A Nonparametric Estimation of the Local Zipf Exponent for all US Cities

González-Val, Rafael

Universitat de Barcelona Institut d'Economia de Barcelona (IEB)

15 November 2010

A Nonparametric Estimation of the Local Zipf Exponent for all US Cities

Rafael González-Val

Universitat de Barcelona & Institut d'Economia de Barcelona (IEB)

Abstract: In this short paper we apply the methodology proposed by Ioannides and Overman (2003) to estimate a local Zipf exponent using data for the entire twentieth century of the complete distribution of cities (incorporated places) without any size restrictions in the US. The results reject Zipf's Law from a long term perspective, as the estimated values are close to zero. However, decade by decade we find evidence in favour of Zipf's Law. We also see how periods in which the Zipf exponent grows with city size are interspersed with others in which the relationship between the exponent and city shares is negative.

Keywords: Zipf's Law; Gibrat's Law; urban growth.

JEL: R00; C14.

Address: Rafael González-Val,

Departamento de Economía Política y Hacienda Pública, Universitat de Barcelona. Facultat d'Economia i Empresa, Av. Diagonal 690, 08034 Barcelona (Spain).

E-mail: r.gonzalez-v@ub.edu

1. Introduction

City size distribution has been the subject of numerous empirical investigations by urban economists, statistical physicists, and urban geographers. One of the stylised facts in Urban Economics is that the city size distribution in many countries can be approximated by a Pareto distribution whose exponent is equal to one. If this is the case, it can be concluded that there is evidence for Zipf's Law[1] (Zipf, 1949) and means that, ordered from largest to smallest, the size of the second city is half that of the first, the size of the third is a third of the first, and so on. Another well-known stylised fact is Gibrat's Law or the Law of Proportionate Growth (Gibrat, 1931), which establishes that the growth rate of a variable is independent of its initial size. Both are considered to be two sides of the same coin. While Gibrat's Law has to do with the population growth process, Zipf's Law refers to its resulting population distribution.

These are extensively studied empirical regularities in many countries, especially in the United States (US); see Ioannides and Overman (2003), Black and Henderson (2003), Eeckhout (2004), or González-Val (2010). Ioannides and Overman (2003) propose a nonparametric procedure to estimate Gibrat's Law for city growth processes as a time-varying geometric Brownian motion and to calculate local Zipf exponents from the mean and variance of city growth rates. They use data from metropolitan areas from 1900 to 1990 (112 to 334 metropolitan areas) and arrive at the conclusion that Gibrat's law holds in the urban growth processes and that Zipf's law is also fulfilled approximately for a wide range of city sizes. Nevertheless, Black and Henderson (2003) arrive at different conclusions for the same period (probably because they use different metropolitan areas; their data goes from 194 metropolitan areas in 1900 to 282 in 1990). Zipf's law holds only for cities in the upper third of the distribution, while Gibrat's law would be rejected for any sample size. These results highlight the extreme sensitivity of conclusions to the geographical unit chosen and to sample size.

Finally, Eeckhout (2004) demonstrates the statistical importance of considering the whole sample, not only the larger cities[2]. The estimated Pareto parameter depends on the truncation point, so when all the cities are considered for the period 1990 to 2000, the empirical city size distribution follows a lognormal rather than a Pareto distribution, and the value of Zipf's parameter is not 1, as earlier works concluded, but is slightly above $1/2$; Gibrat's law holds for the entire sample. Recently, González-Val (2010) generalizes this analysis for all of the twentieth century, extracting long term conclusions: Gibrat's law holds (weakly; growth is proportionate on average but not in variance, as the smallest cities present a clearly higher variance) and Zipf's law holds only if the sample is sufficiently restricted to the top, not for a larger sample, because city size distribution follows a lognormal when we consider all cities with no size restriction.

The nonparametric procedure put forward by Ioannides and Overman (2003) is especially relevant because it is based on the statistical explanation of Zipf's Law for cities offered by Gabaix (1999). Gabaix presents a model based on local random amenity shocks, independent and identically distributed, which through migrations

[1] Although Auerbach had previously observed in 1913 the Pareto pattern of city size distribution.

[2] In the US, to qualify as a metropolitan area a city needs to have 50,000 or more inhabitants, or the presence of an urbanised area of at least 50,000 inhabitants, and a total metropolitan population of at least 100,000 (75,000 in New England), according to the OMB definition. Therefore, data from metropolitan areas impose an implicit truncation point.

between cities generate Zipf's Law. The main contribution of the work is to justify the fulfilment of Zipf's Law in that cities in the upper tail of the distribution follow similar growth processes, so that the fulfilment of Gibrat's Law involves Zipf's Law.

In this paper, the methodology proposed by Ioannides and Overman (2003) to estimate a local Zipf exponent is applied to a new dataset covering the complete distribution of cities in the US (understood as incorporated places) without any size restrictions, for the entire twentieth century. Section 2 presents the dataset and summarises the nonparametric procedure and its statistical foundations. Section 3 offers the results, and Section 4 concludes.

2. Data and Methodology

We use the same data set as González-Val (2010). Table 1 presents the number of cities for each decade, and the descriptive statistics. Our base, created from the original documents of the annual census published by the US Census Bureau, consists of the available data of all incorporated places without any size restriction, for each decade of the twentieth century (decennial data from 1900 to 2000). The US Census Bureau uses the generic term incorporated place to refer to the governmental unit incorporated under state law as a city, town (except in the states of New England, New York and Wisconsin), borough (except in Alaska and New York), or village, and which has legally established limits, powers and functions.

Two details should be noted[3]. First, Alaska, Hawaii, and Puerto Rico have not been considered due to data limitations. And second, for the same reason we also exclude all the unincorporated places (concentrations of population which do not form part of any incorporated place, but which are locally identified with a name), which began to be taken into account after 1950. However, these settlements did exist earlier, so that their inclusion would again present a problem of inconsistency in the sample. Also, their elimination is not quantitatively important; in fact there were 1,430 unincorporated places in 1950, representing 2.36% of the total population of the US, which by 2000 would be 5,366 places and 11.27%.

The empirical strategy commonly used to test Zipf's Law consists in the estimation of log linear regressions of city size (population, P) against rank (R):

$$\log R(P) = \log A - \zeta \log P, \qquad (1)$$

where A and ζ are parameters. Zipf's Law is an empirical regularity, which appears when the Pareto exponent is equal to unity, $\zeta = 1$ (see the surveys of Cheshire, 1999, and Gabaix and Ioannides, 2004, for further explanation). Results are usually presented in double logarithmic graphs of rank compared to population, named Zipf plots, which are used extensively in the specialised literature.

However, this approach has pitfalls, highlighted in the recent literature, and different estimators have been proposed. Gabaix and Ioannides (2004) show that the Hill (Maximum Likelihood) Estimator is more efficient if the underlying stochastic process is really a Pareto distribution, but when the size distribution of cities does not follow a Pareto distribution the Hill estimator may be biased (Soo, 2005). At the same time, the OLS estimate also has some problems, see Goldstein et al. (2004) and Nishiyama et al. (2008). Finally, Gabaix and Ibragimov (2007) proposed subtracting $\frac{1}{2}$

[3] More details about data sources and definitions are given in González-Val (2010).

from the rank to obtain an unbiased estimation of the Pareto exponent using an OLS regression.

In this paper we apply the nonparametric procedure put forward by Ioannides and Overman (2003). This is a completely different empirical strategy, relying on the statistical foundation of Zipf's Law offered by Gabaix (1999). The exposition follows closely Ioannides and Overman (2003); see also Gabaix (1999) and Gabaix and Ioannides (2004) for more details[4].

Let S_i denote the normalised size of city i, that is, the population of city i divided by the total urban population. Following Gabaix (1999), city sizes are said to satisfy Zipf's Law if the countercumulative distribution function, $G(S)$, of normalised city sizes, S, tends to

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

where a is a positive constant and $\zeta=1$. If Gibrat's Law holds for city growth processes, cities grow randomly, with the same expected growth rate and the same standard deviation: then the limit distribution will converge to $G(S)$, given by Eq. (2)[5].

Gabaix also considers the case where cities grow randomly with expected growth rates and standard deviations that depend on their sizes (a weak Gibrat's Law). That is, the size of city i at time t varies according to:

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

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,

$$\zeta(S)=-\frac{S}{p(S)}\cdot\frac{dp(S)}{dS},$$

where $p(S)$ denotes the invariant distribution of S. Starting from the forward Kolmogorov equation associated with Eq. (3), the local Zipf exponent, associated with the limit distribution, can be derived and is given by

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

where $\mu(S)$ is relative to the overall mean for all city sizes. Eq. (4) identifies two possible causes of deviations from Zipf's Law: the means and the standard deviations. If $\zeta<1$ then the distribution has neither finite mean nor finite variance, and if $1<\zeta<2$ it has finite mean but not finite variance.

3. Results

First, we use kernel regression techniques that establish a functional form-free relationship between the mean and the variance of city growth rates and city size for the

[4] Eqs. (3) and (4) replicate, respectively, Eq. (11), p. 756, and Eq. (13), p. 757, in Gabaix (1999).

[5] See Gabaix (1999), p. 744.

entire distribution. This allows us to test whether Gibrat's Law holds. Second, we use Eq. (4) to directly estimate the local Zipf exponents.

In order to analyse the entire twentieth century, all the growth rates are taken between consecutive periods. There are 162,698 population-growth rate pairs in that pool. City size is defined as the normalised size of the city (S), that is, the population of city divided by the total urban population[6], and the growth rate $\mu(S)$ is defined as the difference between each city's growth rate and the contemporary average growth rate, as in Ioannides and Overman (2003). To calculate the conditional mean and variance on city size, we apply the Nadaraya-Watson method[7], exactly as it appears in Härdle (1990). The estimator is very sensitive, both in mean and in variance, to atypical values. Thus, we decided to eliminate the 5% of the smallest distribution observations for each decade, as they are characterised by very high dispersion in mean and in variance, and they distort the results. Therefore, the sample size is reduced to 154,563 observations. Finally, we also eliminate 1,079 observations with a growth rate $\mu(S)$ greater than two. The reason is that we cannot control for change in city boundaries; there are more than twenty thousand different cities in the sample, and information on boundaries is only available for the largest cities in some decades. Then, we decide to eliminate cities with the greatest growth rates to control the most extreme cases, relying on the huge sample size to make the spurious growth produced by change in boundaries irrelevant. Final sample size is 153,484 observations (94.34% of total sample).

Figure 1 shows the nonparametric estimates for the entire twentieth century of the mean growth rate and variance of growth rate conditional on city size, and the local Zipf exponent calculated applying Eq. (4). Graphs also displays bootstrapped 95% confidence bands, calculated using 500 random samples with replacement. Results are shown until city sizes of 0.01; the reason is that there is one technical problem with this procedure, the sparsity of data at the upper tail of the distribution, which produces extreme values of the estimations. This means that, as Ioannides and Overman (2003), we exclude the 76 observations corresponding to the largest cities with shares greater than 0.01.

Results show a very slight increasing behaviour of city growth (observe the very small scale of growth graph), and a negative relationship between variance and city size (although the differences are not significant). Thus, small cities exhibit lower growth rates and higher variances than larger cities, indicating that Gibrat's Law does not hold exactly. Part of this could be explained by the appearance of new cities that enter with small sizes[8]. Moreover, the local estimate of the Zipf exponent is also decreasing with city size, mainly as a consequence of the decreasing variance. Results reject Zipf's Law from this long term perspective, as the estimated values are close to zero.

The variation in the estimates is very small, as a consequence of the huge sample size of the pool. Moreover, most of the observations are concentrated at the lower end of the distribution. So, we have repeated the exercise for each decade, with lower sample sizes. One advantage is that the influence of new entrant cities is lower from one decade to another than in the whole twentieth century. Also, short term estimations could reveal interesting behaviours.

[6] US urban population data from 1900 to 1990 come from Table 1 in Overman and Ioannides (2001). Data for the year 2000 is taken from US Census Bureau (http://www.census.gov).

[7] We use an Epanechnikov kernel and Silverman's kernel bandwidth.

[8] See González-Val (2010) for an analysis of new entrants.

Figure 2 shows the results for four representative decades[9]: 1900-1910 (9,892 observations), 1930-1940 (15,367 observations), 1970-1980 (17,277 observations) and 1980-1990 (17,808 observations). We can observe that decade by decade the estimates of the Zipf exponent are greater than when considering all the twentieth century, probably as a consequence of the lower sample size. In this case we find evidence in favour of Zipf's Law as value one falls within the confidence bands for most of the distribution in most of the decades. We also see how periods in which the Zipf exponent grows with city size are interspersed with others in which relationship between the exponent and city shares is negative. As the variance of growth rate presents a decreasing pattern for almost all decades, the differentiated behaviours of local exponents must be a consequence of differences in the mean growth rates.

4. Conclusions

In this paper, the methodology proposed by Ioannides and Overman (2003) to estimate a local Zipf exponent is applied to a new dataset covering the complete distribution of cities in the US (understood as incorporated places) without any size restrictions for all the twentieth century.

Results reject Zipf's Law from a long term perspective, as the estimated values are close to zero. We also find that small cities exhibit lower growth rates and higher variance than larger cities, indicating that Gibrat's Law does not hold exactly. Part of this could be explained by the appearance of new cities that enter with small sizes. In the short term, decade by decade, we find evidence in favour of Zipf's Law for most of the distribution in most of the decades. We also observe differentiated behaviours: periods in which the Zipf exponent grows with city size are interspersed with others in which relationship between the exponent and city shares is negative.

References

[1] Black, D., and V. Henderson, (2003). Urban evolution in the USA. Journal of Economic Geography, 3(4): 343–372.

[2] Cheshire, P., (1999). Trends in sizes and structure of urban areas. In: Handbook of Regional and Urban Economics, Vol. 3, P. Cheshire and E. S. Mills, eds. Amsterdam: Elsevier, Chap. 35, 1339–1373.

[3] Eeckhout, J., (2004). Gibrat's Law for (All) Cities. American Economic Review, 94(5): 1429–1451.

[4] Gabaix, X., (1999). Zipf's law for cities: An explanation. Quarterly Journal of Economics, 114(3): 739–767.

[5] Gabaix, X., and R. Ibragimov, (2007). Rank-1/2: a simple way to improve OLS estimation of tail exponents. NBER technical working paper, vol. 342.

[6] Gabaix, X., and Y. M. Ioannides, (2004). The evolution of city size distributions. In: Handbook of urban and regional economics, Vol. 4, J. V. Henderson and J. F. Thisse, eds. Amsterdam: Elsevier, 2341–2378.

[7] Gibrat, R., (1931). Les Inégalités Économiques. París: Librairie du Recueil Sirey.

[9] Again, we exclude the observations with shares greater than 0.01: 11 observations in 1900-1910, 10 in 1930-1940, 5 in 1970-1980 and 3 in 1980-1990.

[8] Goldstein, M. L., S. A. Morris and G. G. Yen, (2004). Problems with fitting to the Power-law distribution. The European Physical Journal B – Condensed Matter, 41(2): 255–258.

[9] González-Val, R., (2010). The Evolution of the US City Size Distribution from a Long-run Perspective (1900–2000). Forthcoming in Journal of Regional Science. DOI 10.1111/j.1467-9787.2010.00685.x

[10] Härdle, W., (1990). Applied nonparametric regression. Cambridge,: Cambridge Univ. Press.

[11] Ioannides, Y. M., and H. G. Overman, (2003). Zipf's Law for Cities: an Empirical Examination. Regional Science and Urban Economics, 33: 127–137.

[12] Nishiyama, Y., S. Osada and Y. Sato, (2008). OLS estimation and the t test revisited in rank-size rule regression. Journal of Regional Science, 48(4): 691–715.

[13] 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.

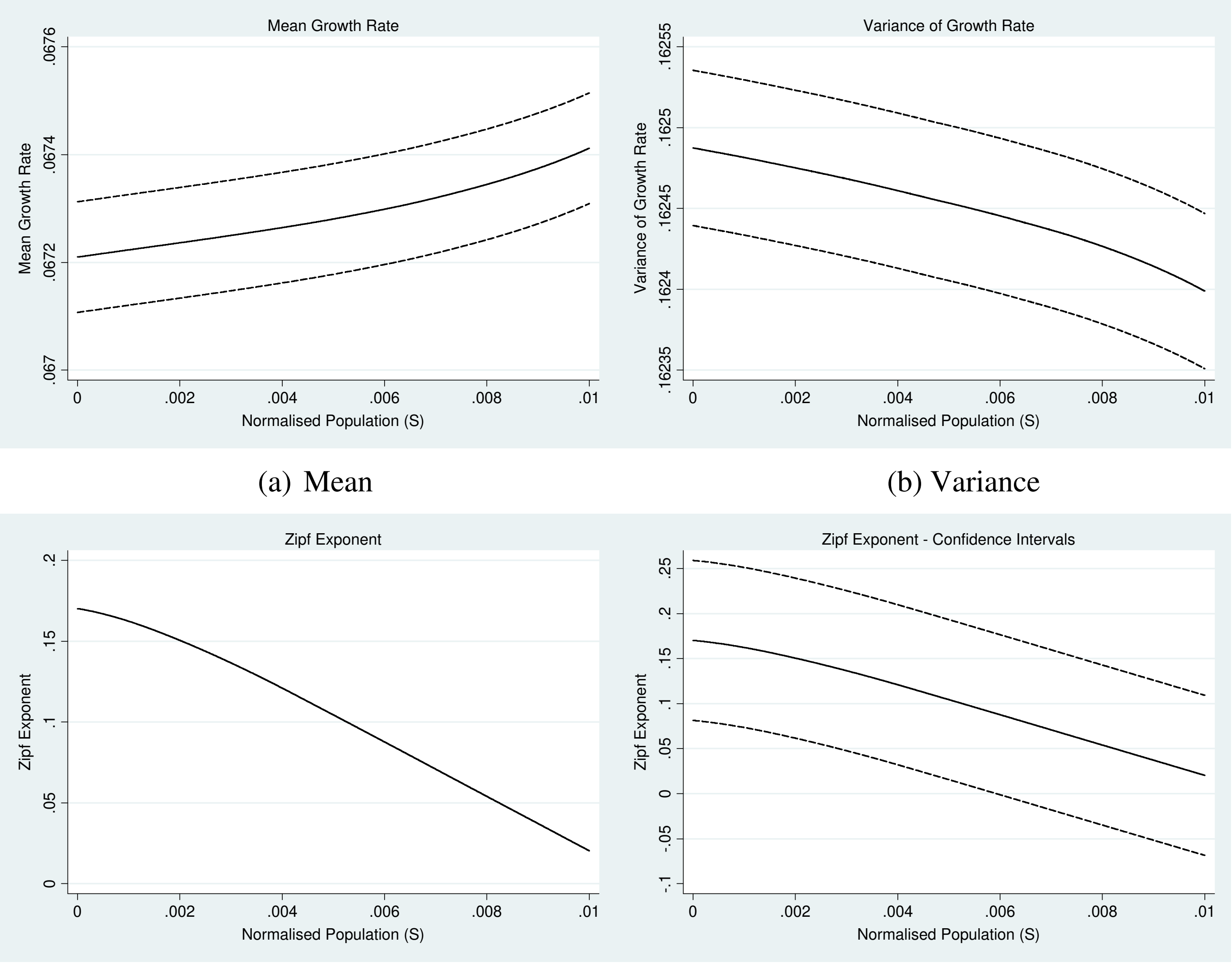
[14] Soo, K. T., (2005). Zipf's Law for cities: a cross-country investigation. Regional Science and Urban Economics, 35: 239–263.

[15] Zipf, G., (1949). Human Behaviour and the Principle of Least Effort. Cambridge, MA: Addison-Wesley.

Table 1. Number of cities and descriptive statistics

Year	Cities	Mean	Standard deviation	Minimum	Maximum
1900	10,596	3,376.04	42,323.90	7	3,437,202
1910	14,135	3,560.92	49,351.24	4	4,766,883
1920	15,481	4,014.81	56,781.65	3	5,620,048
1930	16,475	4,642.02	67,853.65	1	6,930,446
1940	16,729	4,975.67	71,299.37	1	7,454,995
1950	17,113	5,613.42	76,064.40	1	7,891,957
1960	18,051	6,408.75	74,737.62	1	7,781,984
1970	18,488	7,094.29	75,319.59	3	7,894,862
1980	18,923	7,395.64	69,167.91	2	7,071,639
1990	19,120	7,977.63	71,873.91	2	7,322,564
2000	19,296	8,968.44	78,014.75	1	8,008,278

Figure 1. Nonparametric estimates for all the twentieth century (a pool of 153,484 observations)



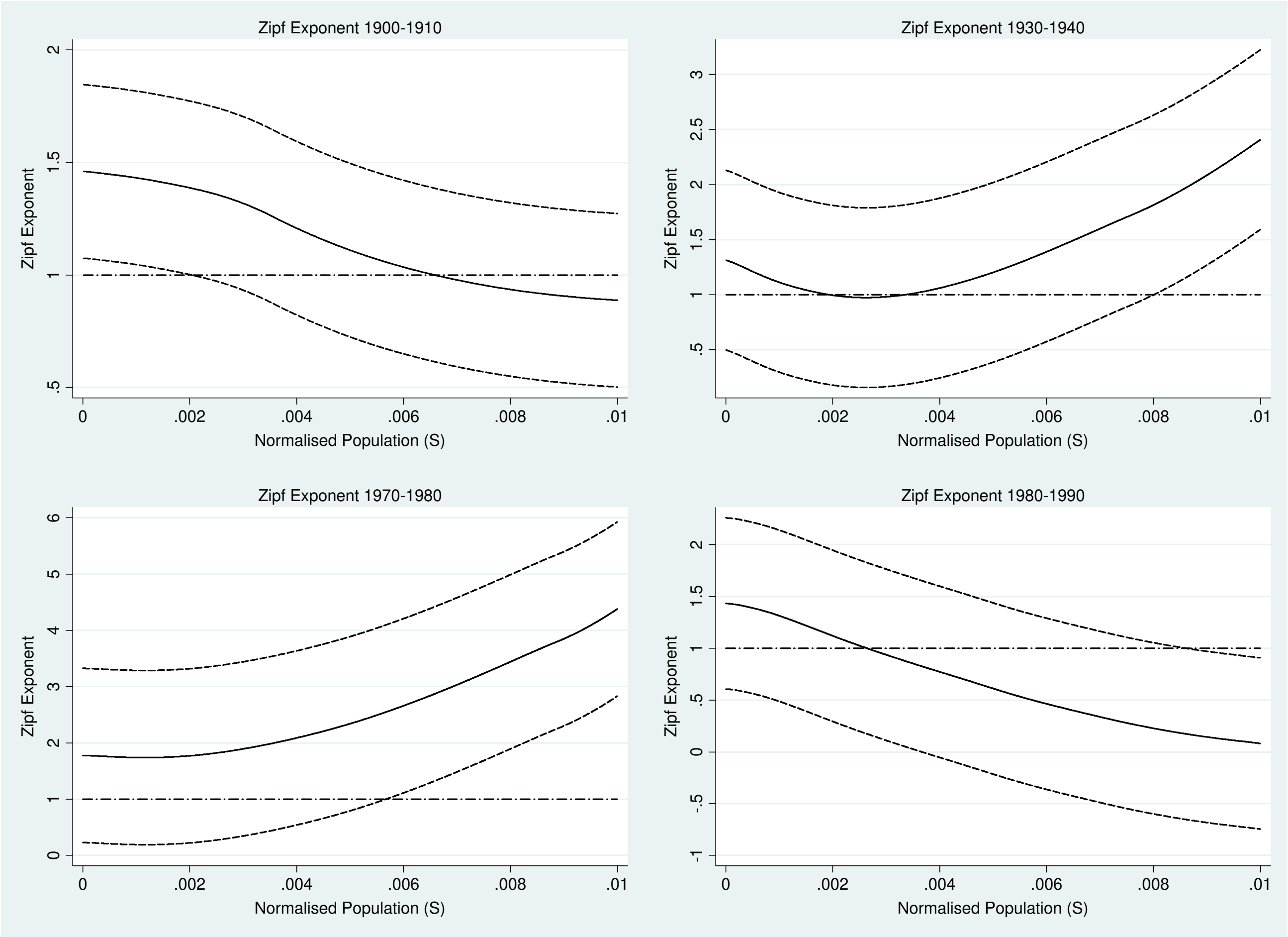
(a) Mean

(b) Variance

(c) Zipf

(d) Zipf (confidence bands)

Figure 2. Nonparametric estimates by decade


Zipf Exponent 1900-1910
Zipf Exponent
Normalised Population (S)
Zipf Exponent 1930-1940
Zipf Exponent
Normalised Population (S)
Zipf Exponent 1970-1980
Zipf Exponent
Normalised Population (S)
Zipf Exponent 1980-1990
Zipf Exponent
Normalised Population (S)