ZIPF'S LAW FOR CITIES: AN EMPIRICAL EXAMINATION[1]

Yannis M. Ioannides

Department of Economics
Tufts University
Medford, Massachusetts 02155, U.S.A.
1-617-627-3294 yioannid@tufts.edu

Henry G. Overman

Department of Geography and Environment
London School of Economics
Houghton Street
London, WC2A 2AE U.K.
44-20-7955-6581 h.g.overman@lse.ac.uk

December 21, 2000

Abstract

We use data for metro areas in the United States, from the US Census for 1900 – 1990, to test the validity of Zipf's Law for cities. Previous investigations are restricted to regressions of log size against log rank. In contrast, we use a nonparametric procedure to estimate Gibrat's Law for city growth processes and to calculate local Zipf exponents from the mean and variance of city growth rates. Despite variation in growth rates as a function of city size, Gibrat's Law does hold. The local Zipf exponents are broadly consistent with Zipf's Law. Deviations from Zipf's Law are easily explained by deviations from Gibrat's Law.

Journal of Economic Literature Codes: R00, C14.
Keywords: Zipf's Law. Gibrat's Law. Urban growth. Estimation of Brownian motion.

[1]Ioannides acknowledges generous research support from John D. and Catherine T. MacArthur Foundation and the National Science Foundation. We thank Xavier Gabaix and Diego Puga for useful comments, and Danny T. Quah for giving us access to his tsrf package.

1. Introduction

This paper reconsiders an alleged 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 Pareto distribution. That is, if we rank cities from largest (rank 1) to smallest (rank N) to get the rank r(p) for a city of size p, then:

$$\log r(p) = \log A \quad - \quad \zeta \log p, \tag{1}$$

where A and ζ are parameters. Zipf (1949) proposed that city sizes follow a special form of the distribution where $\zeta = 1$. This expression of the regularity has become known as Zipf's Law[2].

Gabaix (1999), the latest in a series of notable contributions to this literature, derives a statistical explanation of Zipf's Law for cities. He shows that if different cities grow randomly with the same expected growth rate and the same variance (Gibrat's Law), the limit distribution of city size will converge so as to obey Zipf's Law.

Gabaix's contribution is significant because it addresses the validity of Zipf's Law as the limit of a stochastic process. But the question of the validity of Zipf's Law as an empirical regularity ultimately will rest on reliable econometric findings. Previous empirical investigations have sought to indirectly estimate ζ in Equation (1) by regressing log size against log rank. Obtaining a regression estimate of $\zeta = 1.00$ is then taken as confirmation of Zipf's Law.

Thus, for example, Dobkins and Ioannides (2000a) report OLS estimates of ζ, obtained from repeated cross sections of US Census data for metro areas, that decline from 1.044, in 1900, to .949, in 1990. They also report maximum likelihood estimates for Pareto distributions that decline from .953, in 1900, to .553, in 1990, and for Pareto distributions (but using the upper one-half of the sample only) that decline from 1.212, in 1900, with 112 cities in the entire sample, to .993, in 1990, with 334 cities in the entire sample. Gabaix (1999) obtains an estimate equal to 1.005, using the 135 largest metro areas in 1991. However, despite general satisfaction (and occasional awe) with the fits obtained for Zipf's Law with US city size data, problems remain. Nonparametric results by Dobkins and Ioannides (2000a) and a finding of a significant quadratic term in a log rank regression

[2] Its deterministic equivalent suggests that the second largest city is half the size of the largest, the third largest city a third the size of the largest etc etc. When expressed like this, the regularity is often referred to as the rank size rule.

performed by Black and Henderson (1999) continue to raise genuine doubts about the validity of Zipf's Law, even as an empirical regularity.

In view of Gabaix's results, an econometric examination may rest on *either* the size distribution of cities *or* the growth process of cities. There are a large number of studies based on the former approach. To our knowledge, this paper constitutes the first attempt to use the latter approach to test the validity of Zipf's Law. We believe that in either case an approach is needed which is not confined to linear regression techniques that in effect assume the existence of a representative city and fit the evolution of its mean. It is for these reasons that this paper reconsiders the recent econometric work, which alleges to be supportive of Zipf's Law.

Section 2 of the paper briefly reviews the basic statistical approach of Gabaix to provide the foundation for our econometric findings presented in Sections 3 and 4. Section 5 concludes.

2. Random Growth and Size Distribution of Cities

Let S_i denote the normalized size of city i, that is, the population of city i divided by the total urban population. Following Gabaix, *op. cit.*, city sizes are said to satisfy Zipf's Law if the countercumulative distribution function, $G(S)$, of normalized city sizes, S, tends to

$$G(S) = \frac{a}{S^{\zeta}}, \tag{2}$$

where a is a positive constant and $\zeta = 1$.

Gabaix shows that the distribution of city sizes will converge to $G(S)$, given by equation (2), if Gibrat's Law holds for city growth processes. That is, if city growth rates are identically distributed independent of city size[3]. In Section 4 we test for this independence and show that, despite some variation in growth rates as function of city size, Gibrat's Law does hold for US city growth processes.

Recognizing the possibility that Gibrat's Law might not hold exactly, Gabaix also examines the case where cities grow randomly with expected growth rates and standard deviations that depend

[3]It is straightforward to verify this claim as follows. 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 are independently and identically distributed random variables with density function $f(\gamma)$, and given that the average normalized size must stay constant, $\int_0^\infty \gamma f(\gamma) d\gamma = 1$, then the equation of motion of the distribution of growth rates expressed in term of the countercumulative distribution function of S_t^i, $G_t(S)$, is

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

It is satisfied by $G(S) = \frac{a}{S}$.

on their sizes. That is, the size of city i at time t varies according to Equation (11), *ibid.,* p. 756, replicated here:

$$\frac{dS_t}{S_t} = \mu(S_t)dt + \sigma(S_t)dB_t, \tag{3}$$

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 geometric Brownian motion. In this case, the limit distribution of city sizes will converge to a law with a *local* Zipf exponent, $\zeta(S) = \; - \; \frac{S}{p(S)}\frac{dp(S)}{dS}$, where $p(S)$ denotes the invariant distribution of S. Working with the forward Kolmogorov equation associated with equation (3), the local Zipf exponent, associated with the limit distribution, can be derived and is given by Equation (13) in *ibid.,* p. 757, again replicated here:

$$\zeta(S) = 1 \; - \; 2\frac{\mu(S)}{\sigma^2(S)} + \frac{\partial\sigma^2(S)/\sigma^2(S)}{\partial S/S}, \tag{4}$$

where $\mu(S)$ is relative to the overall mean for all city sizes. This expression for the local Zipf exponent in terms of the mean and variance of growth rates forms the basis of our empirical approach.

Variations of the Zipf exponent from above one to below one are quite critical for the statistical robustness of the finding that the distribution of city sizes obeys a Pareto Law. If ζ is less than one, then the distribution has neither finite mean nor finite variance, and if it is less than 2, but more than 1, it has finite mean but not finite variance. Before any further (nearly) mystical significance is attributed to Zipf's exponent for U.S. (and other) city size data it behooves us to fully explore its origins.

Gabaix's theoretical contribution offers an opportunity for a direct test of a possible foundation for Zipf's Law in the form of Gibrat's Law for city growth rates. Our empirical approach allows for a city's growth rate to depend on city size and to vary according to a law like equation (3) above. To do this, we non-parametrically estimate the mean and variance of city growth rates conditional on size. This allows us to test the validity of Gibrat's Law. We then use equation (4) 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.

3. Nonparametric Estimation of the Distribution of Growth Rates Conditional on City Size

Before we consider conditional means and variances, we briefly consider the entire distribution of growth rates conditional on city size. To do this, we non-parametrically estimate a stochastic kernel — a three dimensional representation of the distribution of growth rates conditional on city size. Figures 1.a and 1.b report the stochastic kernel and contour for the entire sample. Figures 1.c and 1.d report stochastic kernels and contours of stochastic kernels for the top 110 cities, of the largest 135 used by Gabaix. Figures 1.e and 1.f report stochastic kernels and contours of stochastic kernels for top 110 cities but excluding the last two decades.[4] All figures use relative population growth rates against relative population, where relative is meant with respect to total urban population. Both growth rates and populations are normalized by subtracting the average for the sample period and dividing by the appropriate standard deviation.

To better understand the information provided by the stochastic kernel, take any point on the population axis corresponding to a particular city size S, and take a cross-section through the stochastic kernel parallel to the growth axis. This cross-section gives us a (non-parametric) estimate of the distribution of growth rates conditional on city size S. The stochastic kernel just reports this conditional distribution for all values of S.[5]

The plot for the entire sample suggests that, except for the very smallest cities, the conditional *distribution* of growth rates is remarkably stable across city sizes. The plot for the largest 110 metro areas[6], is not quite as clear cut. Growth rates for the middle size cities appear concentrated, but there are some large outlying (negative) growth experiences for these cities. Similarly, growth rates

[4] All stochastic kernels are calculated nonparametrically using a Gaussian kernel with bandwidth set as per section 3.4.2 of Silverman (1986). To estimate the kernel,we first derive the joint distribution of normalised population and growth rates. We then numerically integrate under this joint distribution with respect to growth rates, to get the marginal distribution of population at time t. Finally, we estimate the marginal distribution of growth rates conditional on population size by dividing the joint distribution by the marginal distribution. Calculations were performed with Danny T. Quah's tsrf econometric shell. The contours work exactly like the more standard contours on a map. Any one contour connects all the points on the stochastic kernel at a certain height.

[5] Both population and growth rates are calculated relative to their (time varying) means. In addition, when pooling across years, we normalise by the total standard deviation for each variable. This makes for a clear presentation, but does not artificially induce any of the results which we discuss subsequently.

[6] The plot uses data across all ten time periods for the largest 110 metro areas in 1990, in 1990. In contrast, Gabaix uses data just for 1990

for the top size cities are fairly concentrated except for some large outlying (positive and negative) growth experiences. Interestingly, experimentation suggests that the larger variance for the largest cities is driven by very recent experience. Figures 1.e and 1.f show the plot for the largest 110 cities excluding the last two decades — now, we can see that the variance of the growth rate does decline for the largest cities over most of the sample period, at least to the extent that one can discern from the contour plots. These findings confirm the importance of looking at the entire evolution of city sizes and growth rates, as the snapshot offered by 1990 data is clearly not sufficient to derive firm conclusions. They are clearly more informative than descriptive statistics based on raw data [*c.f.* Dobkins and Ioannides (2000b)], and findings that we have obtained ourselves with less smoothing. Indeed, our results in this section suggest that there are some stable aspects to the distribution of growth rates with respect to city size.

4. Nonparametric Estimation of the Local Zipf Exponent

If the growth process governing the evolution of city sizes is stable overtime, then we can pool data from our panel of cities to calculate city growth rates conditional on normalised city size[7]. We can then directly calculate the value of the Zipf exponent as a function of city size (the local Zipf exponent) as per Equation (13).

Pooling across time gives us 1654 population-growth rate pairs on which to base our estimates. For each population-growth rate pair, normalised population, S, is defined as the city's share of total urban population in the relevant decade. Growth rate, $\mu(S)$, is defined as the difference between a city's growth rate and the mean city growth rate in the relevant decade[8]. The nonparametric estimates of the conditional mean and variances, and the derivates used to calculate the Zipf exponent, are derived according to the Nadaraya-Watson method. Unless otherwise stated, bandwidths are calculated as per Equation 3.31 in Silverman (1986). See Härdle (1990) and Silverman (1986) for details.

Figure 2.a - 2.b give nonparametric estimates of the conditional mean and variance of growth

[7]Results in Black and Henderson (1999) testing for the stability of the Markov-process governing city transitions suggests that such pooling is valid.

[8]One slight modification to Equ. (4) is needed when applied to real data. Namely, as we have done here, we need to normalise by time varying mean city growth rates, rather than a common mean city growth rate.

rates. The figures also show 5% bootstrapped confidence bands[9]. It is immediately apparent that Gibrat's Law does not hold exactly for city growth processes - both the mean and variance vary with city size. However, note that a constant variance and constant (zero) mean growth rate across all city sizes would lie within the 5% confidence bands. This suggests that we cannot formally reject Gibrat's Law for city growth processes. Despite this, the fact that Gibrat's Law does not hold exactly does have interesting implications for Zipf's Law as suggested in our discussion of Equations (3) and (4) above. We return to this issue below.

We can use these nonparametric estimates to calculate the local Zipf exponent as outlined above. The results are presented in Figure 2.c. There is one technical problem with this procedure - the sparsity of data at the upper end of the distribution. Figure 2.d shows just how severe a problem this is at the upper end of the distribution. The figure shows 5% bootstrapped confidence bands for the Zipf coefficient estimate. These bands are so wide at the upper end of the distribution that we have chosen to restrict the sample range. Thus, the figures actually report results for city shares ranging from 0% to 10%. Table 1 shows the number of observations falling in to any given range. From the table, we see that the sample restriction excludes 145 observations corresponding to cities with population shares greater than 10% of the urban population. This is equivalent to excluding approximately 16 cities over the entire sample period[10]. Even with this choice of cut-off, the estimates at the upper end of the range (where the Zipf exponent fluctuates considerably) are based on very few observations. To get round this, Figure 2.e reports results for the Zipf exponent estimated using a larger bandwidth[11]. This oversmooths at the lower end of the distribution, but gives more reasonable values for the Zipf exponent at the upper end of the distribution.

There is actually considerable variation in the estimates of the Zipf exponent. As suggested by Gabaix (1999), we can understand deviations from a Zipf exponent of one, by considering the mean and variance of growth rates for cities in any given range[12]. Thus, for cities around 0.2% of the

[9] The bootstrapped confidence bands are based on 500 samples. Sampling is with replacement and bandwidth is re-calculated for each sample. The bands are based on individual confidence points for each of 1000 grid points on the normalized population axis. See Härdle (1990) Section 4.2-4.3 for details.

[10] The largest cities will have been in from the start of the sample and thus we will have nine data points for each city. However, because even the largest cities change rank over time, see Overman and Ioannides (1999), different cities may be excluded in different years.

[11] The bandwidth that we use is h=0.002 which is approximately double the optimal bandwidth used for Figures 2.a-d.

[12] That is, by considering deviations from Gibrat's Law.

urban population, we can see from Figure 2.a and Figure 2.b, that the mean growth rates are high and the variance in those growth rates is relatively low. When cities have high growth rates, small cities constantly feed the stock of larger cities and we would expect the distribution to decay less quickly. That is, we would expect a Zipf exponent less than one. For cities around 0.45% of the urban population, mean growth rates have fallen somewhat, but the variance of the growth rate is high. Again, this leads to a low Zipf exponent due to both the growth effect, and the fact that high variance of city growth rates leads to mixing of smaller and larger cites. Finally, cities around 0.85% of the urban population have average growth rates, around average variance in those growth rates and, consequently, a Zipf exponent close to one.

Our findings also help explain two interesting features of the size distribution of US cities. First, as outlined above, estimates of the Zipf exponent for US cities decline overtime. Gabaix suggests that one 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. Our calculations 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[13]. Our 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.

5. Conclusion

We have proposed and implemented a methodology for testing for the validity of Zipf's Law for cities and for calculating local Zipf exponents for the US city size distribution. We have two key findings. First, Gibrat's Law broadly holds for city growth processes. Second, Zipf's Law does hold approximately for a large range of city sizes. However, our results suggest that local values of the Zipf exponent can vary considerably across city sizes. As suggested by Gabaix, these variations of the local Zipf exponent can be understood by considering mean growth rate and variances in growth rates conditional on city sizes. Further, our estimates of local Zipf exponents help us to understand several well-documented features of the US city size distribution. More fundamentally, our method

[13] Again, see Dobkins and Ioannides (2000a).

for calculating the Zipf exponent is quite applicable to estimation of geometric Brownian motion models, where the parameters of the stochastic structure are not constant.

6. References

Auerbach, F. (1913), "Das Gesetz der Belvolkerungskoncentration" *Petermanns Geographische Mitteilungen*, 59:74-76.

Black, Duncan, and J. Vernon Henderson (1999), "Urban Evolution in the USA,"working paper, Department of Economics, London School of Economics.

Dobkins, Linda H., and Yannis M. Ioannides (2000a), "Dynamic Evolution of the U.S. City Size Distribution," 217–260, J.-M. Huriot and J.-F. Thisse, eds.,*Economics of Cities,* Cambridge University Press, New York.

Dobkins, Linda H., and Yannis M. Ioannides (2000b), "Spatial Interactions among U.S. Cities," working paper, Tufts University, *Regional Science and Urban Economics,* forthcoming.

Gabaix, Xavier (1999), "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics,* CXIV, August,739 – 767.

Härdle, Wolfgang (1990), *Applied Nonparametric Regression,* Cambridge University Press, Cambridge.

Ioannides, Yannis M., and Henry G. Overman (1999), "Spatial Evolution of the U.S. Urban System," working paper, Tufts University and London School of Economics, November.

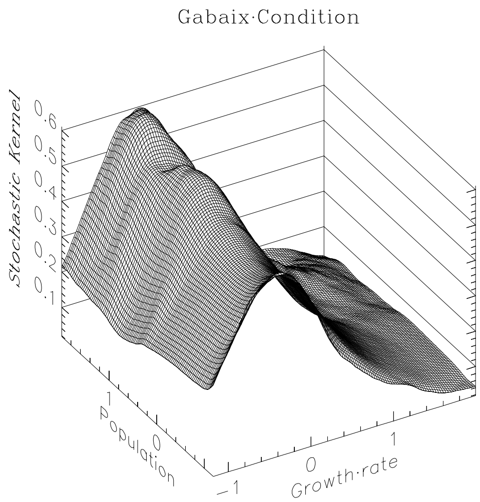
Overman, Henry G. and Yannis M. Ioannides (1999), "Cross-Sectional Evolution of the US City Size Distribution," working paper, Tufts University and London School of Economics, December; *Journal of Urban Economics,* forthcoming.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, New York.

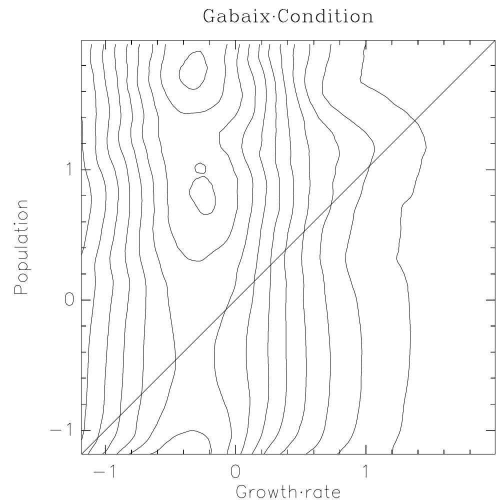
Zipf, G. (1949), *Human Behaviour and the Principle of Least Effort* Reading MA: Addison-Wesley.

Population share	Number of observations
0.000-0.002	734
0.002-0.004	433
0.004-0.006	163
0.006-0.008	114
0.008-0.010	46
0.010-0.012	36
0.012-0.180	109

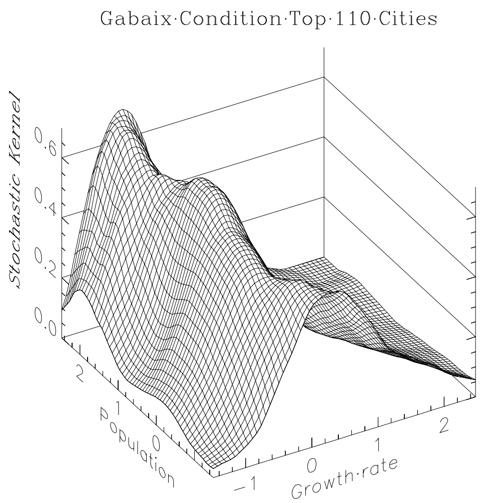
Table 1. Distribution of pooled observations by city sizes



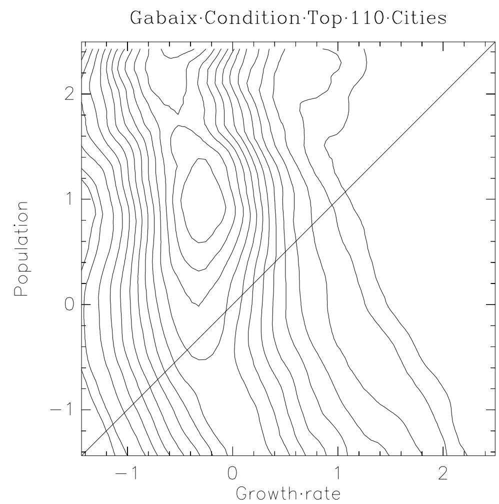
a: All cities - stochastic kernel



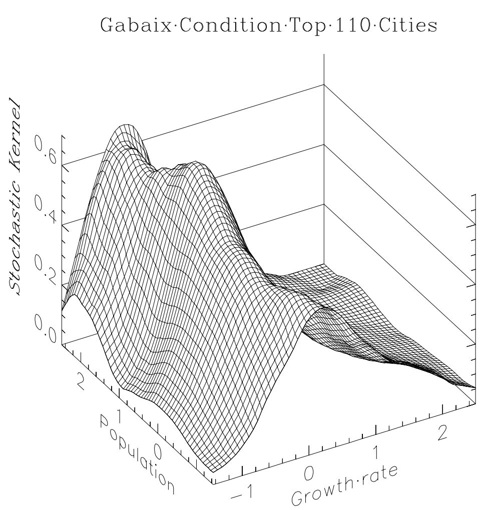
b: All cities - contour



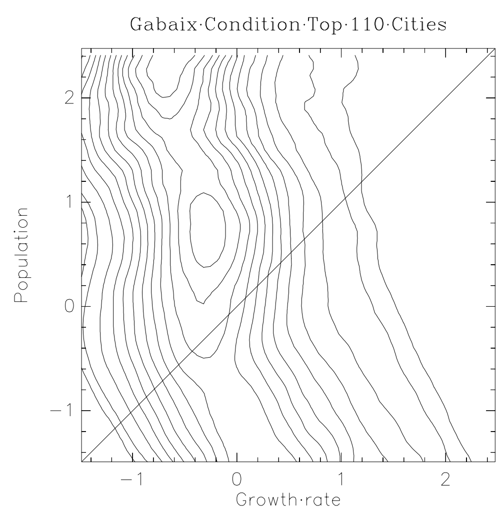
c: Top 110 cities - stochastic kernel



d: Top 110 cities - contour

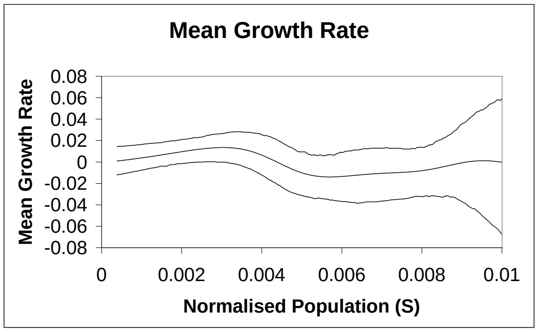


e: Top 110 cities (not 80-90) - stochastic kernel

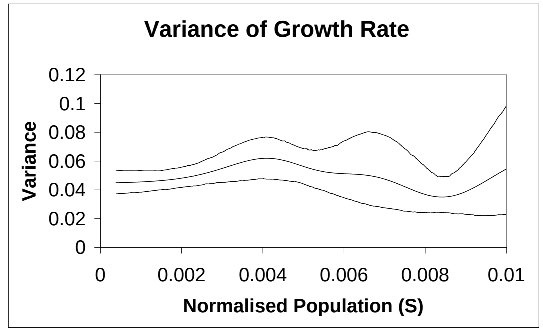


f: Top 110 cities (not 80-90) - contour

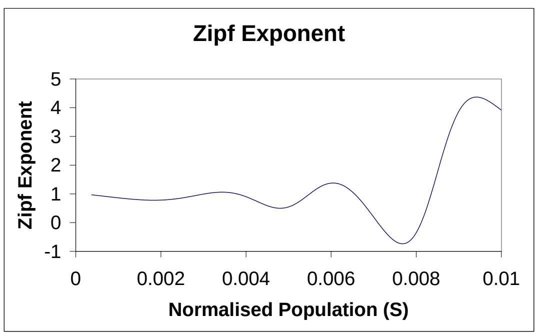
Figure 1. Stochastic Kernel - Population to Growth Rates



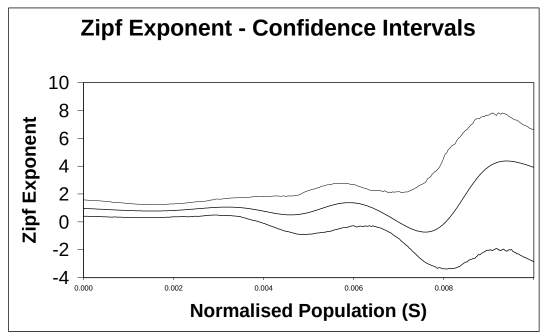
(a) Mean



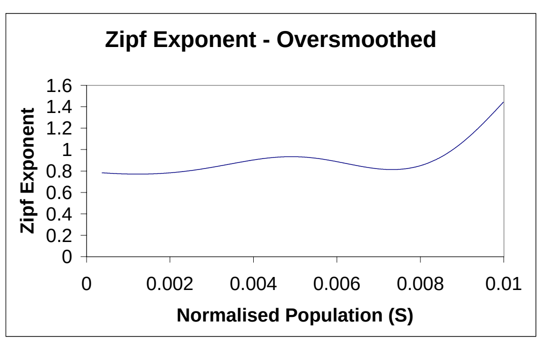
(b) Variance



(c) Zipf



(d) Zipf (confidence bands)



(e) Zipf (oversmoothed)

Figure 2. Nonparametric Estimates