasted roughly through 1800s, in the post World War II period county-level populations have diverged. Their analysis points to the *importance of transitional dynamics* as opposed to *steady state dynamics.* When territories opened up for settlement, growth rates were very high relative to steady states. Once such "frontier effects" have been controlled for, the tendency to divergence in the post war period is clear.

The United States transformed itself from a rural to an urban society over the last three centuries. Kim (2000) emphasizes that after a century of unremarkable growth, the 1700s, the pace of urbanization rose to historically unprecedented levels between the nineteenth and early twentieth centuries. In the twentieth century, the urban population continued to increase but in a much more dispersed manner as the suburban population increased. Throughout these developments, cities also exhibited considerable variation in their population sizes. Kim emphasizes the role of changes in regional comparative advantage and in economies of scale in transportation and local public goods for the patterns of U.S. urban development. He finds that differences in urban sizes are associated with the role of reduced market transaction costs in coordinating greater geographic division of labor. Kim (2002) looks at the dynamic evolution of urban densities. The paper documents the historical changes in population and employment densities in U.S. cities and metropolitan areas, and explores the causes of their rise and decline between the late nineteenth and the twentieth centuries.

The role of urban density has recently attracted attention in relation to the evolution of *other measures of urban size,* such as *employment.* In particular, papers by Carlino and Chatterjee (2001, 2002) point to a pronounced trend towards deconcentration of employment in the U.S. since WWII. That is, the employment share of relatively dense MSAs has declined and the share of less dense MSAs has risen. Similarly, they show that such

effects also apply within MSAs. They explain these trends by means of a density-dependent congestion costs. They do not, however, estimate models for the pattern of transition. Still, these works challenge the view, based on population size studies, that the urban landscape is in some sort of steady state. Instead, they find considerable change.

5.2 Urban Evolution in the U.S.

Dobkins and Ioannides (2000) develop a data set that tracks U.S. cities, actually metropolitan areas, from 1900 to 1990. They use contemporaneous definitions of metropolitan areas, described in detail in the Data section in *ibid.* The number of cities grows from 112 in 1900 to 334 in 1990. Many of the cities that enter the data grow from settlements physically in existence for many years, prior to the time they pass the appropriate threshold of population, that is 50000 inhabitants. Entirely new cities also come into being, the latest one in 1944, and quickly grow large enough to be included in the data. Dobkins and Ioannides find that the U.S. urban system is characterized by *parallel growth, despite its spatial expansion.* They analyze the data in more detail, by constructing transition matrices, and track the movement of each city in the distribution relative to the others.

As noted above, Eaton and Eckstein's selection of France and Japan was motivated by their roughly stable geographical boundaries and the consistent availability of data. In contrast to such "old" countries as France and Japan, the United States has grown by continuously expanding its land mass into a well defined hinterland. New regions and cities have been brought into the U.S. urban system during the nineteenth and twentieth centuries, older regions have grown and declined, and the spatial distribution of economic activity has undergone some remarkable changes. In Europe, almost no new cities were created during the twentieth century. The U.S. urban system has developed with initial conditions quite different from those of other countries.

As Quah (1993) has forcefully argued, typical cross-section or panel data techniques do not allow inference about patterns in the intertemporal evolution of the entire cross-section distribution. They do not allow us to consider the impact over time of one part of the distribution upon another, i.e., of the development of large cities as a group upon smaller cities. Making such inferences requires one to model directly the full dynamics of the entire distribution of cities. The evolution of urbanization and suburbanization may affect individual cities so drastically as to render conventional methods of accounting for attrition totally inappropriate. As smaller urban units fuse

to create larger ones, and given the small number of time series observations, non-parametric or semi-parametric distributional approaches such as the one proposed here would be the only appropriate ones. In fact, these techniques are appropriate when the sample of interest is the entire distribution, and individual observations are used to recover information about the entire distribution. The availability of data are severely restricted both in the time and the cross-section dimensions: there are only ten cross-sections, one for each of the ten census years since 1900, with 112 metropolitan areas and 334 in 1990.

The paucity of the data naturally lends itself to techniques used by Quah (1993) and Eaton and Eckstein (1997). That is, one may construct from population data a fairly low-dimensional vector indicating the frequency of cities in each of a number of suitably defined intervals (cells). Let f_t denote the frequency (density) distribution of P_{it} at time t. Eaton and Eckstein assume that f_t evolves according to a first-order autoregression (that applies to the entire distribution function (rather than scalars or vectors of numbers):

$$f_{t+1} = M \cdot f_t, \tag{15}$$

where M is a matrix of parameters. If f_t were restricted to be measures defined over a discrete set, then M in (15) is a Markov transition matrix. Absence of a random disturbance allows us to iterate (15) forward to get:$f_{t+s} = (M \cdot M \cdot \ldots \cdot M) \cdot f_t = M^s \cdot f_t$. Divergent, convergent or parallel growth may be ascertained by the properties of $f_\infty \equiv \lim_{t\to\infty} : f_t$. If a limit distribution f_∞ exists, then according to the Perron-Frobenius theorem it is given by the eigenvector corresponding to the unique unitary eigenvalue of M, the nonzero solution of $[M - I]f_\infty = 0]$, where 0] denotes a column vector of zeroes. Parallel growth is understood to occur if f_∞ tends to a limit with non-zero probability over the entire support. Convergent growth would occur if f_∞ is a mass point, and divergent growth if f_∞ is a polarized or segmented distribution.

Dobkins and Ioannides (2000) and Black and Henderson (2002) adapt Equ. (15) in order to allow for new cities to enter according to a frequency distribution ε_t. If the number of entrants between t and $t+1$ is I_t^n, $I_{t+1} = I_t + I_t^n$, then

$$f_{t+1} \;=\; \frac{I_t}{I_{t+1}}\, M_t\, f_t \;+\; \frac{I_t^n}{I_{t+1}}\, \varepsilon_t. \tag{16}$$

If M_t and $\iota_t \equiv \frac{I_t^n}{I_{t+1}}$ are time-invariant, then the above equation is amenable to the standard treatment. Letting M and ι be the respective time-invariant

values, we may iterate Equ. (16) backwards to get: $f_t = (1-\iota)^t M^t f_0 + \sum_{\tau=0}^{t}[(1-\iota)M]^{t-\tau}\iota\varepsilon_\tau$, where f_0 denotes the initial distribution of city sizes.

A steady state solution of (16) characterizes the distribution of city sizes in the long run with entry. In general, if there are few or no entrants, $\iota \approx 0$, the homogeneous solution dominates: the invariant (ergodic) distribution is a useful measure of the state of the urban system in the long run. If, on the other hand, ι is non-negligible, then the particular solution may *not* be ignored. In fact, in that case, the magnitude of the largest eigenvalue of $(1-\iota)M$ is $(1-\iota)$, and the impact of the initial conditions would be less important the higher is ι, the number of new cities that have entered over the last decade as a proportion of the new total number of cities.

In the Dobkins and Ioannides data, the values of ι_t are as follows: $\iota_{1910} = .194$, $\iota_{1920} = .067$; $\iota_{1930} = .051$, $\iota_{1940} = .019$, $\iota_{1950} = .012$, $\iota_{1960} = .229$, $\iota_{1970} = .136$, $\iota_{1980} = .245$, and $\iota_{1990} = .036$. These numbers suggest possibly a non-stationary series and the intertemporal variations in ι_t are interesting and worthy of special analysis. We note that in the absence of a theory of entry of new cities, there is rather limited scope for a purely statistical analysis based on such a small number of time series observations. Entry of new cities is pursued further by Dobkins and Ioannides (2001).

The stochastic specification of Equ. (16) is, in general, very complicated, especially when M_t may be time-varying. E.g., forces that cause urban growth and decline may operate quite differently at the upper level of the distribution than at the lower one, and their pattern may change over time. The distribution of new entrants has most of its mass at the lower end, which to large extent reflects the nature of our data. Even if M_t is not time-varying, it could be associated with an invariant distribution that could reflect very different properties.

By coding the position of each city relative to the others within the distribution, we are able to see whether or not specific cities move up or down in the distribution over time. Dobkins and Ioannides constructed transition matrices which are reported in *ibid.,* Appendix A.[23] The empirical transition matrices that are reported suggest that concentration at the upper end of the distribution becomes more pronounced over time: the diagonal entries are higher for higher percentiles. Another observation that follows is that most movements are to nearby cells, with very few big jumps. As one might expect in the U.S. data, there is somewhat more movement off the diagonal (compared to the French and Japanese data). Most of that

[23] De Vries (1984), Ch. 7, appears to have originated the study of urbanization by means of transition matrices.

movement is toward greater concentration in the time period from 1900 to 1990. However, these transition matrices have limitations. They do not pick up the full effect of "entering" cities and they do not offer us any more insight into why such changes might occur. There are undoubtedly other variables that might impact on city size distribution.

Black and Henderson (2002) confirm these results by working with a slightly different data set and a somewhat more general model. Specifically, they work with the steady state solution of Equ. (16) which does account for entry. They also interpret the increasing concentration at the upper end of the distribution as being due to scale economies and changes in technology. Since the mean city size increased four-fold and the median five-fold, medium-size cities have grown substantially. They attribute this growth less to the impact of technology through local knowledge accumulation and improved commuting and more the effect of changes in the national demand for the output of inter-city traded services, which favors large cities. They test for the stationarity of the transition matrices, which is never close to being rejected. They also examine mobility by means of first passage times and find that upward mobility is much stronger than downward mobility. They interpret slow downward mobility as an effect of "established urban scale."

Ioannides and Overman (2000) consider, in the light of recent theoretical advances, the spatial characteristics of the U.S. urban system as it evolved over the twentieth century. These advances have highlighted the importance of spatial dimensions in understanding the evolution of urban systems: Fujita, Krugman and Venables (1999) have added important new spatial insights to the established literature on systems of cities [Henderson (1974; 1988)]. The system of cities approach features powerful models of intrametropolitan spatial structure, but neglected intermetropolitan spatial structure. Intermetropolitan spatial structure plays a key role in the new economic geography literature [Krugman (1991); Fujita *et al.* (1999)]. Further, as shown by Fujita and Thisse (2002), the importance of spatial dimensions is not just restricted to the new economic geography. Rather, it is a general feature of recent theoretical advances in our understanding of the economics of agglomeration.

This recent theorizing has formalized thinking about two fundamental features of any given location – the *first* and *second* "natures" – that determine the extent of development at that location. First nature features are those that are *intrinsic to the physical site itself,* independent of any development that may previously have occurred there. For example, locations on navigable rivers, with favorable climates have first nature features

that might encourage development. The second nature features of a location are those that are *dependent on the spatial interactions between economic agents.*

However, these theories do not offer very precise predictions, and especially of the type that may be used to structure empirical investigations. Real life geography, the tendency for all cities to grow, the gradual convergence to some kind of equilibrium in the westward expansion of the country, the movement of population towards the sunbelt and changes in the U.S. urban system induced by a shift over the period in industrial structure away from manufacturing and towards services are all important features in the spatial evolution of the US urban system that have not yet been elaborated in the formal theory. Thus, Ioannides and Overman seek to understand first and second nature features of the U.S. urban system without restricting analysis to specific functional forms. Instead, they choose to focus predominantly on non-parametric methods proposed by Quah, *op. cit.*, that is, non-parametric estimations of stochastic kernels for the distributions of city sizes and growth rates, conditional on various measures of market potential. They show that while these relationships evolve during the twentieth century, by 1990 they stabilize so that the size distribution of cities conditional on a range of spatial variables are all roughly independent of these conditioning variables. In contrast, similar results suggest that there is a spatial element to the city wage distribution.

Their parametric estimations for growth rates against market potential, entry of neighbors, and own lagged population imply a negative effect of market potential on growth rates, unless own lagged population is also included, in which case market potential has a positive effect and own lagged population a negative one. Cities grow faster when they are small relative to their market potential.

Overman and Ioannides (2001) report non-parametrically estimated stochastic transition kernels for the evolution of the distribution of U.S. metropolitan area populations, for the period 1900 to 1990. These suggest a fair amount of uniformity in the patterns of mobility during the study period. The distribution of city sizes is predominantly characterized by persistence. Additional kernel estimates do not reveal any stark differences in intra-region mobility patterns. They characterize the nature of intra-size distribution dynamics by means of measures that do not require discretization of the city size distribution. They employ these measures to study the degree of mobility within the U.S. city size distribution and, separately, within regional and urban subsystems. They find that *different regions* show *different degrees of intra-distribution mobility. Second-tier cities show more mobility than*

top-tier cities.

The results of Dobkins and Ioannides (2001) may also be considered as supportive of parallel growth. They test implications of economic geography by exploring spatial interactions among U.S. cities. They augment the data set developed in Dobkins and Ioannides (2000) by means of spatial measures including distance from the nearest larger city in a higher- tier, adjacency, and location within U.S. regions. They also date cities from their time of settlement. They find that among cities which enter the system, larger cities are more likely to locate near other cities. Moreover, older cities are more likely to have neighbors. Distance from the nearest higher-tier city is not always a significant determinant of size and growth. They find no evidence of persistent nonlinear effects on urban growth of either size or distance, although distance is important for city size for some years.

6 The Empirical Evidence on the Determinants of Urban Growth

6.1 Determinants of Urban Growth

Madden (1956) provides an interesting non-parametric analysis of urban growth in the United States. He emphasizes stability features in the distribution of growth rates and their evolution over time, where he notes that great dispersion coexists with considerable intertemporal variation for individual cities.

Henderson (1988), Glaeser, Kallal, Scheinkman and Shleifer (1992) and Glaeser, Scheinkman, and Shleifer (1995) examine the role of socioeconomic characteristics of city populations and of city industrial structures in economic growth. The results are detailed in this Handbook by Moretti (2004).

Black and Henderson (2002) also estimate an equation for Gibrat's Law, that is for the growth rate as a function of lagged size, which yields a *statistically significant estimate for the mean reversion coefficient,* the coefficient of the logarithm of size, from $-.022$ to $-.039$. However, the finding of significant mean reversion may be an artifact of *measurement error.* That is, measurement error of 10% along with a standard deviation of the logarithm of size of .7, for the fifty largest cities, would imply an estimate coefficient of .02. Also, positive autocorrelation in the residual of the regression could also show up as mean reversion. In fact, the studies of Davis and Weinstein (2001) and Brakman *et al.* (2002), discussed below, do estimate generally positive autocorrelation for the error in such a regression.

Black and Henderson also report regressions with additional explanatory variables, that is spatially varying geographical variables like temperature, precipitation, proximity to coast (including proximity to the Great Lakes, regional dummies, and market potential variables (which are defined in an ad hoc fashion). They find that cities in warmer, drier and coastal locations do grow faster, and that regional dummies have little additional impact. Market potential has a quadratic effect on growth that diminishes as market potential rises, but it has a large effect around its mean value. Having neighbors nearby enhances growth, an effect of intercity trade. They interpret the diminution of the effect for large market potential as an outcome of *competition.* If a city is in a very high market potential area it suffers from competition: *Los Angeles benefits by being far from New York.* Still, high market potential helps large cities maintain their relative positions.

Black and Henderson are particularly careful with the estimation of the relative growth equation. Noting that the lagged city size and spatial interactions introduce endogeneity, they use lagged instruments with GMM for the unbalanced panel data for the estimation. Allowing for fixed effects and using GMM increase the absolute value of the mean reversion coefficient nearly ten-fold. This implies mean reversion that is so much stronger than is typically found in the growth literature, that it raises doubts for the reliability of these estimates. Black and Henderson also examine city sizes in relation to city types, defined in terms of industrial compositions, and find that different city types have different absolute sizes. Therefore, changes in industrial compositions change relative sizes. These results confirm important features of the system-of-cities approach.

Finally, Florida (2002) studies the impact of hard-to-measure variables such as the openness to new ideas and creativity. He uses measures and proxies, such as the fraction of the population who is foreign born or gay, a coolness and a bohemian index, all of which are not commonly used, and finds that they have a high predictive power.

We expect more studies such as those to arise, especially in a non-U.S. contexts. A tighter link with the evidence on the stability of the city size distribution, such as along the lines of the distinction in section 3.2.2 deserves serious empirical attention.

6.2 The Determinants of Urban Primacy

Rosen and Resnick (1980) and Wheaton and Shishido (1981) show that urban concentration is negatively correlated with a country's population. Ades and Glaeser (1995) offer an empirical analysis that shows that high tariffs,

high costs of internal trade and low level of international trade increase the degree of urban concentration. Interestingly, a very good predictor is a political variable: dictatorships have central cities that are, on average, 50 percent larger than their democratic counterpart. Their evidence suggests that the causation goes from political factors to urban concentration rather than the opposite.

6.3 Studies of Urban Growth Based on Quasi "Natural Experiments"

Davis and Weinstein (2002) and Brakman *et al.* (2002) offer a completely different viewpoint on the robustness of city size distributions when they are subject to unusually large aggregate shocks. These papers rely on the quasi "natural experiments" provided by the strategic bombing of, respectively, Japan and Germany, during World War II. The two studies differ, however, in their time horizons. In the former paper, the case of Japan, the time span ranges over the past 8,000 years. In the latter paper, the case of Germany, in the latter, it ranges from the beginning to the end of the 20th century.

These studies examine the performance of three, possibly not mutually exclusive, theories of economic geography and urban development. These are: first, *increasing returns,* defined as the combined effects on city size of knowledge spillovers, labor market pooling and costly intercity transportation either as modelled by the system of cities literature [Henderson (1974)] or by new economic geography [Krugman (1991)]; second, *random growth processes;* and third, *locational fundamentals,* by which they mean that random growth results from randomness in the physical and economic characteristics of locations themselves. Davis and Weinstein argue that these three theories, which we discuss earlier in the chapter, have very *different testable predictions* for the impact on the size distribution from a powerful but *temporary shock*.

Davis and Weinstein (2002) argue that the great deal of variation in regional densities suggests factors other than increasing returns are important in determining regional densities. The extraordinary changes in technology over the length of the study would have produced radical shifts in the urban structure over time, which are not observed. Random growth, on the other hand, is consistent with the facts, provided that the underlying stochastic process satisfies certain conditions. The locational fundamentals theory could easily explain persistence, as certain physical features of the landscape, like proximity to waterways and the ocean, have not been altered even with the intense bombing that Japan (and Germany) suffered. They

interpret the great deal of persistence in population densities over time that they find as strong support of the locational fundamentals theory.

They interpret the evidence on the robustness of Japan's urban system as against the increasing returns theory and in favor of the locational fundamentals theory. They conclude that the evidence is consistent with a hybrid theory whereby locational advantages help establish basic patterns of regional densities and increasing returns, or random growth, help determine the degree of concentration. Davis and Weinstein interpret the remarkable recovery of Japan's urban system fact as evidence against random growth. While their results are very interesting, they need not warrant this conclusion. In the terms of section 3.2.2 of this chapter, they show evidence for a mean-reverting component in the growth process. This is still a priori compatible with the main condition of random growth models, the presence of a unit root. They do not purport to reject the existence of such a unit root.

The same caveat applies to Brakman *et al.* (2002). Like Davis and Weinstein, they estimate an equation for the growth rate during $1946+t$ and 1946, where t assumes alternative values 4, 17 and 18 in order to distinguish between short-term and long-term effects, with German city data. They separate the sample into West and East Germany, for $t = 4, 17$ and $t = 4, 18$, respectively. They conclude that when the entire of Germany and West Germany only are studied, the impact of the bombing is significant but temporary. The East German urban system, if treated separately, obeys a random walk. They attribute this difference to the different socioeconomic systems prevailing in the two parts of Germany following WW II, with East Germany having been a communist state from 1949 until its absorption into the Federal Republic of Germany in 1989. The post World War II division of Germany might have created border effects for those cities near the border. It is interesting that in spite of the prevalence of central planning in East Germany, the East German urban system might not have been altered and thus remained affected by its state at the end of WW II. In contrast, the outcome for the urban system within the free market system in West Germany was not conditioned by the bombing. They find this to be consistent with the locational fundamentals theory but not necessarily with the increasing returns theory. It is interesting that the study of the urban dynamics in Germany and Japan provides evidence for such different results. Perhaps this is due to the very different geographies of those two countries, which might have prevented the operation of market forces' altering the urban system in Japan but not in Germany.

These historical studies have clearly opened up new horizons for eco-

nomic research. Still, they may be providing additional evidence on the resilience of urban systems in-the-large. For example, repeated destructions of urban settlements in Europe have always been followed by reconstruction along the earlier patterns. But it is also true, as the chapter by Hohenberg in this handbook [Hohenberg (2003)] articulates, that *persistence* of the urban structure historically must always be studied in terms of *fully dynamic models*. Such an approach finds us in full agreement.

7 Conclusion

Because Zipf's law appears to be a quite robust empirical regularity, this survey put some emphasis on it. Two related empirical regularities are Gibrat's law for means and Gibrat's law for variances. They have been less systematically studied, so more research is warranted to study their empirical validity (though initial assessments appear favorable to Gibrat's laws). These three laws offer a strong benchmark against which to measure theories of urban evolution and to organize an up to date look at the literature. The robustness of Zipf's law has also served to attract attention to the need for microfoundations.

The paper reviews a number of theories, some of them very recent, whose implications match those laws quite closely, at least approximately and within the confidence intervals in which those laws themselves hold. Most existing theories until very recently did not easily accommodate these laws. The classical urban (system of cities) theory may accommodate it as outcome of very special assumptions about preferences and technology. Some of its recent variants offer much more precise predictions and, notably, also explain departures from Zipf's law that we observe at the extremes of the city size distribution. The new economic geography literature may also accommodate it, albeit in very simple models. We do not know whether this accommodation would survive in more complex models. As the revival of interest in these topics fosters additional research with enriched theories of urban growth and development, we think that several important issues deserve attention. Notable such issues are the robustness of urban evolution, in spite of the presence of stochastic forces, and the role of economic integration and international trade.

8 Appendix: Zipf's Law and Urban Primacy

Ordering cities by size, the $k-$primacy π_k is the ratio between the size of the largest city and the sum of the population of the k largest cities. Formally:

$$\pi_k = \frac{S_{(1)}}{\sum_{i=1}^{k} S_{(i)}} \tag{17}$$

It lies between 0 and 1. A large π_k indicates that the largest city is quite large. In this appendix we describe the predictions of Zipf's law for this. The Rényi representation theorem cited in section 2.2.2 gives that, for $i < j$, the difference $\ell n S_{(i)} - \ell n S_{(j)}$ can be written:

$$\ell n S_{(i)} - \ell n S_{(j)} = \sum_{h=i}^{j-1} \frac{\tau_h}{h} \tag{18}$$

where the τ_k are independent draws of an exponential distribution $P(\tau_k > \tau) = e^{-\tau}$ for $\tau \geq 0$. A consequence of this is the distribution of the k first cities in a sample of n cities depends only on the ratios $\left(S_{(1)}/S_{(k)}, ..., S_{(k-1)}/S_{(k)}\right)$ and doesn't depend on the specific value of n. So, to sample $k-$primacies (and most statistics), it is enough to draw k cities from a Zipf distribution, rather than draw n cities and take the k biggest cities. One can do that by drawing k i.i.d. random variables u_i from a uniform distribution in $[0, 1]$ and sets the sizes as $S_i = 1/u_i$. One sorts them to get the ordered sizes $S_{(i)}$. One gets the corresponding $k-$primacy ratio π_k. The results are reported in Table 3.

k	2	3	5	10	50
Mean π_k	0.693	0.590	0.502	0.424	0.323
S.D. of π_k	0.140	0.172	0.193	0.205	0.208
95% C.I.	[0.506,0.975]	[0.361,0.961]	[0.251,0.944	[0.168,0.922]	[0.090,0.879]

Table 3 Statistics on the $k-$primacy ratio π_k.
π_k is the ratio between the size of the largest city and the sum of the population of the k largest cities. The table reports the mean of π_k, its standard deviation, and a 95% confidence interval for π_k. Source: Authors' calculations, based on 1 million Monte-Carlo simulations for each k.

A conclusion from Table 3 is that the confidence intervals are extremely wide. This could be guessed from the result cited in footnote 3, that under

Zipf's law, the ratio of the largest city to second largest city has smallest 95% confidence interval equal to $[1, 20]$. Various authors look to explain variations in urban primacy. But Table 3 suggests that the large variations in urban primacy across countries are just what Zipf's law predicts. In a sample of 44 countries, Rosen and Resnick (1980) finds a 5-primacy of 0.49 (S.D. 0.12), which is very close to the Zipf prediction, and a 50-primacy of 0.24 (S.D. 0.098). Thus Rosen and Resnick's 50-primacy number is a bit less than expected from Zipf. This maybe be due to the quality of their data. Soo (2003) finds a for the 5-primacy mean 0.50 with a standard deviation of 0.13. For the 10-primacy, he finds a mean of 0.39, with a mean of 0.13. The results are thus extremely close to the Zipf predictions.

References

[1] Ades, A. and E. Glaeser (1995),"Trade and circuses: explaining urban giants", Quarterly Journal of Economics 110:195–228.

[2] Anas, A. (2001), "The spatial economy: cities, regions and international trade", Regional Science and Urban Economics 31:601–15.

[3] Auerbach, F. (1913), "Das Gesetz der Bevölkerungskonzentration", Petermanns Geographische Mitteilungen 59:74–76.

[4] Axtell, R. L. (2001), "Zipf distribution of U.S. firm sizes", Science 293:1818–1820.

[5] Axtell, R.L., and R. Florida (2001), "Emergent cities: a microeconomic explanation of Zipf's Law", paper presented at the Society for Computational Economics, Yale University.

[6] Bairoch, P., (1988), Citites and Economic Development: From the Dawn of History to the Present, Christopher Braider, tr.(University of Chicago Press, Chicago).

[7] Bairoch, P., J. Batou, P. Chevre (1988) " The population of european cities 800- 1850: Data bank and short summary of results", Centre of International Economic History series, no. 2 (Geneva University).

[8] Beeson, P. E., and D. N. DeJong (2002), "Divergence", Contributions to Macroeconomics 2:1049.

[9] Beeson, P. E., D. N. DeJong, and W. Troesken (1999), "Population growth in U.S. counties: 1840–1990", Department of Economics, University of Pittsburgh, mimeo.

[10] J. Beirlant, G. Dierckx, Y. Goegebeur, G. Matthys (1999), "Tail index estimation and an exponential regression model", Extremes: 2177.

[11] Black, D., and J. V. Henderson (1999), "A theory of urban growth", Journal of Political Economy 107: 252–284.

[12] Black, D., and J. V. Henderson (forthcoming), "Urban Evolution in the USA", Journal of Economic Geography.

[13] Brakman, S., H. Garretsen, and C. van Marrewijk (2001), An Introduction to Geographical Economics (Cambridge University Press, Cambridge and New York).

[14] Brakman, S., H. Garretsen, C. Van Marrewijk and M. van den Berg (1999), "The return of Zipf: a further understanding of the rank-size distribution", Journal of Regional Science 39:183–213.

[15] Brakman, S., H. Garretsen, M. Schramm (2002), "The strategic bombing of german cities during WWII and its impact on cities growth", CESifo working paper no. 808, CESifo, Munich.

[16] Carlino, G., and S. Chatterjee (2001), "Aggregate metropolitan employment growth and the deconcentration of metropolitan employment", Journal of Monetary Economics 48:549–583.

[17] Carlino, G., and S. Chatterjee (2002), "Employment deconcentration: a new perspective on America's postwar urban evolution", Journal of Regional Science 42: 455–475.

[18] Champernowne, D. (1953), "A model of income distribution", Economic Journal 83: 318-51.

[19] Córdoba, J.-C. (2003), "On the distribution of city sizes", working paper, Rice University.

[20] Davis, D. R. (2002), Review of The Spatial Economy, Cities, Regions and International Trade, by M.Fujita, P. R. Krugman, and A. J. Venables, in: Journal of International Economics 57: 247-251.

[21] Davis, D. R., and D. E. Weinstein (2002), "Bones, bombs and break points: the geography of economic activity", American Economic Review 92: 1269–1289.

[22] De Vries, J. (1984), European Urbanization, 1500– 1800, (Harvard University Press,Cambridge, MA).

[23] Dobkins, L. H., and Y. M. Ioannides (2000), "dynamic evolution of the U.S. city size distribution", in : J. Huriot and J. Thisse, eds., The Economics of Cities,Theoretical Perspectives (Cambridge University Press, Cambridge) 217–260.

[24] Dobkins, L. H., and Y. M. Ioannides (2001), "Spatial interactions among U.S. cities", Regional Science and Urban Economics 31:701–731.

[25] Duranton, G. (2002), "City size distribution as a consequence of the growth process", Department of Geography and Environment, London School of Economics.

[26] Eaton, J., and Z. Eckstein (1997), "Cities and growth: theory and evidence from France and Japan", Regional Science and Urban Economics 27:443–474.

[27] Embrechts, P., C. Kluppelberg, T. Mikosch (1997), Modelling Extremal Events for Insurance and Finance (Springer, New York).

[28] Feuerverger, A. and P. Hall (1999), "Estimating a tail exponent by modelling departure from a pareto distribution", Annals of Statistics 27:760-781.

[29] Florida, R. (2002) The Rise of the Creative Class: And How Its Transforming Work, Leisure Community and Everyday Life, (Basic Books, New York).

[30] Fujita, M., and T. Mori (1997), "Structural Stability and Evolution of Urban Systems", Regional Science and Urban Economics 27:399– 442.

[31] Fujita, M., P. Krugman, and T. Mori, (1999), "On the Evolution of hierarchical urban systems", European Economic Review 43:209–251.

[32] Fujita, M., P. Krugman and A. J. Venables (1999), The Spatial Economy (MIT Press, Cambridge, MA).

[33] Fujita, M., and J-F. Thisse (2002), Economics of Agglomeration (Cambridge University Press, Cambridge).

[34] Gabaix, X. (1999a), "Zipf's Law and the growth of cities", American Economic Review, Papers and Proceedings 89:129-32.

[35] Gabaix, X. (1999b), "Zipf's Law for cities: an explanation", Quarterly Journal of Economics 114: 739–767.

[36] Gabaix, X., P. Gopikrishnan, V. Plerou, H. E. Stanley (2003),"A theory of power laws in financial fluctuations", Nature 423:267-70.

[37] Gabaix, X. and Y. Ioannides (2003), "The properties of the least squares estimates of power law exponents", (MIT and Tufts University).

[38] Gibrat, R., (1931), "Les inégalités économiques", Librairie du Recueil Sirey, Paris, France.

[39] Glaeser, E. L., J. A. Scheinkman, A. Shleifer (1995), "Economic growth in a cross-section of cities, Journal of Monetary Economics 36:117-143.

[40] Grossman, G. M. and E. Helpman (1991), "Quality ladders in the theory of growth", Review of Economic Studies 58:43–61.

[41] Henderson, J. V. (1974), "The types and size of cities", American Economic Review, 64: 640–656.

[42] Henderson, J. V. (1988), Urban Development: Theory, Fact and Illusion (Oxford University Press, Oxford).

[43] Hill, B. M. (1975), "A simple approach to inference about the tail of a distribution", Annals of Statistics 3:1163–1174.

[44] Hohenberg, Paul M. (2003), "The historical geography of European cities: an interpretive essay," this Handbook.

[45] Ioannides, Y. M., and H. G. Overman (2003)," Zipf's Law for cities: an empirical examination", Regional Science and Urban Economics 33: 127–137.

[46] Ioannides, Y. M., and H. G. Overman, (2000), "Spatial evolution of the U.S. urban system", Journal of Economic Geography, forthcoming (Tufts University, Medford MA) http://ase.tufts.edu/econ/papers/index.html.

[47] Kesten, H. (1973) "Random difference equations and renewal theory for products of random matrixes", Acta Mathematica 131: 207-48.

[48] Kim, S.(2000), "Urban development in the United States, 1690-1990", Southern Economic Journal.

[49] Kim, S. (2002), The reconstruction of the American urban landscape in the twentieth century Working paper no. w8857 (National Bureau Economic Research).

[50] Krugman, P. (1991), "Increasing returns and economic geography", Journal of Political Economy 99: 483–499.

[51] Krugman, P. (1996a), The Self-Organizing Economy (Blackwell Publishers Oxford, UK and Cambridge, MA).

[52] Krugman, P. (1996b), "Confronting the mystery of urban hierarchy", Journal of the Japanese and the International Economies 10: 399–418.

[53] Leamer, E. and J. Levinsohn (1995), "International trade theory: the evidence", in: G. Grossman and K. Rogoff, eds., Handbook of International Economics, Vol III (North-Holland, Amsterdam) 1339-1394.

[54] Levy, M. and S. Solomon (1996), "Dynamical explanation for the emergence of power law in a stock market model", International Journal of Modern Physics C 7:65-72.

[55] Lucas, Jr., R. E. (1988), "On the mechanics of economic development", Journal of Monetary Economics, 22: 3-42.

[56] Madden, C. H. (1956), "On some indications of stability in the growth of cities in the United States", Economic Development and Cultural Change 4: 236–252.

[57] Malcai, O., O. Biham and S. Solomon (1999), "Power-law distributions and lévy-stable intermittent fluctuations in stochastic systems of Many Autocatalytic Elements", Physical Review E 60:1299.

[58] Marsili, M. and Y.C. Zhang (1998), "Interacting individuals leading to Zipf's Law", Physical Review Letters 80:2741-44.

[59] Moretti, E. (2004), "Human capital externalities and cities", in: Henderson and J. Thisse, eds., Handbook of Urban and Regional Economics, vol. 4 (North Holland, Amsterdam).

[60] Mori, T., K. Nishikimi and T. Smith (2003), "Some Empirical Regularities of Spatial Economies: A Relationship between Industrial Location and City Size", Kyoto University Mimeo.

[61] Neary, P. (2001) "Of hype and hyperbolas: introducing the new economic geography", Journal of Economic Literature 39: 536–561.

[62] Okuyama, K., M. Takayasu and H. Takayasu, "Zipf's Law in income distribution of companies", Physica A 269:125–131.

[63] Overman, H. G., and Y. M. Ioannides (2001) "Cross-sectional evolution of the U.S. city size distribution", Journal of Urban Economics 49: 543–566.

[64] Quah, D. (1993), "Empirical cross-section dynamics and economic growth", European Economic Review 37:426-434.

[65] Reed, W. (2001) "The Pareto, Zipf and other power law", Economics Letters, 74:15-19.

[66] Reed, W. (2002) "On the rank-size distribution for human settlements", J. Regional Science, 41: 1-17

[67] Reed, W. and B. Hughes (2002), "From gene and genera to incomes and internet files: why power laws are so common in nature", Physical Review E 66:067103/1-4.

[68] Reiss, R. (1989), Approximate Distributions of Order Statistics, (Springer Verlag, Berlin).

[69] Romer, P. M. (1990), "Endogenous technological change", Journal of Political Economy 98: S71-102.

[70] Rosen, K. and M. Resnick (1980), "The size distribution of cities: an examination of the Pareto Law and primacy", Journal of Urban Economics 8:165-186.

[71] Rossi-Hansberg, E., and M. L. J. Wright (2003), "Urban structure and growth", Working paper (Stanford University).

[72] Simon, H.(1955), "On a class of skew distribution functions", Biometrika, 44:425-440, and reprinted in: (1957) Models of Man: Social and Rational. Mathematical Essays on Rational Human Behavior in a Social Setting, (Wiley and Sons, New York).

[73] Soo, K. T. (2003), "Zipf's Law for cities: a cross country investigation", Working paper (Centre for Economic Performance, London School of Economics).

[74] Sornette, D. (2001) Critical Phenomena in Natural Sciences (Springer Verlag, Berlin and New York).

[75] Stock, J. (1994) "Unit roots, structural breaks, and trends", ch. 46 in: R. Engle and D. McFadden, eds., Handbook of Econometrics, volume IV (Amsterdam: Elsevier) 2740-2843.

[76] Suarez-Villa, L. (1988), "Metropolitan evolution, sectoral economic change, and the city size distribution", Urban Studies 25:1–20.

[77] Tabuchi, T. (1986), "Existence and stability of city-size distribution in the gravity and logit models", Environment and Planning A 18:1375–1389.

[78] Van der Woude, A., J. de Vries, A. Hayami (1990), "The hierarchies, provisioning, and Demographic patterns of cities", in: Urbanization in History: A Process of Dynamic Interactions (Oxford and New York, Oxford University Press)1-19.

[79] Wheaton, W. and H. Shishido (1981), "Urban concentration, agglomeration economies and the level of economic development", Economic Development and Cultural Change 30: 17–30.

[80] Zanette, D. H. and S. C. Manrubia (1997), "Role of Intermittency in Urban Development: A Model of Large-Scale City Formation", Physical Review Letters 79:523–6

[81] Zipf, G. K. (1949), Human Behavior and the Principle of Least Effort (Addison-Wesley, Cambridge, MA).

files/eudoraold/attach/ZipfPlotUSA1991.wmf

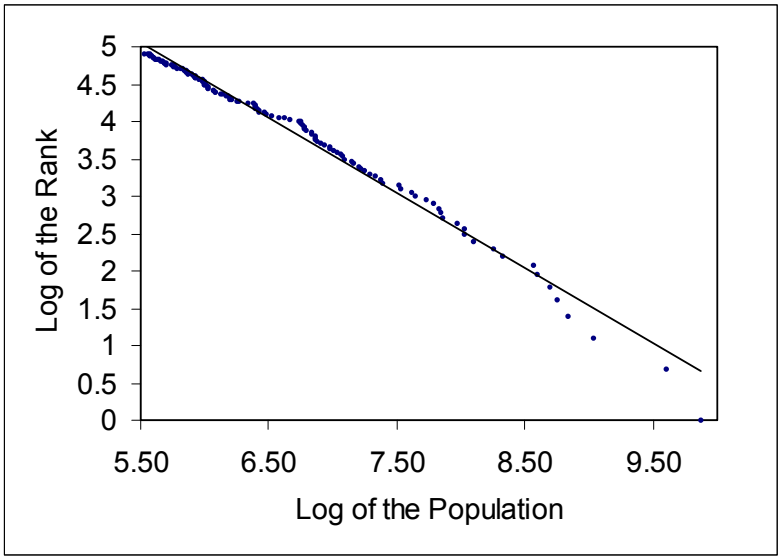

Figure 1: Log Size vs Log Rank of the 135 U.S. Metropolitan Areas in 1991 listed in the Statistical Abstract of the United States (1993).

files/eudoraold/attach/FirmsSizeAxtellScience.bmp

Log frequency $\ln g(S)$ vs log size $\ln S$ of U.S. firm sizes (by number of employees) for 1997. OLS fit gives a slope of 2.059 (s.e.= 0.054; R^2 =0.992). This corresponds to a frequency $g(S) \sim S^{-2.059}$. Source: Axtell (2001).

files/eudoraold/attach/Fig2a.bmp

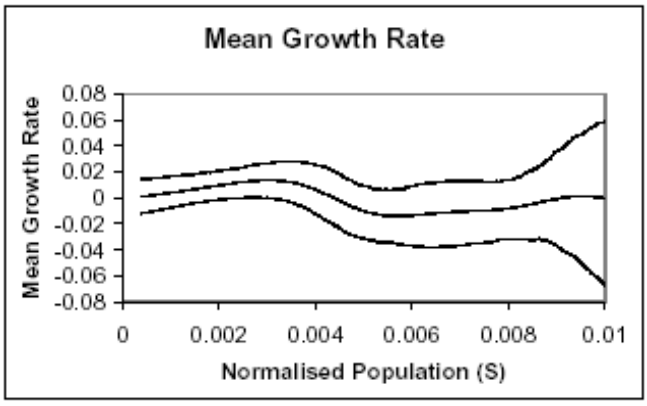

Figure 2: Figure 2(a)

Figure 3: Non-parametric estimates of the mean and variance of the growth rate of a city of size S as a function of the size S. The figure plots the bootstrapped 95% confidence intervals. Source: Ioannides and Overman (2003).

[Note to NorthHolland: The two figures below should be displayed next to each other in one Figure 3]

files/eudoraold/attach/Fig2b.bmp

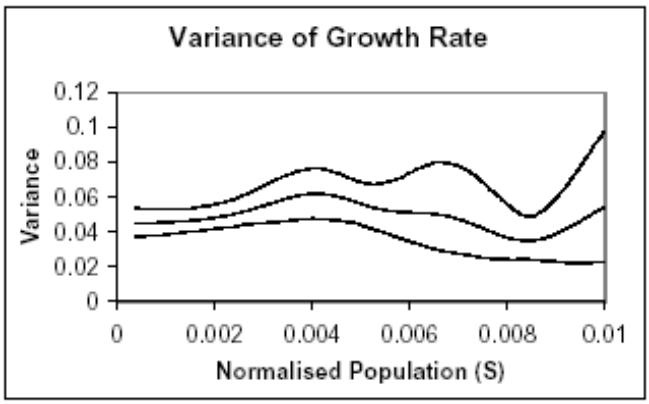

Figure 3: Figure 2(b)