UNR Joint Economics Working Paper Series
Working Paper No. 07-006

Further Analysis of the Zipf's Law: Does the Rank-Size Rule Really Exist?

Fungisai Nota and Shunfeng Song

Department of Economics /0030
University of Nevada, Reno
Reno, NV 89557-0207
(775) 784-6850 | Fax (775) 784-4728
email: fnota@cabnr.unr.edu; song@unr.edu

November, 2007

Abstract

The widely-used Zipf's law has two striking regularities. One is its excellent fit; the other is its close-to-one exponent. When the exponent equals to one, the Zipf's law collapses into the rank-size rule. This paper further analyzes the Zipf exponent. By changing the sample size, the truncation point, and the mix of cities in the sample, we found that the exponent is close to one only for some selected sub-samples. Small samples of large cities alone provide higher value of the exponent whereas small cities introduce high variance and lower the value of the exponent. Using the values of estimated exponent from the rolling sample method, we obtained an elasticity of the exponent with respect to sample size. We concluded that the rank-size rule is not an economic regularity but a statistical phenomenon.

JEL Classification: C1, R11, R12

Keywords: Zipf's law; Rank-size rule; Rolling sample method

Further Analysis of the Zipf's Law: Does the Rank-Size Rule Really Exist?

Fungisai **Nota** (Corresponding Author)
Department of Resource Economics/ MS 204
University of Nevada, Reno
Reno, NV 89503
Email: fnota@cabnr.unr.edu
Tel: 1-775-784-1344
Fax: 1:775-784-1342

And

Shunfeng Song
Department of Economics
University of Nevada, Reno
Reno, NV 89557
Email: song@unr.edu

Abstract:

The widely-used Zipf's law has two striking regularities. One is its excellent fit; the other is its close-to-one exponent. When the exponent equals to one, the Zipf's law collapses into the rank-size rule. This paper further analyzes the Zipf exponent. By changing the sample size, the truncation point, and the mix of cities in the sample, we found that the exponent is close to one only for some selected sub-samples. Small samples of large cities alone provide higher value of the exponent whereas small cities introduce high variance and lower the value of the exponent. Using the values of estimated exponent from the rolling sample method, we obtained an elasticity of the exponent with respect to sample size. We concluded that the rank-size rule is not an economic regularity but a statistical phenomenon.

JEL classification: C1; R11; R12
Keywords: Zipf's law; Rank-size rule; Rolling sample method

Further Analysis of the Zipf's Law: Does the Rank-Size Rule Really Exist?

I. INTRODUCTION:

Zipf's law is a common regularity in natural and social sciences (e.g., Shiode and Batty, 2000; Sinclair, 2001; Li and Yang, 2002; Tachimori and Tahara, 2002). It states that the rank associated with some size S is proportional to S to some negative power. If this power is equal to one, the Zipf's law collapses into the commonly named rank-size rule. This implies that in the case of cities, the second largest city is half the size of the first and the third largest city is one third the size of the largest and so on. In cases where this power is greater than one, it suggests that the second largest city is more than half of the largest city and the third largest city is more than a third of the largest city and so on. An exponent less than one suggests that the second largest city is smaller than half of the largest city and so on. Linearizing this relationship between the rank and size using a log transformation makes it easy to estimate the negative exponent.

There are two striking observations about the Zipf's law. One is its excellent fit. Numerous empirical studies have shown that a linear regression of log-rank on log-size generates an excellent fit (very high R^2-value). For example, Rosen and Resnick (1980) used data from 44 countries and found that R^2-values were above 0.95 for 36 countries, with only Thailand having an R^2-value lower than 0.9 (0.83). Mills and Hamilton (1994) found a R^2-value of 0.99 using 1990 data on 366 U.S. urbanized areas. Song and Zhang (2002) found an R^2-value of 0.91 for 665 Chinese cities in 1998. This astonishing regularity led Krugman (1995, p.44) to say that the rank-size rule is "a major embarrassment for economic theory: one of the strongest statistical phenomenon we know, lacking any clear basis in theory." Fujita et al. (1999, p. 219) stated "the

regularity of the urban size distribution posses a real puzzle, one that neither our approach nor the most plausible alternative approach to city sizes seems to answer."

The second striking observation is about the Zipf coefficient. In the urban literature, the coefficient is very close to 1, thus the rank-size rule holds. Theoretically, Gabaix (1999a, b) has argued that the rank-size rule is natural result of a growth process which is independent of the size of the city. Fujita et al. (1999) suggested that the rank-size rule does indeed approximate to the long-run spatial distribution of a mature spatial system. Empirically, for the 44 countries studied by Rosen and Resnick (1980), the estimated coefficient ranges from 0.809 for cities in Morocco to 1.963 for cities in Australia. Nitsche (2005) analyzed 515 estimates from 29 studies of the rank-size relationship and found that two-third of the estimated coefficients are between 0.80 and 1.20, with a median estimate of 1.09. This finding implies that cities are on average more evenly distributed than suggested by the rank-size rule.

Several explanations have attempted to explain why Zipf's law holds, including economic explanation, the Gibrat's law, and pure statistical phenomenon. The economic explanation of Zipf's law relies on a delicate balance between transportation costs, positive and negative externalities, and productivity differences (Gabaix, 1999b; Fujita et al., 1999). However, this approach has some inherent problems. For example, it is difficult to see how radically different economies such as the U.S., China, and India would hold the same delicate balance of forces across time and space. It is clear that the U.S has different transportation costs, externalities, and productivity differences with India and China; therefore, looking at the economies this way does not give a good reason why the Zipf's law should hold. The second explanation was offered by Gabaix (1999a). Gabaix proved that the Zipf's law derives from the Gibrat's law, which states that the growth process is independent of size. Following this explanation, Zipf's law becomes

the steady state distribution. “The existence of power law can be thought of as due to a simple principle: the scale invariance. Because the growth process is the same at all scales, the final distribution process should be scale invariant. This forces it to be a power law” Gabaix (1999a, p. 744). A more recent and compelling explanation of why the Zipf’s law holds was provided by L. Gan et al. (2006). They proved that a high R^2-value exists because the dependent variable (rank) is generated from the independent variable (size). Using data from China, the U.S., and randomly generated data, they concluded that the Zipf’s law does not need a basis in economic theory to show a high degree of explanatory power; in other words, Zipf’s law is a statistical phenomenon. This explanation is supported by the fact that Zipf's law holds in many other cases such as firm size, web server domains, and in fields that range from economics to physics.

However, the striking observation of Zipf’s coefficient close to 1 remains a puzzle. Is it an economic regularity or a statistical phenomenon? This paper attempts to solve this puzzle. Using the data from Chinese cities (1985 and 1999) and the U.S urbanized areas data (1980, 1990 and 2000); this research investigates what drives the distribution of the estimated coefficient. We chose the US and China for two main reasons. One is because both countries have many cities. More cities allow us to do more rigorous statistical analysis. The other reason is because they have two very different economic systems. We want to see if the Zipf’s coefficient is sensitive to economic systems.

A better understanding of the Zipf's exponent is crucial. The validity of rank-size rule hinges on this exponent having a value close to one. If the value of this exponent is not really close to one this might put the fate of the rule in jeopardy and highlight the explanation offered by Gan et al. (2006) that the Zipf’s law is a statistical phenomenon, not an economic regularity.

The next section outlines the methodologies used in this analysis, including the rolling sample method and the random sampling method with replacement. The third section provides the results and cross country comparisons. The final section summarizes the empirical results and discusses their economic significance. Succinctly, this paper seeks to find the impact of sample size, truncation point and the mix of cities on the estimated exponent of the Zipf's law.

II. THE MODEL AND METHODOLOGY

Before setting up the model, it might be helpful to visualize the Zipf's law. Following the example from Gabaix (1999), we take a country such as U.S and order the cities with population: number 1 New York, number 2 Los Angeles and so on. We then draw a graph; on the y-axis we place the log of rank and on the x-axis we place the log of population. We see a straight line as shown below with a slope of close to -1.

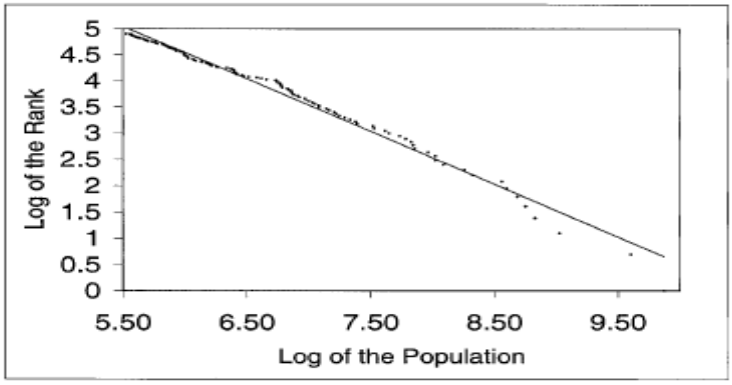


Figure 1: Log size versus log lank for 140 largest U.S cities in 1990.

The common way of estimating the slope of the line in the graph above is as follows:

$$\log(R_i) = \alpha - \beta \log(S_i) + \varepsilon \qquad [1]$$

where R_i is the rank of the ith city, S_i is the city's size and α is a constant term. Here we note that equation 1 is simply a log transformation of the regular Zipf's law which is normally expressed in the following form:

$$R_i = AS_i^{-\beta} \qquad [2]$$

Several studies have noted that estimating this expression of the Zipf's law yields an OLS bias through the standard errors. To correct for this bias, Gabaix and Ibragimov (2006) offered the following version that corrects for the bias, giving the standard errors of the exponent the form $(2/\bar{n}_i)^{0.5}\hat{\beta}$: where $\bar{n}_i$ is the corresponding sub-sample size. Instead of equation (1) they offer the following:

$$\log(R_i - 0.5) = \alpha - \beta \log(S_i) + \varepsilon \qquad [3]$$

This corrected version has come to be known as the rank minus half rule. Throughout this analysis, we will provide results from both versions and comment on the differences that exist between them.

The first method we will use is the rolling sample method. That is, we will estimate the exponent coefficient β using OLS, and in addition, we will repeat the estimation process using a moving truncation point. The start point of each sub-sample is fixed at the largest city and the truncation point moves down by one city every time, thereby increasing the sub-sample size by one each time. For example, the full sample size of U.S. urbanized areas for 1990 is 396. These urbanized areas will be ordered decreasingly from the largest urbanized area of New York (16,044,012 persons) to the smallest urbanized area of Brunswick, GA (50,066 person). The first sub-sample size of the regression [1] and [3] is $\bar{n}_1$, the 10 largest cities for example; then the second sub-sample is $\bar{n}_2 = \bar{n}_1 + 1$, the 11 largest cities, the third sub-sample size is $\bar{n}_3 = \bar{n}_2 + 1$, or

12 largest cities, and so on. We continue this process until the last sub-sample is equal to the full sample size of 396. The advantage of this methodology is that it captures the variation pattern of the exponent coefficients. Some studies have offered various theoretical explanations for the variation pattern of the exponent coefficients. Gabaix (1999a, b) suggested that the variation of the estimated power coefficients is due to bigger variance while Eeckhout (2004) thought that it is because of the underlying lognormal distribution. The rolling sample method captures the variation of the coefficient as both the sample size and truncation point changes. The results are explained fully in the next section.

The second method, random sampling with replacement, set out to separate some of the simultaneous effects captured with the rolling sample. The rolling sample does provide the gross variation in the estimated coefficient as three factors change simultaneously (sample size, truncation point and the variation in city sizes). To untangle these effects, we use our original data for each year as a pool to select from; this original data is not ordered in any fashion. We then randomly select the first sub-sample $\tilde{n}_1$, 5 random cities, and then rank them. The first sub-sample size of the regression [3] is $\tilde{n}_1$, the first 5 cities randomly selected. The second sub-sample $\tilde{n}_2$ is independent of the first sub-sample. However, it contains one more city than the first sub-sample, and the third sub-sample is also randomly selected and contains one more city than the second sub-sample, and so on. We continue this process until the last sub-sample is the same as the full sample size. Since this is a random process, we repeat the same process 100 times for each sample collecting the estimated coefficient. We average these series to get the distribution of the coefficient when we account for the size of the cities and the truncation point, the biases we expected under the rolling sample method.

The third method is to further test the effect of sample size on the distribution of the estimated coefficient. For this, we randomly generate 1000 numbers from a normal distribution. We then apply the random sampling technique and repeat the process we did above. After 100 iterations we average the series of $\hat{\beta}$'s. Finally, we examine the relationship between the average and sample size.

III. RESULTS

Table 1 shows the full-sample results of Zipf's law. As expected, all cities in both countries for all the years show high R^2-value (0.857 to 0.989). It is interesting to note that the estimated coefficients ($\hat{\beta}$'s) are slightly higher for the OLS bias corrected model than the original uncorrected version. Therefore, we can conclude that the uncorrected Zipf's law has a downward bias on the estimated coefficient.

Table 1: Regression results on Zipf's law using City Size Data from China and the U.S.

Nation	Year	$\hat{\beta}$	OLS Bias Corrected $\hat{\beta}$	R^2 (from unadjusted)	Sample Size
U.S.	1980	0.91	0.925	0.989	366
U.S.	1990	0.895	0.913	0.989	396
U.S.	2000	0.875	0.895	0.989	452
China	1985	0.856	0.875	0.857	324
China	1999	1.075	1.09	0.927	667

Data Sources: U.S Bureau of Census and Urban Statistical Yearbook of China (1986, 2000).

A. Rolling Sample Results for U.S. Urbanized Areas

The rolling sample results for samples from both countries show a negative relationship between the estimated coefficient and sample size. This implies that small samples of big cities yield higher coefficients ($\beta>1$) than large samples that also include smaller cities. These results are intuitive and hold true for most samples. When rank-size holds, the coefficient is close to 1, or the 95% confidence interval includes 1. This implies that the size of the second largest city is half the size of the first, the third largest city is one third the size of the first and so on. However, the results from the rolling sample size, especially from the small samples of large cities, suggests that the second largest city is greater than half of the first largest (Los Angeles has a population greater than half that of New York). This is what is implied by $\beta>1$. There is a truncation point and sample size bias in estimating the exponent. An explanation of this intuitive result is that most large cities enjoy the same economies of scale and have almost the same level of diversity, and productivity levels, and therefore do not have different population levels. As we include more cities into the sub-samples, we increase the variance: an explanation offered by Gabaix (1999a, b) for the variation of the estimated exponent. The differences between cities become significant as small cities significantly differ from other small cities in terms of the economies of scale they enjoy, the diversity they house, and productivity levels they have. This large difference between small cities contributes to the low estimated coefficients.

For U.S urbanized areas, we note that the rank-size rule only holds for certain sub-samples. For the 1980 cities, rank-size rule holds only for sub-samples between 180 and 205; for 1990, 140 to 195; and for year 2000, 140 to 205. These are the ranges where the 95% confidence interval includes $\beta=1$. This finding suggests that the rank-size rule (i.e., $\beta=1$) does not holds always hold.

Figure 2

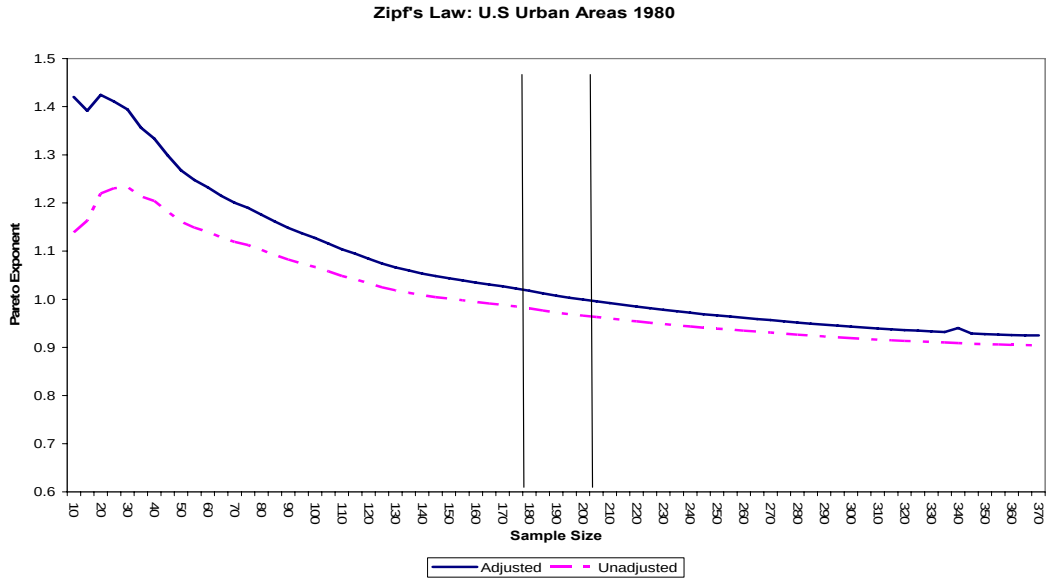


Figure 3

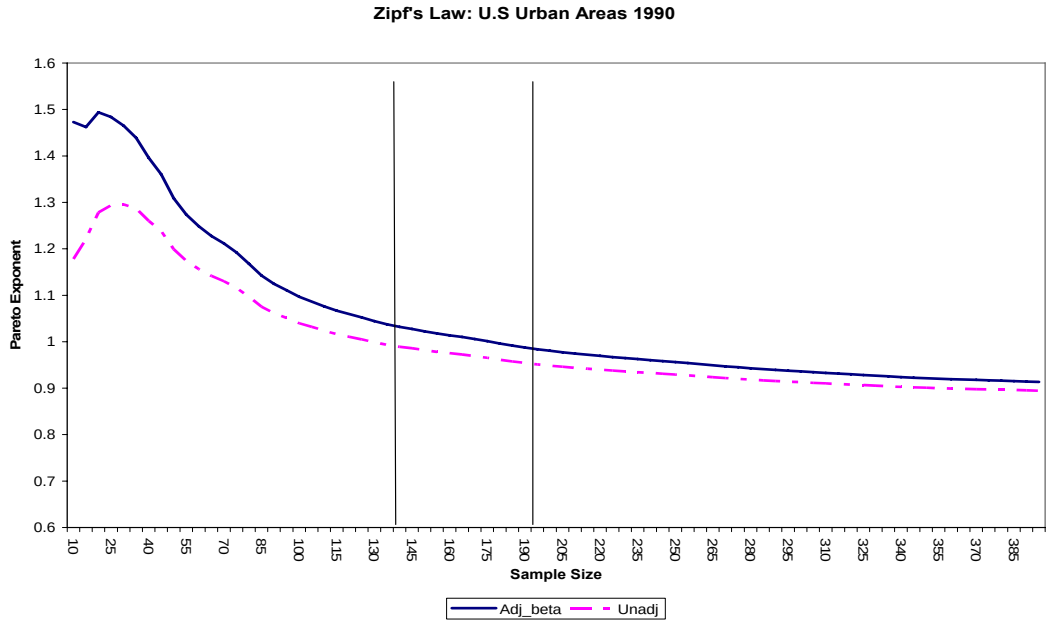
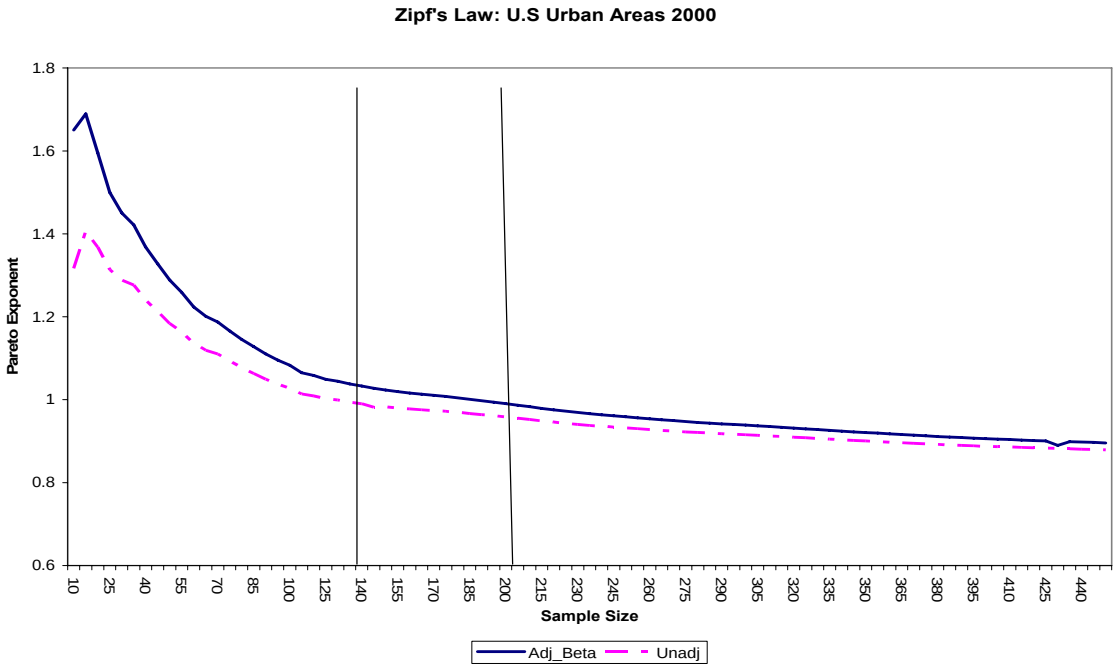


Figure 4



Figures 2-4 show that the unadjusted model has a downward bias in terms of the magnitude of the estimated exponent. For all the three years shown for U.S. urbanized areas, the estimated exponent seems to be following a lognormal distribution when the rolling sample technique is used.

B. Rolling Sample Results for Chinese Cities

The two graphs below show the distribution of Zipf coefficient for the Chinese data using the rolling sample approach. Like the U.S. case, we also found a negative relationship between the estimated coefficient and the sample size. For the 1985 cities, the rank-size rule holds only for sub-samples between 315 and 320 cities. Therefore, the rank-size rule does not hold for a sample of the largest 314 cities in China for 1985. Because of China's rapid urbanization and city reclassification, the number of Chinese cities was more than doubled between 1985 and 1999, from 324 to 667. For Chinese cities in 1999, the rank-size rule does not hold for any sub-

sample. All the estimated Zipf exponents using the rolling sample approach are significantly greater than one.

In the graphs below, we show the gap between Gabaix and Ibragimov (2006) adjusted results from the original biased model. The unadjusted model has a downward bias on the estimated coefficient. We also notice that the estimated coefficients are generally higher for Chinese cities than U.S. urbanized areas. This difference might be attributed to the differences in definitions of the 'city or urbanized area' in the two countries and the economic systems.

Figure 5

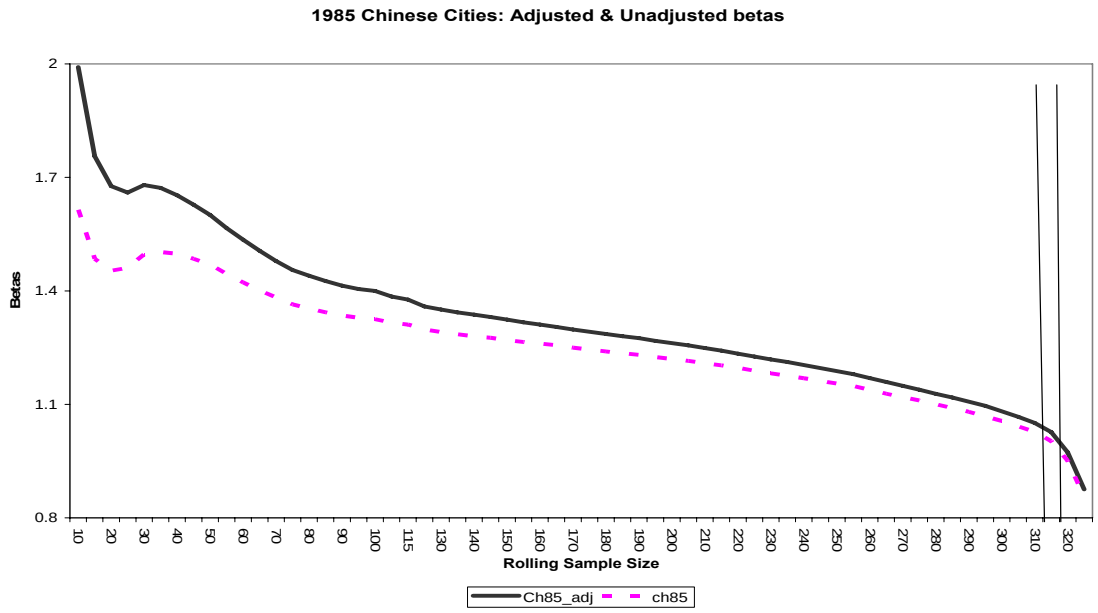
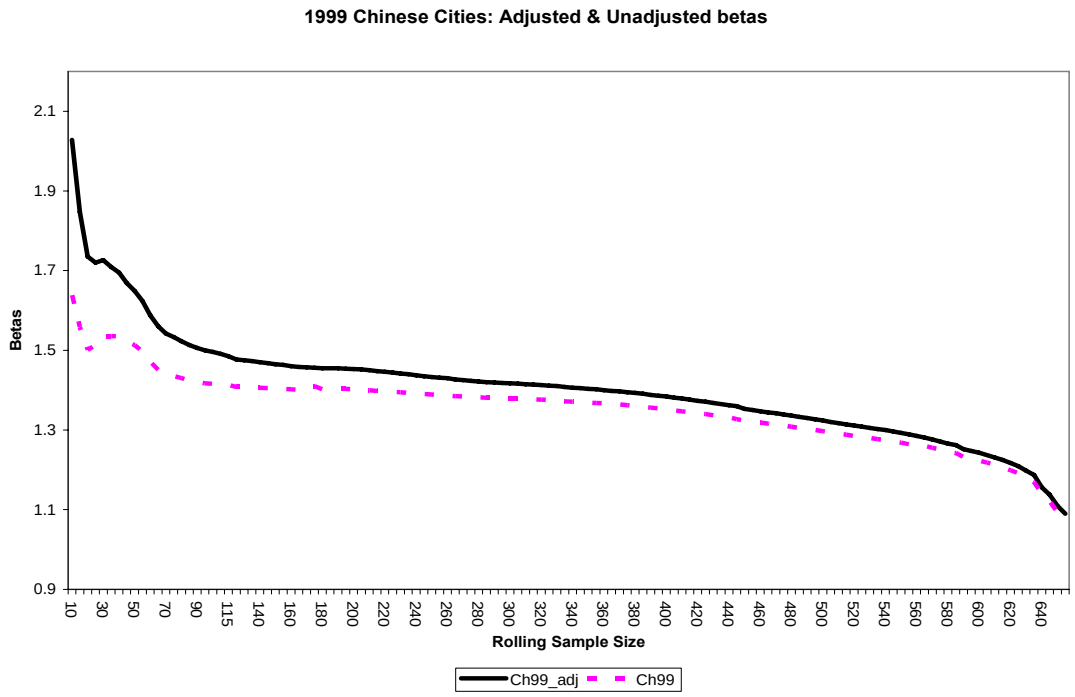


Figure 6

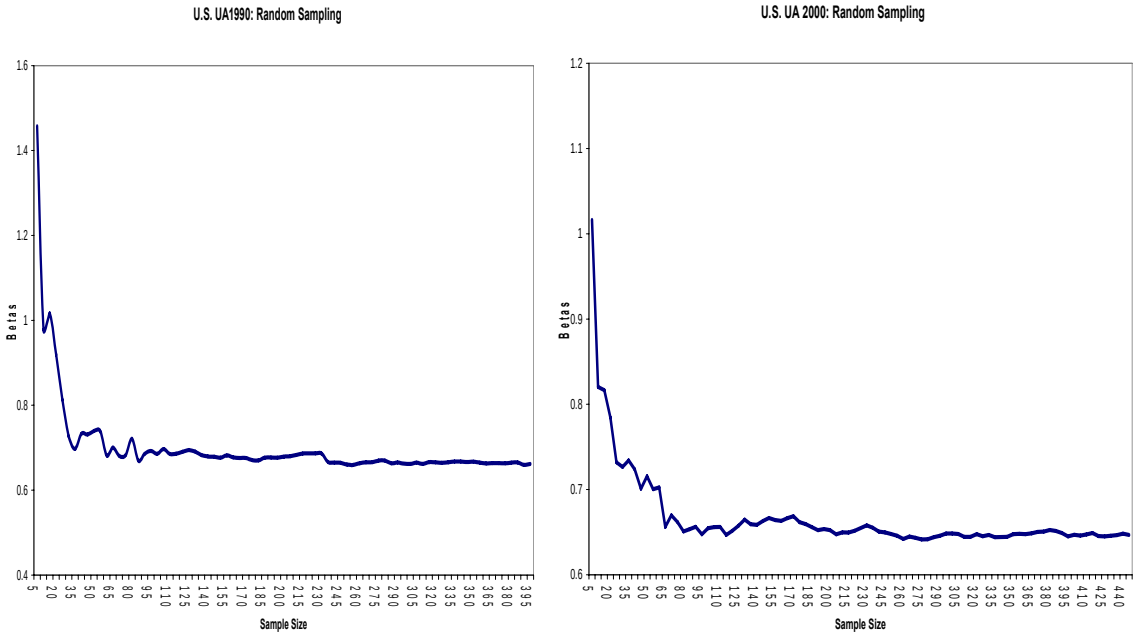


C. Results from Random Sampling

As we discussed in the methodology section, there is a dilemma with the rolling sample technique in our analysis. It captures the truncation point effect, sample size effect, and also the assortment of cities in the sample. In the rolling sample method, the assortment is biased since it starts with largest cities and increases the sub-sample systematically. Using the random sampling with replacement technique while increasing the sub-sample, we captured an assortment of cities that can include all sizes from the beginning. This eliminates the upward bias due to large cities in the first sub-samples. By randomly sampling each time, the truncation point also randomly changes. This eliminates the systematically changing truncation point bias inherent in the rolling sample technique.

As shown in the graphs below, sample size alone has a small upward bias mainly for sub-samples below 100. For sample sizes greater than 100, the effect of sample size disappears as we increase the sample size (that is, the estimated coefficient stays almost constant).

Figure 7



D. Simulation Results Using Random Sampling with Replacement

To further test the effect of sample size on the distribution of the estimated coefficient, we randomly generate 1000 numbers from a normal distribution. We then apply the random sampling technique and repeat the process we did above. After 100 iterations we average the series of $\hat{\beta}$'s and the average is shown in the graph below. Surprisingly, we still capture the effect of very small sample sizes below 100.

Figure 8

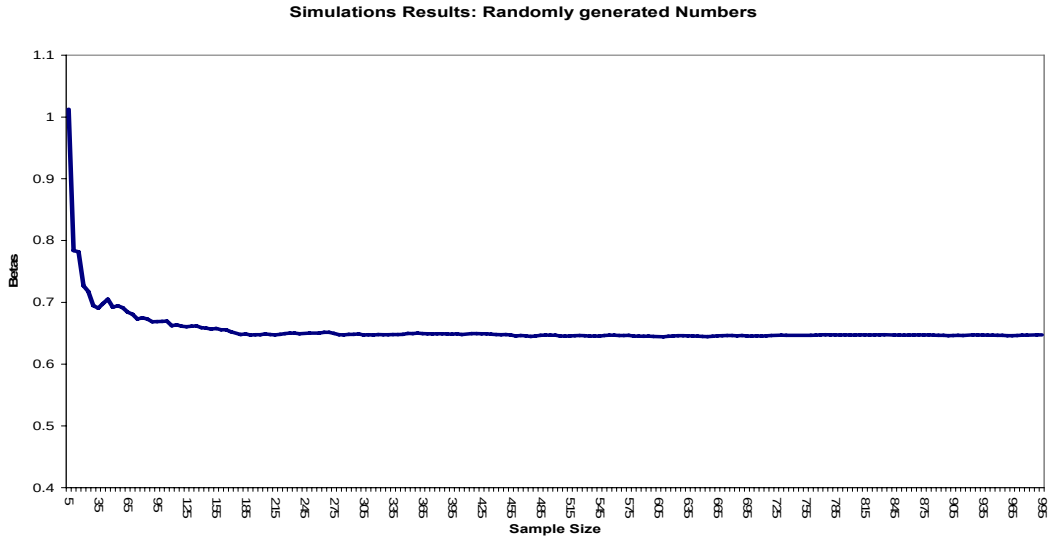


Figure 8 confirms the upward bias of samples less than 100. Therefore, when working with sample sizes greater than 100, the sample size does not influence the value of the estimated coefficient. This result confirms the conclusion reached by Gabaix (1999a, b).

Interestingly, the graphs, especially for the U.S samples, suggest a lognormal distribution. Therefore, we run a regression to check the relationship between the estimated exponent ($\hat{\beta}$) and the sample size (SS). For this analysis we run the following equation and present the results in Table 2.

$$\log(\hat{\beta}_i) = \alpha - \delta \log(SS_i) + \varepsilon \qquad [4]$$

Table 2: The relationship between the log of the estimated Pareto exponent and the log of the sample size

Nation	Year	$\hat{\delta}$	OLS Bias Corrected $\hat{\delta}$	R^2 (from adjusted)	Number of observations
U.S.	1980	-0.10***	-0.15***	0.98	355
U.S.	1990	-0.11***	-0.16***	0.96	385
U.S.	2000	-0.13***	-0.17***	0.97	441
Random	--	--	-0.03***	0.57	1000

***: significant at 1%

The number of observations is the number of estimated exponents ($\hat{\beta}$'s) obtained using the rolling sample method. For the United States, the lognormal regression (equation 4) yields a very high R^2-value, 0.96 or higher. In 1980, a one percent increase in the number of urban areas led to a 0.15 percent reduction in the value of the estimated exponent (for the adjusted model). The unadjusted model shows a smaller percentage decrease in the value of the estimated exponent as the sample size increase; this explains why we have the unadjusted model converging with the adjusted model in figures 2-4.

These results are important. If the value of estimated exponent is greatly impacted by sample size, we cannot expect the value of this exponent to be close to one in all cases. Therefore, the validity of the rank-size rule largely depends on the sample size used in a study. In other words, the rank-size is not an economic regularity but a statistical phenomenon.

IV. CONCLUSIONS

This paper has examined the validity of the rank-size rule based on estimated Zipf exponent. Using the rolling sample technique, we proved that small samples with large cities only tend to generate high values of the estimated coefficient compared to samples dominated with small cities. We showed that the rank-size rule holds only for some selected sub-samples. For the U.S. samples, the estimated coefficient is close to one between 180 and 205 cities for 1980 data, between 140 and 195 for 1990 data, and between 140 and 205 for 2000. For the Chinese cities, the estimated coefficient is close to one only for sub-samples that contain between 315 and 320 cities for the 1985 data; it is never close to one for the 1999 data. This finding raises questions on the general application of the rank-size rule. From the random sampling technique we concluded that small samples in general produce higher value of the estimated coefficient.

The double log regression model of estimated exponents and sample size yielded a very high R^2-value, 0.96 or greater. It also produced an elasticity of the estimated exponent with respect to sample size. If US urbanized areas are used, our results indicated that a one percent increase in the sample size led to about 0.15 percent reduction in the value of the estimated exponent (for the adjusted model). Therefore, we conclude that the Zipf exponent depends on the sample size used in a study and the rank-size rule does not hold in general. In other words, the rank-size is not an economic regularity but a statistical phenomenon. If the rank-size rule could be derived from the Gibrat's law, our conclusion implies that the Gibrat's law in the city size distribution does not hold either. Thus, the urban growth process is not independent of city size.

REFERENCES:

Eeckhout, J., 2004. Gibrat's law for (all) cities. American Economic Review 94, 1429-1451.

Gabaix, X., 1999a. Zipf's law for cities: an explanation. Quarterly Journal of Economics CXIV (3), 739—767.

Gabaix, X., 1999b. Zipf's law and the growth of cities. American Economic Review, Vol. 89 (2), 129—132.

Gabaix, X., Ibragimov, R., 2006. Rank – ½: A simple way to improve the OLS estimation of tail exponents. Working Paper.

Gabaix, X., Ioannides, Y.M., 2004. The evolution of city size distributions. In Vernon Henderson, J., Thisse, J.F., (Eds.), Handbook of Regional and Urban Economics, vol. 4. North Holland, Amsterdam, pp. 2341—2378. Chapter 53.

Gan, L., Li, D., and Song, S., 2006. Is the Zipf's law spurious in explaining city-size distributions? Economic Letters 92, 256—262.

Guerin-Pace, F., 1995. Rank-size distribution and the process of urban growth. Urban Studies 32 (3), 551—562.

Krugman, K., 1995. Development, Geography, and Economic Theory. The MIT Press, Cambridge, MA.

Li, W., Yang, Y. 2002. Zipf's law in importance of genes for cancer classification using microarray data. Theoretical Biology 219 (4), 539-551.

Mills, E.S., Hamilton, B.W., 1994. Urban Economics, 5th ed. Harper Collins College Publishers, New York.

Nitsche, V., 2005. "Zipf Zipped." Journal of Urban Economics 57, pp. 86-100.

Rosen, K., Resnick, M., 1980. The size distribution of cities: An explanation of the Pareto law and primacy. Journal of Urban Economics 8, pp. 165-186.

Shiode, N. Batty, M., 2000. Power Law Distribution in Real and Virtual Worlds (http:www.isoc.org/inet2000/cdproceedings/2a/2a_2.htm, also presented at INET 2000, Yokohama, Japan, 2000).

Sinclair, R. 2001. Examining the Growth Model's Implications: The World Income Distribution, Working paper. Department of Economics, Syracuse University.

Song, S., Zhang, K.H., 2002. Urbanization and city size distribution in China. Urban Studies 39 (12), 2317—2327.

Tachimori, Y., Takashi, T. 2002. Clinical diagnoses following Zipf's Law. Fractals 10 (3), 341-351.

Zipf, G., 1949. Human behavior and the principle of last effort. Cambridge, MA: Addison Wesley Press.