Available online at www.sciencedirect.com
SCIENCE DIRECT®

Regional Science and Urban Economics 35 (2005) 239–263
www.elsevier.com/locate/econbase

Zipf's Law for cities: a cross-country investigation

Kwok Tong Soo*

Centre for Economic Performance, London School of Economics, Houghton Street, London WC2A 2AE, UK

Received 22 March 2004; accepted 14 April 2004
Available online 17 July 2004

Abstract

This paper assesses the empirical validity of Zipf's Law for cities, using new data on 73 countries and two estimation methods—OLS and the Hill estimator. With either estimator, we reject Zipf's Law far more often than we would expect based on random chance; for 53 out of 73 countries using OLS, and for 30 out of 73 countries using the Hill estimator. The OLS estimates of the Pareto exponent are roughly normally distributed, but those of the Hill estimator are bimodal. Variations in the value of the Pareto exponent are better explained by political economy variables than by economic geography variables.

JEL classification: C16; R12
Keywords: Cities; Zipf's Law; Pareto distribution; Hill estimator

1. Introduction

One of the most striking regularities in the location of economic activity is how much of it is concentrated in cities. Since cities come in different sizes, one enduring line of research has been in describing the size distribution of cities within an urban system.

The idea that the size distribution of cities in a country can be approximated by a Pareto distribution has fascinated social scientists ever since Auerbach (1913) first proposed it. Over the years, Auerbach's basic proposition has been refined by many others, most notably Zipf (1949), hence the term "Zipf's Law" is frequently used to refer to the idea that city sizes follow a Pareto distribution. Zipf's Law states that not only does the size

* Tel.: +44-207-955-7080.
E-mail address: k.t.soo@lse.ac.uk (K.T. Soo).

0166-0462/$ - see front matter
doi:10.1016/j.regsciurbeco.2004.04.004

distribution of cities follow a Pareto distribution, but that the distribution has a shape parameter (henceforth the Pareto exponent) equal to 1.[1]

The motivation for this paper comes from several recent papers,[2] which seek to provide theoretical explanations for the "empirical fact" that Zipf's Law holds in general across countries. The evidence they present for the existence of this fact comes in the form of appeals to past work such as Rosen and Resnick (1980), or some regressions on a small sample of countries (mainly the US). One limitation of such appeals to the Rosen and Resnick result is that their paper is over 20 years old, and is based on data that dates from 1970. Thus, one pressing need is for newer evidence on whether Zipf's Law continues to hold for a fairly large sample of countries.

The present paper sets out to do four things: the first is to test Zipf's Law, using a new data set that includes a larger sample of countries. The second is to perform the analysis using the Hill estimator suggested by Gabaix and Ioannides (in press), who show that the OLS estimator is downward biased when estimating the Zipf regression, and that the Hill estimator is the maximum likelihood estimator if the size distribution of cities follows a Pareto distribution. Third, it nonparametrically analyses the distribution of the Pareto exponent to give an indication of its shape and to yield additional insights. Finally, this paper sets out to explore the relationship between variation in the Pareto exponent, and some variables motivated by economic theory.

Compared to Rosen and Resnick (1980), we find, first, that when we use OLS, for cities, Zipf's Law fails for the majority of countries. The size distribution often does not follow a Pareto distribution, and even when it does, the Pareto exponent is frequently statistically different from 1, with over half the countries exhibiting values of the Pareto exponent significantly greater than 1. This is consistent with Rosen and Resnick's earlier result. However, our result for urban agglomerations differs from their results. We find that, for agglomerations, the Pareto exponent tends to be significantly less than 1 using OLS (Rosen and Resnick find that, for agglomerations, the Pareto exponent is equal to 1). This could indicate the impact of increasing suburbanisation in the growth of large cities in the last 20 years. The OLS estimates of the Pareto exponent are unimodally distributed, while the Hill estimates are bimodal; this may indicate that at least one of the estimators is not appropriate. Finally, we show that political variables appear to matter more than economic geography variables in determining the size distribution of cities.

The next section outlines Zipf's Law and briefly reviews the empirical literature in the area. Section 3 describes the data and the methods, and Section 4 presents the results, along with nonparametric analysis of the Pareto exponent. Section 5 takes the analysis further by seeking to uncover the relationship between these measures of the urban system and some economic variables, based on models of the size distribution of cities. The last section concludes.

---

[1] Although to be clear, it is not a "Law", but simply a proposition on the size distribution of cities.

[2] A partial list includes Krugman (1996), Gabaix (1999), Axtell and Florida (2000), Reed (2001), Cordoba (2003), Rossi-Hansberg and Wright (2004). In addition, Brakman et al. (1999) and Duranton (2002) seek to model the empirical city size distribution, even if it does not follow Zipf's Law.

2. Zipf's Law and related literature

The form of the size distribution of cities as first suggested by Auerbach in 1913 takes the following Pareto distribution:

$$y = Ax^{-\alpha} \tag{1}$$

or

$$\log y = \log A - \alpha \log x \tag{2}$$

where x is a particular population size, y is the number of cities with populations greater than x, and A and α are constants (A, α>0). Zipf's (1949) contribution was to propose that the distribution of city sizes could not only be described as a Pareto distribution but that it took a special form of that distribution with $\alpha = 1$ (with the corollary that A is the size of the largest city). This is Zipf's Law.

The key empirical article in this field is Rosen and Resnick (1980). Their study investigates the value of the Pareto exponent for a sample of 44 countries. Their estimates ranged from 0.81 (Morocco) to 1.96 (Australia), with a sample mean of 1.14. The exponent in 32 out of 44 countries exceeded unity. This indicates that populations in most countries are more evenly distributed than would be predicted by the rank-size rule. Rosen and Resnick also find that, where data was available, the value of the Pareto exponent is lower for urban agglomerations as compared to cities.

More detailed studies of the Zipf's Law (e.g. Guerin-Pace's (1995) study of the urban system of France between 1831 and 1990 for cities with more than 2000 inhabitants) show that estimates of α are sensitive to the sample selection criteria. This implies that the Pareto distribution is not precisely appropriate as a description of the city size distribution. This issue was also raised by Rosen and Resnick, who explored adding quadratic and cubic terms to the basic form, giving

$$\log y = (\log A)' + \alpha' \log x + \beta' (\log x)^2 \tag{3}$$

$$\log y = (\log A)'' + \alpha'' \log x + \beta'' (\log x)^2 + \gamma'' (\log x)^3 \tag{4}$$

They found indications of both concavity ($\beta' < 0$) and convexity ($\beta' > 0$) with respect to the pure Pareto distribution, with more than two thirds (30 of 44) of countries exhibiting convexity. As Guerin-Pace (1995) demonstrates, this result is also sensitive to sample selection.[3]

[3] The addition of such terms can be viewed as a weak form of the Ramsey (1969) RESET test for functional form misspecification. In our sample, we find that the full RESET test rejects the null of no omitted variables almost every time.

There have also been papers which seek to test directly some of the theoretical models of Zipf's Law; in particular, the idea, associated with Gabaix (1999) and Cordoba (2003), that Zipf's Law follows from Gibrat's Law. Black and Henderson (2003), for example, test whether the growth rate of cities in the US follows Gibrat's Law. They conclude that neither Zipf's Law nor Gibrat's Law apply in their sample of cities. On the other hand, Ioannides and Overman (2003), using similar data but a different method, find that Gibrat's Law holds in the US. This is an interesting development; however, data limitations prevent us from being able to test for Gibrat's Law, as the test requires data on the growth rate of cities.

While obtaining the value for the Pareto exponent for different countries is interesting in itself, there is also great interest in investigating the factors that may influence the value of the exponent, for such a relationship may point to more interesting economic and policy-related issues. Rosen and Resnick (1980), for example, find that the Pareto exponent is positively related to per capita GNP, total population and railroad density, but negatively related to land area. Mills and Becker (1986), in their study of the urban system in India, find that the Pareto exponent is positively related to total population and the percentage of workers in manufacturing. Alperovich's (1993) cross-country study using values of the Pareto exponent from Rosen and Resnick (1980) finds that it is positively related to per capita GNP, population density, and land area, and negatively related to the government share of GDP, and the share of manufacturing value added in GDP.

3. Data and methods

3.1. Data

This paper uses a new data set, obtained from the following website: Thomas Brinkhoff (2004): City Population, http://www.citypopulation.de. This site has data on city populations for over 100 countries. However, we have only made use of data on 75 countries, because for smaller countries the number of cities was very small (less than 20 in most cases). For each country, data is available for one to four census periods, the earliest record being 1972 and the latest 2001. This gives a total number of country-year pairs of observations of 197. For every country (except Peru and New Zealand), data is available for administratively defined cities; but for a subset of 26 countries (including Peru and New Zealand), there is also data for urban agglomerations, defined as a central city and neighbouring communities linked to it by continuous built-up areas or many commuters.

The precise definition of cities is an issue that often arises in the literature. Official statistics, even if reliable, are still based on the statistical authorities' definition of city boundaries. These definitions may or may not coincide with the economically meaningful definition of "city" (see Rosen and Resnick, 1980 or Cheshire, 1999). Data for agglomerations might more closely approximate a functional definition, as they typically include surrounding suburbs where the workers of a city reside.

To alleviate fears as to the reliability of online data, we have cross-checked the data with official statistics published by the various countries' statistical agencies, the UN Demographic Yearbook and the Encyclopaedia Britannica Book of the Year (2001). The data in every case matched with one or more of these sources.[4]

The lower population threshold for a city to be included in the sample varies from one country to another—on average, larger countries have higher thresholds, but also a larger number of cities in the sample. The countries chosen all have minimum thresholds of at least 10,000. Our sample of 75 countries includes all the countries in the Rosen and Resnick sample, except for Ghana, Sri Lanka and Zaire.

Some discussion of the sample selection criteria used here is in order. Cheshire (1999) raises this issue. He argues that there are three possible criteria: a fixed number of cities, a fixed size threshold, or a size above which the sample accounts for some given proportion of a country's population. He objects to the third criterion as it is influenced by the degree of urbanisation in the country. However, it is simple to see that the other two criteria he prefers are also problematic: the first because for small countries a city of rank *n* might be a mere village indistinguishable from the surrounding countryside, whereas for a large country the *n*th city might be a large metropolis. While the limitation of the second criterion is that when countries are of different sizes, a fixed threshold would imply that a different fraction of the urban system is represented in the sample. The data as we use it seems in our opinion to represent the best way of describing the reality that large countries do have more cities than small countries on average, however, what is defined as a city in a small country might not be considered as such in a larger country.

As an additional test, data was kindly provided by Paul Cheshire on carefully defined Functional Urban Regions (FURs), for 12 countries in the EC and the EFTA. This data set, by more carefully defining the urban system, might be viewed as a more valid test of Zipf's Law. However, because the minimum threshold in the data set is 300,000, meaningful regressions were run for only the seven largest countries in the sample (France, West Germany, Belgium, the Netherlands, Italy, Spain, and the United Kingdom). This serves as an additional check on the validity of the results obtained using the main data set. The results using Cheshire's data set are similar to those obtained using Brinkhoff's data set and are not reported for brevity.

Data for the second stage regression which seeks to uncover the factors which influence α is obtained from the World Bank World Development Indicators CD-ROM, the International Road Federation World Road Statistics, the UNIDO Industrial Statistics Database, and the Gallup et al. (1999) geographical data set. The GASTIL index is from Freedom House.

[4] For example, the figures for South Africa, Canada, Colombia, Ecuador, Mexico, India, Malaysia, Pakistan, Saudi Arabia, South Korea, Vietnam, Austria and Greece are the same as those from the United Nations Demographic Yearbook. The figures for Algeria, Egypt, Morocco, Kenya, Argentina, Brazil, Peru, Venezuela, Indonesia, Iran, Japan, Kuwait, Azerbaijan, Philippines, Russia, Turkey, Jordan, Bulgaria, Denmark, Finland, Germany, Hungary, the Netherlands, Norway, Poland, Portugal, Romania, Sweden, Switzerland, Spain, Ukraine and Yugoslavia are the same as those from the Encyclopaedia Britannica Book of the Year. It should be noted that the Encyclopaedia Britannica Book of the Year (2001) lists Brinkhoff's website as one of its data sources, thus adding credibility to the data obtained from this website.

3.2. Methods

Two estimation methods are used in this paper: OLS and the Hill (1975) method. Using OLS, two regressions are run:

$$\log y = \log A - \alpha \log x \tag{2}$$

$$\log y = (\log A)' + \alpha' \log x + \beta' (\log x)^2 \tag{3}$$

Eq. (2) seeks to test whether $\alpha = 1$ and A = size of largest city, while Eq. (3) seeks to uncover any nonlinearities that could indicate deviations from the Pareto distribution. Both these regressions are run for each country and each time period separately, using OLS with heteroskedasticity-robust standard errors. This is done for all countries although a Cook-Weisberg test for heteroskedasticity has mixed results. As an additional check, the regressions were also run using lagged population of cities as an instrument for city population, to address possible measurement errors and endogeneity issues involved in running such a regression. The IV estimators passed the Hausman specification test for no systematic differences in parameter values, as well as the Sargan test for validity of instruments. Results using IV are very similar to the ones obtained using OLS, and are not reported.[5]

One potentially serious problem with the Zipf regression is that it is biased in small samples. Gabaix and Ioannides (in press) show using Monte Carlo simulations that the coefficient of the OLS regression of Eq. (2) is biased downward for sample sizes in the range that is usually considered for city size distributions. Further, OLS standard errors are grossly underestimated (by a factor of at least 5 for typical sample sizes), thus leading to too many rejections of Zipf's Law. They also show that, even if the actual data exhibit no nonlinear behaviour, OLS regression of Eq. (3) will yield a statistically significant coefficient for the quadratic term an incredible 78% of the time in a sample of 50 observations.

This clearly has serious implications for our analysis. Gabaix and Ioannides (in press) propose the Hill (1975) estimator as an alternative procedure for calculating the value of the Pareto exponent. Under the null hypothesis of the power law, it is the maximum likelihood estimator. Thus, for a sample of n cities with sizes $x_1 \geq \ldots \geq x_n$, this estimator is:

$$\hat{\alpha} = \frac{n-1}{\sum_{i=1}^{n-1} (\ln x_i - \ln x_n)} \tag{5}$$

[5] However, using lagged values as instruments is appropriate only if shocks to city populations are not correlated accross time. Otherwise the instruments would be correlated with the errors and would yield inconsistent estimates.

while the standard error is given by:

$$\sigma_n(\hat{\alpha}) = \hat{\alpha}^2 \left(\frac{\sum_{i=1}^{n-1} (\ln x_i - \ln x_{i+1})^2}{n-1} - \frac{1}{\hat{\alpha}^2} \right)^{\frac{1}{2}} n^{-\frac{1}{2}} \tag{6}$$

The best known paper that has used the Hill estimator for estimating Zipf's Law is Dobkins and Ioannides (2000), who find that the Pareto exponent is declining in the US over time, using either OLS or the Hill method. However, they also find that the Hill estimate of the Pareto exponent is always smaller than the OLS estimate, thus calling into question the appropriateness of the Hill method, at least for the US. Additional evidence from Black and Henderson (2003), who use a very similar data set, suggests that the reliability of the Hill estimate is dependent on the curvature of the log rank–log population plot, something which we return to in Section 4.3 below.

As an aside, it should be noted that, in comparing the two alternative estimators, the OLS estimator is a bit heuristic, since it simply finds the best fit line to a plot of the log of city rank to the log of city population. On the other hand, the Hill estimator starts out by assuming a Pareto distribution for the data, and finds the best (maximum likelihood) estimator for that distribution. However, if the distribution does not follow a Pareto distribution, then the Hill estimator is no longer the maximum likelihood estimator.

We plot the kernel density functions for the estimates of the Pareto exponent using the OLS and Hill estimators to give a better description and further insights of the distribution of the values of the exponent across countries. The Pareto exponent is then used as the dependent variable in a second stage regression where the objective is to explain variations in this measure using variables obtained from models of political economy and economic geography.

4. Results

In this section, we discuss only the results for the latest available year for each country, for the regressions (2) and (3) for Zipf's Law and the Hill estimator. This is to reduce the size of the tables. Full details are available from the author upon request.

4.1. Zipf's Law for cities

Table 1 presents the detailed results of the OLS regressions of Eqs. (2) and (3) and the Hill estimator for cities. For OLS, the largest value of the Pareto exponent (1.719) is obtained for Kuwait, followed by Belgium, whereas the lowest value is obtained for

Table 1
Results of OLS regression of Eqs. (2) and (3) and the Hill estimator, for the sample of cities, for latest year of each country

Country	Year	Cities	OLS				Hill
			α	α′	β′	log*A*	α
Algeria	1998	62	1.351**	− 2.3379	0.0408	18.7999**	1.3586*
Egypt	1996	127	0.9958	− 2.9116**	0.0781**	15.0635	1.0937
Ethiopia	1994	63	1.0653	− 4.3131**	0.1425**	14.2275	1.3341*
Kenya	1989	27	0.8169**	− 1.9487**	0.0486**	11.2945**	1.0060
Morocco	1994	59	0.8735**	− 1.0188	0.006	13.0697**	0.9295
Mozambique	1997	33	0.859**	1.0146**	− 0.0811**	12.1286**	0.8107
Nigeria	1991	139	1.0409**	− 0.9491	− 0.00375	15.9784**	1.0459
South Africa	1991	94	1.3595**	− 1.1031	0.01076	19.1221**	1.2679*
Sudan	1993	26	0.9085	− 0.2142	− 0.0283	13.0723*	1.0066
Tanzania	1988	32	1.01	− 1.8169	0.0348	13.6915	0.9089
Australia	1998	131	1.2279**	7.8935**	− 0.4055**	17.6039**	0.8012**
Argentina	1999	111	1.0437	2.9939**	− 0.1652**	16.1345**	0.9670
Brazil	2000	411	1.1341**	− 0.0963**	− 0.0418**	18.3681**	1.0607
Canada	1996	93	1.2445**	0.4273	− 0.0689	18.0872**	1.2526
Chile	1999	67	0.8669**	− 0.6516	− 0.00915	13.0195**	0.7908*
Colombia	1999	111	0.9024**	− 0.804	− 0.00404	14.0252**	0.9345
Cuba	1991	55	1.09	− 3.6859**	0.1093**	15.1299	1.3177
Dominican Republic	1993	23	0.8473	− 2.6376*	0.0749*	11.6874**	0.8029
Ecuador	1995	42	0.8083**	− 1.4086	0.0255	11.6871**	0.9015
Guatemala	1994	13	0.7287**	− 3.6578**	0.1249**	9.71255**	1.2074
Mexico	2000	162	0.9725	1.9514**	− 0.1172*	15.8281	0.8127**
Paraguay	1992	19	1.0137	− 1.9584	0.0415	13.1465	1.2571
USA	2000	667	1.3781**	− 1.9514**	0.0235**	21.3849**	0.9339
Venezuela	2000	91	1.0631*	− 0.7249	− 0.0139	15.8205**	1.4277**
Azerbaijan	1997	39	1.0347	− 5.2134**	0.1812**	13.6575	1.3605
Bangladesh	1991	79	1.0914	− 4.1878**	0.1274**	15.6311	1.3545*
China	1990	349	1.1811**	1.4338**	− 0.1008**	19.5678**	0.9616
India	1991	309	1.1876**	− 0.7453	− 0.0170**	19.3916**	1.2178**
Indonesia	1990	235	1.1348**	− 2.6325**	0.0610**	17.4209**	1.2334**
Iran	1996	119	1.0578**	− 1.5539	0.01985	16.2499**	1.0526
Israel	1997	55	1.0892*	1.4982**	− 0.1148**	14.8869**	1.0409
Japan	1995	221	1.3169**	− 0.6325	− 0.02655	20.6491**	1.2249**
Jordan	1994	34	0.8983**	− 2.4831**	0.0699**	12.0845**	1.0629
Kazakhstan	1999	33	0.9615	4.8618**	− 0.2444**	13.8818	0.8653
Kuwait	1995	28	1.719**	5.8975**	− 0.3547**	20.5508**	1.6859*
Malaysia	1991	52	0.8716*	2.8194**	− 0.1622**	12.6602**	0.8419
Nepal	2000	46	1.1870**	− 2.0959	0.0405	15.5832**	1.2591
Pakistan	1998	136	0.9623	− 2.4838**	0.0607**	15.0410**	1.0626
Philippines	2000	87	1.0804	3.4389**	− 0.1838**	16.4972**	0.8630
Saudi Arabia	1992	48	0.7824**	0.02426**	− 0.0333*	11.9143**	0.7302**
South Korea	1995	71	0.907**	− 0.3178	− 0.02251	14.5804**	0.6850**
Syria	1994	10	0.7442*	− 1.4709	0.02796	10.8967**	1.0862
Taiwan	1998	62	1.0587**	0.1482**	− 0.0487**	15.7536**	0.9294
Thailand	2000	97	1.1864**	− 4.9443**	0.1553**	16.6797	1.4184**
Turkey	1997	126	1.0536	− 2.6659**	0.0642**	16.1683	1.1850
Uzbekistan	1997	17	1.0488	− 8.9535**	0.3048**	14.7941	1.5111*
Vietnam	1989	54	0.9756**	− 1.4203	0.0184**	14.1331*	0.8028

Table 1 (*continued*)

Country	Year	Cities	OLS				Hill
			α	α'	β'	log$A$	α
Austria	1998	70	0.9876	−3.9862**	0.1358**	13.0823	1.4226**
Belarus	1998	41	0.8435**	0.6492**	−0.0639**	12.2363**	0.7503*
Belgium	2000	68	1.5895**	−2.1862	0.02647	20.5048**	1.8348*
Bulgaria	1997	23	1.114	−4.8424**	0.1531**	15.1382	1.2862
Croatia	2001	24	0.9207	−1.7693	0.03769	12.0916**	0.9551
Czech Republic	2001	64	1.1684**	−3.5189**	0.1029**	15.6961**	1.2669
Denmark	1999	58	1.3608**	−2.7601**	0.06274*	17.5639**	1.3753*
Finland	1999	49	1.1924**	−2.468**	0.0569**	15.6367**	1.3462
France	1999	104	1.4505**	−4.1897**	0.1137**	20.2497**	1.6388**
Germany	1998	190	1.238**	−0.3019**	−0.0384**	18.6477**	1.2548**
Greece	1991	43	1.4133**	−6.2019**	0.2036**	18.5979**	1.4804*
Hungary	1999	60	1.124**	−4.0186**	0.1254**	15.1636	1.2789
Italy	1999	228	1.3808**	−3.9073**	0.1064**	19.8143**	1.4967**
Netherlands	1999	97	1.4729**	−0.4333	−0.04491	20.0318**	1.4436**
Norway	1999	41	1.2704**	−4.5945**	0.1481**	16.2593**	1.4026
Poland	1998	180	1.1833**	0.3931**	−0.0679**	17.2931**	1.0908
Portugal	2001	70	1.382**	−4.1362**	0.1241**	17.7945**	1.6703**
Romania	1997	70	1.1092*	−0.05598	−0.0445	15.9369**	1.0598
Russia	1999	165	1.1861**	1.2459*	−0.0942*	18.9423**	1.0344
Slovakia	1998	42	1.3027**	−4.4861**	0.1428**	16.5644**	1.4810*
Spain	1998	157	1.1859**	−0.06586	−0.04697	17.5737**	1.0969
Sweden	1998	120	1.4392**	−1.2181	−0.00991	19.1777**	1.2867**
Switzerland	1998	117	1.4366**	−6.1258**	0.2229**	17.8549**	1.7386**
Ukraine	1998	103	1.0246	1.5787	−0.1058**	15.7615**	1.0197
Yugoslavia	1999	60	1.1827*	−2.2817	0.04839	15.8798**	1.1670
United Kingdom	1991	232	1.4014**	−3.5503**	0.0894**	20.3123**	1.3983**

*Significant at 5%; **significant at 1%; for α, significantly different from 1; for α', significantly different from −1; for β', significantly different from 0; for log$A$, significantly different from the log of the population of the largest city. α is defined as a positive value; to compare the coefficients of log$x$ in Eq. (2) and (log x)′ in Eq. (3), we compare $-\alpha$ with α'.

Guatemala at 0.7287, followed by Syria and Saudi Arabia. Unsurprisingly, the former two countries are associated with a large number of small cities and no primate city, whereas in the latter three countries one or two large cities dominate the urban system. The left side of Table 2 summarises the statistical significance of the Pareto exponent, using both OLS and the Hill estimator for cities. Using OLS, α is significantly greater than 1 for 39 of our 73 countries, while a further 14 observations are significantly less than one. This is consistent with Rosen and Resnick's result, as they find that 32 of their 44 countries had a Pareto exponent significantly greater than 1, while 4 countries had the exponent significantly less than 1.

For the Hill estimator, the country with the largest value of the Pareto exponent is Belgium with a value of 1.742, followed by Switzerland and Portugal. The lowest values were obtained for South Korea, Saudi Arabia and Belarus. It is clear that the identity of the countries with the highest and lowest values for the Pareto exponent differ between the OLS and the Hill estimators. In fact, the correlation between the OLS estimator and the Hill estimator is not exceptionally high, at 0.7064 for the latest available period (the

Table 2
Breaking down the results of OLS regressions (2) and (3) and the Hill estimator: statistical significance (5% level) in the latest available observation, for cities and urban agglomerations

Summary results: OLS estimates of α

Cities				Agglomerations			
Continent	$\alpha<1$	$\alpha=1$	$\alpha>1$	Continent	$\alpha<1$	$\alpha=1$	$\alpha>1$
Africa	3	4	3	Africa	1	1	
N America		1	2	N America	2	1	
S America	4	4	2	S America	3	2	
Asia	5	8	10	Asia	3	2	
Europe	2	3	21	Europe	5	2	2
Oceania			1	Oceania	2		
Total	14	20	39	Total	16	8	2

Summary results: OLS estimates of β'

Cities				Agglomerations			
Continent	$\beta'<0$	$\beta'=0$	$\beta'>0$	Continent	$\beta'<0$	$\beta'=0$	$\beta'>0$
Africa	1	6	3	Africa	1		1
N America		1	2	N America	2	1	
S America	3	4	3	S America		5	
Asia	11	5	8	Asia	2	2	1
Europe	4	7	14	Europe	3	4	2
Oceania	1			Oceania	1	1	
Total	20	23	30	Total	9	13	4

Summary results: OLS estimates of A (compared to largest city)

Cities				Agglomerations			
Continent	Less than	Equal to	Greater than	Continent	Less than	Equal to	Greater than
Africa	3	4	3	Africa	1	1	
N America		1	2	N America	1	2	
S America	5	2	3	S America	5		
Asia	6	7	10	Asia	2	3	
Europe	2	3	21	Europe	5	3	1
Oceania			1	Oceania	2		
Total	16	17	40	Total	16	9	1

Summary results: Hill estimator for α

Cities				Agglomerations			
Continent	$\alpha<1$	$\alpha=1$	$\alpha>1$	Continent	$\alpha<1$	$\alpha=1$	$\alpha>1$
Africa		7	3	Africa	1	1	
N America	1	1	1	N America	1	2	
S America	1	9		S America	1	4	
Asia	2	14	7	Asia		5	
Europe	1	12	13	Europe	1	8	
Oceania	1			Oceania	1	1	
Total	6	43	24	Total	5	21	

Spearman rank correlation is 0.6823). This can be interpreted as saying that, because we use a different number of cities for each country, and since the OLS bias is larger for small samples, we should not expect the results of the OLS and Hill estimators to be perfectly correlated. Indeed, we find a weak negative correlation between the difference in estimates using the two methods, and the number of cities in the sample (corr = – 0.2575).

For statistical significance of the Hill estimator, one key result of Gabaix and Ioannides (in press) is that the standard errors of the OLS estimator are grossly underestimated. Thus, Table 2 shows that using the Hill estimator, 43 of the 73 countries (or 59%) in our sample for cities have values of the Pareto exponent that are not significantly different from the Zipf's Law prediction of 1, with 24 countries having values significantly higher than 1, while only 6 countries have values significantly less than 1. Hence, the overall pattern of statistical significance of the Pareto exponent for the Hill estimator follows that of the OLS estimator, except that there are fewer significant values for the Hill estimator because the (correct) standard errors are larger than those estimated using OLS.

The top half of Table 3 summarises the results of both OLS and Hill estimators for cities. The first set of observations labelled Full Sample shows the summary statistics for α for the latest available observation in all countries. We see that the mean of the Pareto exponent for cities using OLS is approximately 1.11. This lends support to Rosen and Resnick's result (they obtain a mean value for the Pareto exponent of 1.13). For the Hill estimator, the mean of the Pareto exponent is 1.167, which is statistically different from the mean for the OLS estimator at the 5% level. This is consistent with the argument in Gabaix and Ioannides (in press), that OLS is biased downward in small samples. However, we also find that for 34 of the 73 countries, the Hill estimate of the Pareto exponent is smaller than the OLS estimate, which may indicate a bias in the Hill estimator (recall that the Hill estimator is supposed to overcome the downward bias of the OLS estimator; Section 4.3 discusses this further).

Breaking down the results by continents, we find that, for both OLS and Hill estimators, there seems to be a clear distinction between Europe, which has a high average value of the Pareto exponent (the average being above 1.2 using OLS) and Asia, Africa, and South America, which have low average values of the exponent (below 1.1 using OLS).[6] This indicates that populations in Europe are more evenly spread over the system of cities than in the latter three continents. Indeed, 21 of the 26 European countries in our sample had α significantly greater than 1 using OLS. These findings raise the interesting question of why these differences exist between different continents. Could it be the different levels of development, or institutional factors? The next section will seek to identify the reasons for these apparently systematic variations.

Table 1 also provides the results of the value of the intercept term of the linear regression (2). As Alperovich (1984, 1988) notes, a proper test of Zipf's Law should not

[6] A two-sample *t*-test shows that the average Pareto exponent for Europe is significantly different from that for the rest of the world as a whole.

Table 3
Summary statistics: by continent; values of α using OLS and Hill estimators, for cities and agglomerations

OLS for cities	Obs	Mean	Std. dev.	Min	Max
Full sample	73	1.1114	0.2042	0.7287	1.719
Africa	10	1.0280	0.1910	0.8169	1.3595
North America	3	1.2008	0.1705	1.0127	1.3451
South America	10	0.9531	0.1363	0.7287	1.1391
Asia	23	1.0633	0.2027	0.7442	1.719
Europe	26	1.2306	0.1735	0.8435	1.540
Oceania	1	1.2685		1.2685	1.2685
Hill for cities	Obs	Mean	Std. dev.	Min	Max
Full sample	73	1.1667	0.2583	0.6850	1.7422
Africa	10	1.0762	0.1868	0.8107	1.3586
North America	3	1.1772	0.2724	0.8751	1.4039
South America	10	1.0255	0.1819	0.8028	1.3177
Asia	23	1.1226	0.2602	0.6850	1.6859
Europe	26	1.3063	0.2542	0.7503	1.7422
Oceania	1	0.8398		0.8398	0.8398
OLS for agglomerations	Obs	Mean	Std. dev.	Min	Max
Full sample	26	0.8703	0.1526	0.5856	1.2301
Africa	2	0.8661	0.3374	0.6275	1.1047
North America	3	0.8941	0.0648	0.8345	0.9631
South America	5	0.8510	0.1065	0.7025	0.9904
Asia	5	0.8778	0.1316	0.6813	1.0001
Europe	9	0.9111	0.1725	0.6349	1.2301
Oceania	2	0.6844	0.1399	0.5856	0.7833
Hill for agglomerations	Obs	Mean	Std. dev.	Min	Max
Full sample	26	0.8782	0.2276	0.5058	1.5897
Africa	2	1.0477	0.7665	0.5058	1.5897
North America	3	0.7202	0.1714	0.5225	0.8273
South America	5	0.8812	0.2084	0.5229	1.0567
Asia	5	0.8837	0.1133	0.7286	1.0384
Europe	9	0.9402	0.1178	0.6778	1.0903
Oceania	2	0.6458	0.1939	0.5087	0.7829

only consider the value of the Pareto exponent, but also whether the intercept term A is equal to the size of the largest city. We find, perhaps unsurprisingly, that whenever the Pareto exponent is significantly greater than 1, the intercept term is also greater than the size of the largest city (this is almost by construction: in a log-rank–log-population plot, the largest city enters on the horizontal axis, so that, provided the largest city is not too far from the best-fit line, if the line has slope equal to 1, it must be that the vertical intercept is equal to the horizontal intercept). A comparison of the first and third panels of Table 2 confirms this result, as the estimates of the Pareto exponent and the intercept follow almost identical patterns.

For values of the quadratic term, the patterns are less strong. Recalling that a significant value for the quadratic term represents a deviation from the Pareto distribution, we find the following results. For the cities sample, 30 observations or 41% display a value for the quadratic term significantly greater than zero, indicating convexity of the log-rank–log-population plot, while 20 observations (26%) have a value for the quadratic term significantly less than zero, indicating concavity of the log-rank–log-population plot. These results are again in the same direction as those obtained by Rosen and Resnick (1980), but less strong (they find that the quadratic term is significantly greater than zero for 30 out of 44 countries).

One additional result that arises out of the quadratic regression (3) is that including the quadratic term often dramatically changes the value or even the sign of the coefficient of the linear term. This is actually a fairly common result in the literature; Rosen and Resnick (1980) find that, in the quadratic regression (3), the linear term is positive for 6 of their 44 countries; this compares with 17 of our 73 countries (in Table 1, α is a positive value, but the coefficient on the term ($\log x$) in the linear specification (2) is $-\alpha$). This sign change in the linear term can be explained by the different interpretations of the linear term in Eqs. (2) and (3). In a linear regression, the linear term gives the slope of the best-fit line; but in a quadratic regression, the linear term gives the location of the maximum or minimum point of the best-fit line.[7]

Figs. 1 and 2 graph the estimates for the Pareto exponent for all countries using the latest available observation, using the OLS and Hill estimators, respectively, including the 95% confidence interval and sorting the sample according to values of the Pareto exponent (the confidence intervals do not form a smooth series since each country has a different standard error). The figures show graphically what the tables summarise. We find that the confidence intervals for the Hill estimator are larger than for the OLS estimator, and hence that we reject the null hypothesis that the Pareto exponent is equal to 1 more frequently using the OLS estimator (in the figures, a rejection occurs when no portion of the vertical line indicating the confidence interval intersects the horizontal line at 1.00).

4.2. Zipf's Law for urban agglomerations

It is frequently claimed (see e.g. Rosen and Resnick, 1980 or Cheshire, 1999) that Zipf's Law holds if we define cities more carefully, by using data on urban agglomerations rather than cities. To see if this is in fact the case, we also run the OLS regressions (2) and (3), and the Hill estimator, for our sample of 26 countries for which data on urban agglomerations is available.

The results for the latest available period for urban agglomerations are presented in Table 4, and are summarised in the lower half of Table 3. Using either OLS or the Hill estimator, the mean value of the Pareto exponent is lower for agglomerations than for cities

[7] If the function is $y=a+bx+cx^2$, then y is maximised when $x=-(b/2c)$. Since our data points have values for x (the log of city size) between 9 and 17, it is possible that, if the quadratic term is negative, the maximum of y occurs at a positive value of x, thus implying a positive value of b.

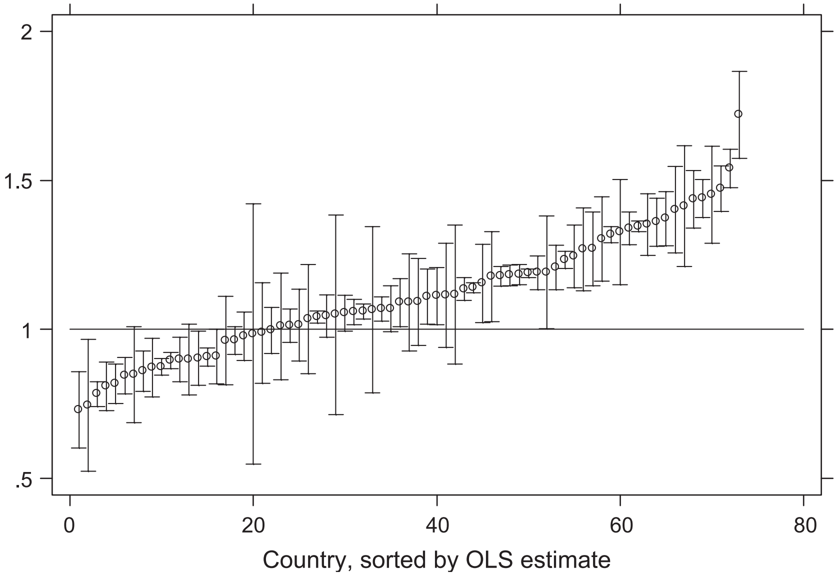

Fig. 1. Values of the OLS estimate of the Pareto exponent with the 95% confidence interval, for the full sample of 73 countries for the latest available period, sorted according to the Pareto exponent.

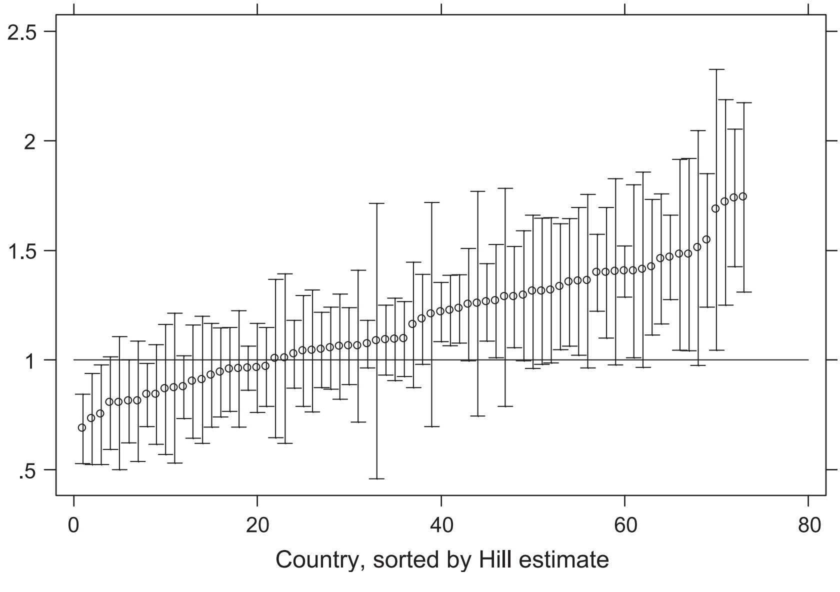

Fig. 2. Values of the Hill estimate of the Pareto exponent with the 95% confidence interval, for the full sample of 73 countries for the latest available period, sorted according to the Pareto exponent.

Table 4
Results of OLS regression of Eqs. (2) and (3), and the Hill estimator, for the sample of urban agglomerations, for latest year of each country

Country	Year	AGG	OLS				Hill
			α	α'	β'	log$A$	α
Morocco	1982	10	1.10466	− 14.207**	0.48473**	15.8475	1.5897
South Africa	1991	23	0.6275**	3.8188**	− 0.1747**	10.1609**	0.5058**
Australia	1998	21	0.5855**	0.9107	− 0.05806*	9.4412**	0.5087**
New Zealand	1999	26	0.7833**	− 0.8086	0.0011	10.8562**	0.7830
Argentina	1991	19	0.7025**	− 1.1177	0.01527	11.1267**	0.5229**
Brazil	2000	18	0.9904	− 1.1245	0.00444	16.5577	0.9737
Canada	1996	56	0.8345**	− 0.2635	− 0.0225	13.0979**	0.8273
Colombia	1993	16	0.8278**	− 0.2378	− 0.02141	12.9431**	1.0567
Ecuador	1990	43	0.9046	− 2.0169	0.0474	12.7637**	0.9573
Mexico	2000	38	0.9631	− 1.3863	0.01501	15.6724	0.8107
Peru	1993	65	0.8295**	− 1.5843	0.03171	12.3510**	0.8955
USA	2000	336	0.8847**	3.4992**	− 0.1669**	16.1013	0.5225**
Bangladesh	1991	43	0.8068**	− 2.9315**	0.08399**	12.1569**	0.9141
India	1991	178	0.9579**	0.1559**	− 0.0419**	16.2945	0.9001
Indonesia	1990	193	1.0001	− 1.1315	0.00532	15.8411	1.0384
Jordan	1994	10	0.6813**	0.2377	− 0.03703	9.7100**	0.7286
Malaysia	1991	71	0.9429	3.3355**	− 0.1872**	13.7914	0.8370
Austria	1998	34	0.7501**	− 0.6338	− 0.0051	10.6591**	0.6778**
Denmark	1999	27	0.8166**	− 3.7224**	0.1235**	11.2213**	1.0903
France	1999	114	1.02332	− 1.5263	0.02014	15.7905	1.0643
Germany	1996	144	0.8902**	0.5697**	− 0.0578**	14.6429**	0.8886
Greece	1991	15	0.6349**	− 3.987**	0.1324**	9.2190**	0.9499
Netherlands	1999	21	1.2301*	0.83	− 0.08044	17.5350**	0.9703
Norway	1999	19	0.8828*	− 1.7724	0.03853	11.7679**	0.9212
Switzerland	1998	48	0.9847	− 0.1671	− 0.0356**	13.7188	0.9557
United Kingdom	1991	151	1.0303*	− 0.9192	− 0.0045	16.0465	0.9438

AGG: number of urban agglomerations. *Significant at 5%; **significant at 1%; for α, significantly different from 1; for α', significantly different from − 1; for β', significantly different from 0; for log$A$, significantly different from the log of the population of the largest city. α is defined as a positive value; to compare the coefficients of log$x$ in Eq. (2) and $(\log x)'$ in Eq. (3), we compare $-\alpha$ with α'.

(the value is 0.870 for OLS and 0.8782 for the Hill estimator). This is to be expected, since the Pareto exponent is a measure of how evenly distributed is the population (the higher the value of the exponent, the more even in size are the cities), and urban agglomerations tend to be larger relative to the core city for the largest cities than for smaller cities. Once again a slight pattern can be observed across continents; the small sample size however does not make this result particularly strong.

The right side of Table 2 summarises the statistical significance of both OLS and the Hill estimator for agglomerations. Using OLS, the Pareto exponent for agglomerations is significantly greater than one for only two countries (the Netherlands and the United Kingdom), while fully 16 of the 26 observations for agglomerations were significantly less than one (a similar result albeit with weaker significance is obtained using the Hill estimator). Results for the intercept term of the linear regression (2) tracks the results for the Pareto exponent very closely. For the quadratic regression (3), we find that half of the observations (13 out of 26) have a value for the quadratic term not significantly different from zero, with 9 or 35% having a quadratic term significantly less than zero.

Therefore, the claim that Zipf's Law holds for urban agglomerations (see Rosen and Resnick, 1980; Cheshire, 1999) is strongly rejected for our sample of countries in favour of the alternative that agglomerations are more uneven in size than would be predicted by Zipf's Law. Our interpretation of this finding is that, in more recent years, the growth of cities (especially the largest cities) has mainly taken the form of suburbanisation, so that this growth is not so much reflected in administratively defined cities, but shows up as increasing concentration of population in larger cities when urban agglomerations are used.

4.3. Nonparametric analysis of the distribution of the Pareto exponent

An additional way of describing the distribution of the Pareto exponent across countries is to construct the kernel density functions. The advantage of doing so is that it gives us a more complete description of how the values of the Pareto exponent are distributed—whether it is unimodal or bimodal, or whether it is normally distributed or not. In implementing this method, we use the latest available observation for each country. We construct the efficient Epanechnikov kernel function for the Pareto exponent for both the OLS and Hill estimators, using the "optimal" window width (the width that minimises the mean integrated square error if the data were Gaussian and a Gaussian kernel were used), and including an overlay of the normal distribution for comparative purposes.

Fig. 3 shows the kernel function for the OLS estimator. It is slightly right skewed relative to the normal distribution, but is clearly unimodal (with the mode approximately

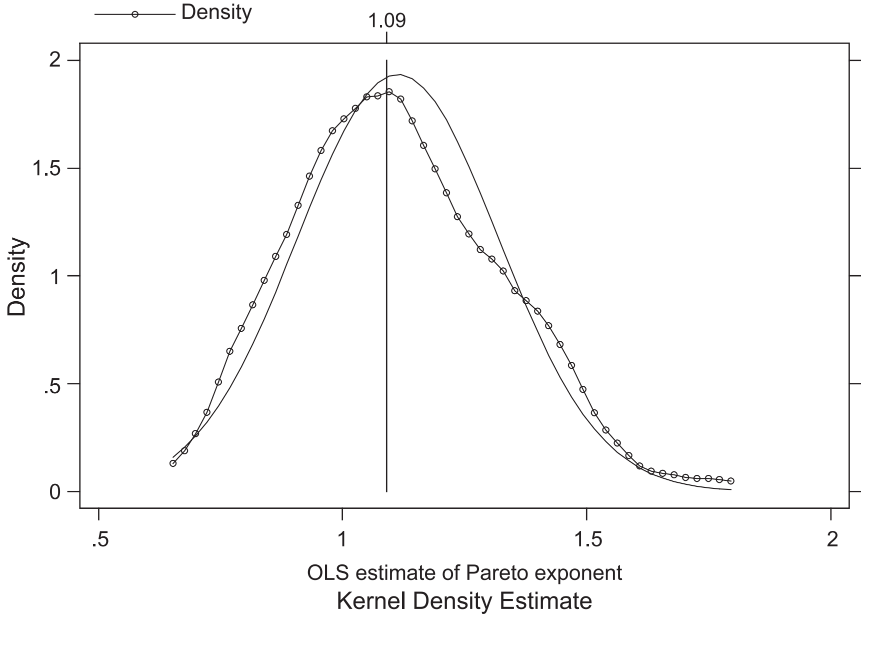

Fig. 3. Kernel density function for Pareto exponent using the OLS estimator (optimal window width = 0.076).

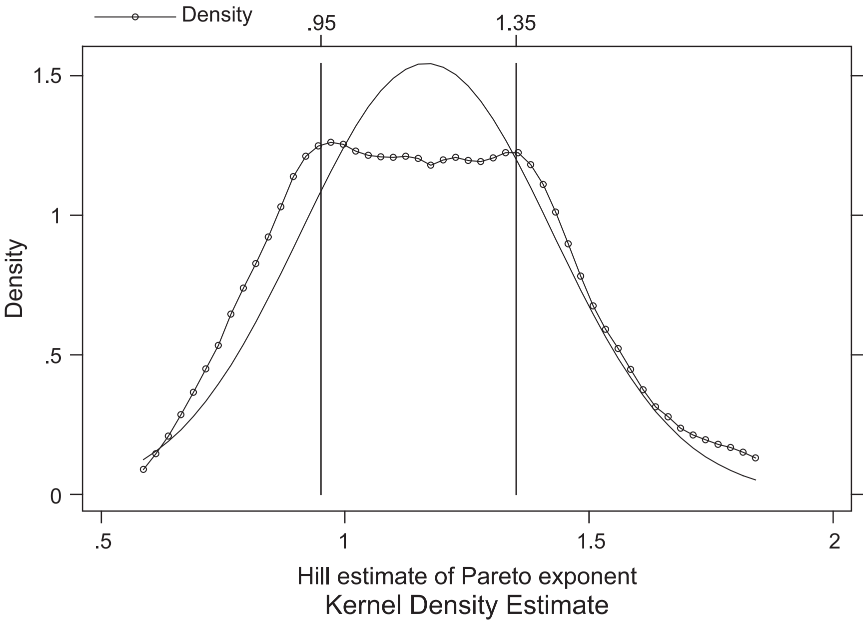

Fig. 4. Kernel density function for the Pareto exponent using the Hill estimator (optimal window width = 0.098).

equal to 1.09) and its distribution is quite close to the normal distribution. Fig. 4 shows the kernel function for the Hill estimator. What is interesting (and a priori unexpected) is that the distribution is not unimodal. Instead, we find that there is no clearly defined mode, rather that observations are spread roughly evenly across ranges of the Pareto exponent between 0.95 and 1.35. Experimenting with narrower window widths (Fig. 5, where the window width is 0.06)[8] shows that the distribution is in fact bimodal, with the two modes at approximately 1.0 and 1.32.

Closer inspection of the relationship between the OLS estimator and Hill estimator of the Pareto exponent, and the value of the coefficient for the quadratic term in the OLS regression Eq. (3), reveals further insights as to what is actually happening. We find that, while the correlation between the OLS estimator of the Pareto exponent and the quadratic term is very low (corr = – 0.0329 for the latest available period), the correlation between the Hill estimator and the quadratic term is high (corr = 0.5063). Further, the correlation between the difference between the Hill estimator and the OLS estimator, and the quadratic term, is even higher (corr = 0.7476; see Fig. 6). What we find is that, in general, the Hill estimator is larger than the OLS estimator if the quadratic term is positive (i.e. the log-rank–log-population plot is convex), while the reverse is true if the quadratic term is negative. In other words, when the size distribution of cities does not follow a Pareto distribution, the Hill estimator may be biased. These results are similar to those obtained

[8] While the “optimal” window width exists, in practice choosing window widths is a subjective exercise. Silverman (1986) shows that the “optimal” window width oversmooths the density function when the data are highly skewed or multimodal.

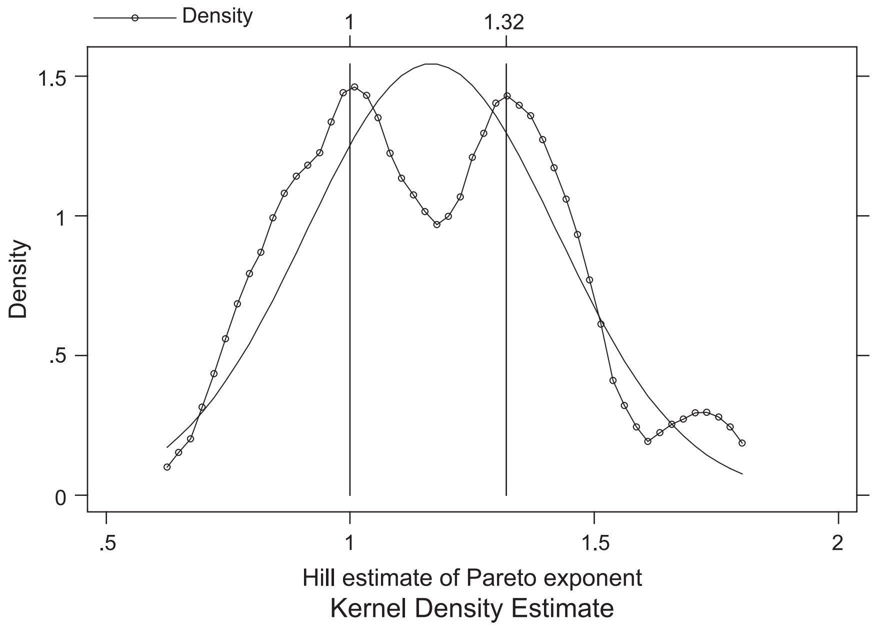

Fig. 5. Kernel density function for the Pareto exponent using the Hill estimator (window width = 0.006, vertical lines at $x = 1.00$ and $x = 1.32$).

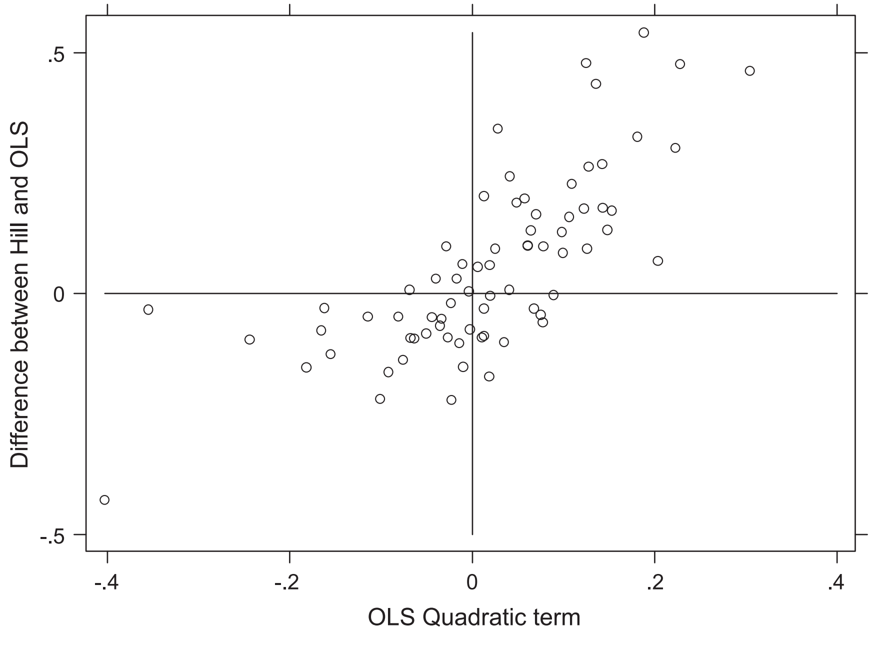

Fig. 6. Relationship between difference between Hill and OLS estimators, and the value of the quadratic term in Eq. (3).

by Dobkins and Ioannides (2000) and Black and Henderson (2003) for US cities (see the brief discussion in Section 3.2 above). Therefore, we should tread carefully in drawing conclusions from the results of the Hill estimator.

5. Explaining variation in the Pareto exponent

The Pareto exponent α can be viewed as a measure of inequality: the larger the value of the Pareto exponent, the more even is the populations of cities in the urban system (in the limit, if $\alpha = \infty$, all cities have the same size). There are many potential explanations for variations in its value. One possibility is a model of economic geography, as exemplified by Krugman (1991) and Fujita et al. (1999). These models can be viewed as models of unevenness in the distribution of economic activity. For certain parameter values, economic activity is agglomerated, while for other parameter values, economic activity is dispersed. The key parameters of the model are: the degree of increasing returns to scale, transport costs and other barriers to trade within a country, the share of mobile or footloose industries in the economy. From Chapter 11 of Fujita et al. (1999), there will be a more uneven distribution of city sizes (smaller Pareto exponent), the greater are scale economies, the lower are transport costs, the smaller the share of manufacturing in the economy, and the lower the share of international trade in the economy. These results can be explained as follows. The greater are scale economies in each manufacturing industry, the fewer the number of cities that will be formed, so that the greater is the average difference in sizes between cities. Similarly, lower transport costs imply that the benefits of locating close to the agricultural periphery are reduced, so that fewer cities are formed. Also, the smaller the share of manufacturing in the economy, the more cities will be formed, as the desire to serve the agricultural periphery induces firms to locate away from existing cities (these conclusions are reached from an analysis of Fujita et al., 1999, Eq. (11.12)). In addition, Chapter 18 of Fujita et al. (1999) shows that a greater extent of international trade weakens the force for agglomeration and leads to a more even distribution of economic activity.[9]

But we can also think of political factors that could influence the size distribution of cities. Ades and Glaeser (1995) argue that political stability and the extent of dictatorship are key factors that influence the concentration of population in the capital city. They develop a model to justify this line of reasoning in terms of the size of the capital city, but their model can be reinterpreted in terms of the urban system as a whole. Political instability or a dictatorship should imply a more uneven distribution of city sizes (i.e. a smaller Pareto exponent). Thus, a dictatorship would be more likely to have a large capital city since rents are more easily obtainable in the national capital. However, regional capitals would also be a source of rents (albeit at a smaller scale than in the national capital). We should therefore see a hierarchy of cities where cities at each tier of the hierarchy are much larger in size than

[9] Strictly speaking, to the best of our knowledge, existing models of economic geography are not able to generate a size distribution of cities that follows a Pareto distribution, without making additional assumptions (c.f. Brakman et al., 1999). They are however able to generate cities of different sizes, and here we seek to explore whether the variables associated with models of economic geography, impact on the size distribution of cities, in the way that is predicted by the models.

cities at a lower tier. Similarly, if the country is politically unstable, then if the government is unwilling or unable to protect the population outside large cities, we should find a more uneven distribution of city sizes since the population would flock to the larger cities.

We also control for other variables that could influence the size distribution of cities, including the size of the country as measured by population, land area or GDP, and also for possible effects of being located in different continents.

Thus our estimated equation is:

$$\alpha_{it} = \delta_0 + \delta_1 \text{GEOG} + \delta_2 \text{POLITIC} + \delta_3 \text{CONTROL} + \delta_4 \text{DUMMIES} + u_{it} \quad (7)$$

where α_{it} is the Pareto exponent, GEOG is the vector of economic geography variables: scale economies, transport costs, nonagricultural economic activity, and trade as a share of GDP (a detailed definition of the variables is given in the data appendix). POLITIC is a group of political variables: the GASTIL index of political rights and civil liberties, total government expenditure as a share of GDP, an indicator variable for the time the country achieved independence, and an indicator variable for whether the country had an external war between 1960 and 1985. The GASTIL index is our measure of dictatorship, while the timing of independence and external war are our measures of political stability.[10] Government expenditure can be interpreted in two ways: either as a dictatorship indicator, or as an indicator of stability (the greater the share of government in the economy, the smaller the effect of market forces on the economy. The government can redistribute tax revenues to reduce regional inequalities). CONTROL is a set of variables controlling for the size of the country; here the control variables used are the log of per capita GDP in constant US dollars, the log of the land area of the country, and the log of population. Finally, DUMMIES is the set of continent dummies.

One potential concern is the effect of using an estimated coefficient from a first stage regression as a dependent variable in a second stage regression. Lewis (2000) shows that the danger in doing so is that there could be measurement error in the first stage estimate, leading to inefficient estimates in the second stage. Heteroskedasticity might also arise if the sampling uncertainty in the (second stage) dependent variable is not constant across observations. He advocates the use of feasible GLS (FGLS) to overcome this problem. However, Baltagi (1995) points out that FGLS yields consistent estimates of the variances only if $T \rightarrow \infty$. This is clearly not the case for our sample; hence, FGLS results are not reported. In addition, Beck and Katz (1995) show that FGLS tends to underestimate standard errors, and that the degree of underestimation is worse the fewer the time periods in the panel. They propose an alternative estimator using panel corrected standard errors with OLS, which they show to perform better than FGLS in the sense that it does not underestimate the standard errors, but still takes into account the panel structure of the data and the fact that the data could be heteroskedastic and contemporaneously correlated across panels. The regressions using panel-corrected standard errors are those that are reported below.

[10] Following Ades and Glaeser (1995), we would liked to use as the measure of political instability, the number of attempted coups, assassinations or revolutions from the Barro and Lee (1994) data set. However, the years of their data do not match ours.

Table 5 presents the results using the OLS estimate of the Pareto exponent as the dependent variable (running the regression with the Hill estimate as the dependent variable yields almost identical results). The number of observations is somewhat less than the full sample because data is not available for all countries in all years. Columns (1) to (3) present the results using all available observations. Column (1) is the model without size and continent controls. Of the economic geography variables, transport cost

Table 5
Panel estimation of Eq. (5) (dependent variable = OLS coefficient of α)

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	OLS	OLS	OLS
Transport cost	− 0.6151	− 0.2763	− 0.4064	− 0.8702	− 0.5014	− 0.6386
	(3.00)***	(1.13)	(1.36)	(3.48)***	(2.56)**	(2.31)**
Trade (% of GDP)	− 0.0928	0.0370	− 0.0240	− 0.0459	0.0532	− 0.0177
	(1.71)*	(0.51)	(0.30)	(0.89)	(0.81)	(0.25)
Nonagricultural economic activity	− 0.2411	− 1.0137	− 0.5644	− 0.6002	− 1.4002	− 0.7731
	(0.73)	(2.37)**	(1.69)*	(1.99)**	(3.37)***	(2.10)**
Scale economies	0.4467	0.4462	0.4057	0.4993	0.4756	0.4284
	(2.25)**	(2.14)**	(1.77)*	(2.30)**	(2.14)**	(1.75)*
GASTIL index of dictatorship	− 0.0375	− 0.0145	− 0.0369	− 0.0307	− 0.0028	− 0.0284
	(1.96)*	(1.32)	(1.97)**	(1.59)	(0.21)	(1.67)*
Total government expenditure	0.7837	0.8013	0.7500	1.0097	0.9598	0.9154
	(6.08)***	(6.30)***	(2.56)**	(6.74)***	(5.68)***	(2.90)***
Timing of independence	− 0.0596	− 0.0686	− 0.1429	− 0.0974	− 0.0984	− 0.1692
	(2.36)**	(2.82)***	(3.96)***	(3.80)***	(3.52)***	(4.75)***
War dummy	0.2211	0.1410	0.1474	0.2437	0.1425	0.1659
	(3.71)***	(3.03)***	(2.36)**	(4.42)***	(3.54)***	(3.05)***
ln(land area)		0.0066	0.0288		0.0097	0.0239
		(0.39)	(1.59)		(0.64)	(1.33)
ln(population)		0.0548	0.0100		0.0459	0.0032
		(3.50)***	(0.49)		(2.81)***	(0.16)
ln(GDP per capita)		0.0959	0.0585		0.1053	0.0467
		(4.45)***	(2.05)**		(4.23)***	(1.34)
Africa dummy			0.1306			0.0967
			(1.24)			(0.97)
Asia dummy			0.2069			0.1898
			(1.85)*			(1.92)*
North America dummy			− 0.0655			− 0.0184
			(0.59)			(0.16)
South America dummy			− 0.1304			− 0.1459
			(1.30)			(1.32)
Oceania dummy			− 0.0804			− 0.0375
			(1.02)			(0.50)
Constant	1.1638	− 0.1307	0.3961	1.4082	0.1885	0.8256
	(3.96)***	(0.24)	(0.69)	(5.69)***	(0.38)	(1.57)
R-squared	0.4702	0.5778	0.6587	0.5403	0.6254	0.7007
Observations	79	79	79	72	72	72
Countries	44	44	44	40	40	40

z statistics in parentheses. *Significant at 10%; **significant at 5%; ***significant at 1% OLS with panel-corrected standard errors results reported.

and the degree of scale economies are highly significant. However, they enter with the opposite signs to what we expect from theory. The political variables fare better, with all variables being significant. The coefficients on the GASTIL index of political rights and the timing of independence enter with the theoretically predicted signs. However, the war dummy enters with the wrong sign; this could be explained by suggesting that large cities are more highly prized targets in a war, so that people will tend to leave large cities. Total government expenditure enters with a very strong positive coefficient, which indicates that greater government expenditure is associated with a more even distribution of cities.

Including controls for country size and continent dummies (columns (2) and (3)) shows that the results of the economic geography variables are not robust. This contrasts with the strong robustness of the political variables. The only robustly significant economic geography variable is the degree of scale economies, and this enters with the opposite sign to what we would expect from existing theoretical models. The political variables remain highly significant. Therefore, our results suggest that politics plays a more important role than economy-wide economic geography variables in explaining variation in the Pareto exponent across countries.

Columns (4) to (6) of Table 5 present results of the same regression, run for the sample excluding former communist countries, in the belief that in the rest of the world, free market forces play a more important role than political forces. Dropping the former communist countries improves the overall fit of the estimated equation, since *R*-squared increases. The signs of all significant variables remain unchanged. We do indeed find that the economic geography variables have increased significance; however, as noted before they enter with the wrong sign vis-à-vis the theoretical model. Also, while the GASTIL index becomes less significant, the rest of the political variables remain highly significant although the war dummy continues to enter with the wrong sign.

Of the control variables and the continent dummies, not much need be said. In the full specifications (3) and (6), they are mainly insignificant. This indicates that the economic geography and the political variables account for most of the variation in the Pareto exponent across continents noted in Section 4.

Comparing our results to previous findings, we find that our results for columns (3) and (6) of Table 5 (including all the variables and controls) are broadly in line with those of Alperovich (1993). However, we get somewhat different results from those of Rosen and Resnick, as they find that the Pareto exponent is positively related to per capita GNP, total population and railroad density, and negatively related to land area. One likely explanation for this difference in results is that our specification is more complete than the one used by Rosen and Resnick; this can also be seen from the larger R^2 that we obtain (0.66) compared to their largest R^2 of 0.23.

6. Conclusion

This paper set out to test Zipf's Law for cities, using a new data set and two alternative methods—OLS and the Hill estimator. Using either method, we reject Zipf's Law much more often than we would expect based on random chance. Using OLS, we reject the

Zipf's Law prediction that the Pareto exponent is equal to 1, for 53 of the 73 countries in our sample. This result is consistent with the classic study by Rosen and Resnick (1980), who reject Zipf's Law for 36 of the 44 countries in their sample. We get a different result using the Hill estimator, where we reject Zipf's Law for a minority of countries (30 out of 73). Therefore, the results we obtain depend on the estimation method used, and in turn, the preferred estimation method would depend on our sample size and on our theoretical priors—whether or not we believe that Zipf's Law holds.

One new result which we obtain is that the average value of the Pareto exponent for urban agglomerations is less than 1 (and significantly so for over half the sample using OLS); Zipf's Law fails for urban agglomerations. This is a new result, as previous work (e.g. Rosen and Resnick, 1980) have tended to find that the Pareto exponent is equal to 1 if data on urban agglomerations are used. This could be an indication of the increasing suburbanisation of large cities in the last 20 years, which would show up as increasing inequality between urban agglomerations.

In attempting to explain the observed variations in the value of the Pareto exponent, we sought to relate the value of the Pareto exponent to several variables used in models of the size distribution of cities. The data appears to be more consistent with a model of political economy as the main determinant of the size distribution of cities. Economic geography variables are important as well, but tend to enter with coefficients which are opposite in sign to theoretical predictions.

Acknowledgements

I am very grateful to Alejandro Cunat, Gilles Duranton, Xavier Gabaix, Henry Overman, Steve Redding, Martin Stewart, Tony Venables, David Cuberes, seminar participants at the CEP International Economics Field Seminar, and two anonymous referees for valuable comments and suggestions, and to Paul Cheshire and the LSE Research Lab Data Library for access to data. Financial support from the Overseas Research Student Award Scheme and the LSE are gratefully acknowledged. All remaining errors are mine.

Appendix A. Data appendix

This appendix describes the variables used in the regressions (the full list of data sources is given in the text). Unless otherwise mentioned, all data are from the World Bank World Development Indicators CD-ROM.

Scale economies is the degree of scale economies, constructed as the share of industrial output in high-scale industries where the definition of high-scale industries is obtained from Pratten (1988). The method used is to obtain the output of three-digit industries from the UNIDO 2001 Industrial Statistics Database, then use Table 5.3 in Pratten (1988) to identify the industries that have the highest degree of scale economies, and divide the output of these industries by total output of all manufacturing industries.

Transport cost is transport cost, measured using the inverse of road density (total road mileage divided by land area). Source: United Nations WDI CD-ROM and International Road Federation World Road Statistics.

Nonagricultural economic activity is the share of nonagricultural value-added in GDP.

GASTIL index is a combination of measures for political rights and civil liberties, and ranges from 1 to 7, with a lower score indicating more freedom. Source: Freedom House.

Total government expenditure is total government expenditure as a share of GDP.

War dummy is a dummy indicating whether the country had an external war between 1960 and 1985. Source: Gallup et al. (1999).

Timing of independence is a categorical variable taking the value 0 if the country achieved independence before 1914, 1 if between 1914 and 1945, 2 if between 1946 and 1989, and 3 if after 1989. Source: Gallup et al. (1999).

Trade (% of GDP) is the ratio of total international trade in goods and services to total GDP.

ln(GDP per capita) is the log of per capita GDP, measured in constant US dollars.

ln(land area) is the log of land area, measured in square kilometers.

ln(population) is the log of population.

References

Ades, A.F., Glaeser, E.L., 1995. Trade and circuses: explaining urban giants. Quarterly Journal of Economics 110, 195–227.

Alperovich, G.A., 1984. The size distribution of cities: on the empirical validity of the rank-size rule. Journal of Urban Economics 16, 232–239.

Alperovich, G.A., 1988. A new testing procedure of the rank size distribution. Journal of Urban Economics 23, 251–259.

Alperovich, G.A., 1993. An explanatory model of city-size distribution: evidence from cross-country data. Urban Studies 30, 1591–1601.

Auerbach, F., 1913. Das gesetz der bevolkerungskoncentration. Petermanns Geographische Mitteilungen 59, 74–76.

Axtell, R.L., Florida, R., 2000. Emergent cities: a microeconomic explanation of Zipf's law, mimeo. The Brookings Institution.

Baltagi, B.H., 1995. Econometric Analysis of Panel Data. Wiley, Chichester, England.

Barro, R.J., Lee, J.-W., 1994. Data Set for a Panel of 138 Countries. NBER, Cambridge, MA.

Beck, N., Katz, J.N., 1995. What to do (and not to do) with time-series cross-section data. American Political Science Review 89, 634–647.

Black, D., Henderson, J.V., 2003. Urban Evolution in the USA. Journal of Economic Geography 3, 343–372.

Brakman, S., Garretsen, H., van Marrewijk, C., van den Berg, M., 1999. The return of Zipf: towards a further understanding of the rank-size distribution. Journal of Regional Science 39, 183–213.

Brinkhoff, T., 2004. City Populations [online]. Available from http://www.citypopulation.de[Accessed May-December 2001].

Cheshire, P., 1999. Trends in sizes and structures of urban areas. In: Cheshire, P., Mills, E.S. (Eds.), Handbook of Regional and Urban Economics, vol. 3. Elsevier, Amsterdam, pp. 1339–1372.

Cordoba, J.C., 2003. On the Distribution of City Sizes, mimeo. Rice University.

Dobkins, L.H., Ioannides, Y.M., 2000. Dynamic evolution of the size distribution of US cities. In: Huriot, J.-M., Thisse, J.-F. (Eds.), Economics of Cities: Theoretical Perspectives. Cambridge Univ. Press, Cambridge, UK, pp. 217–260.

Duranton, G.. 2002. City Size Distribution as a Consequence of the Growth Process, mimeo. London School of Economics.

Encyclopaedia Britannica Inc. 2001.Encyclopaedia Britannica Book of the Year. Encyclopaedia Britannica, Chicago, IL.

Fujita, M., Krugman, P., Venables, A.J., 1999. The Spatial Economy: Cities, Regions and International Trade. MIT Press, Cambridge, MA.

Gabaix, X., 1999. Zipf's Law for cities: an explanation. Quarterly Journal of Economics 114, 739–767.

Gabaix, X., Ioannides, Y.M., 2004. The evolution of city size distributions in press. In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, 2341–2378.

Gallup, J.L., Sachs, J.D., Mellinger, A.D., 1999. Geography and Economic Development, Center for International Development Working Paper No. 1, Harvard University.

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

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Annals of Statistics 3, 1163–1174.

Ioannides, Y.M., Overman, H.G., 2003. Zipf's law for cities: an empirical investigation. Regional Science and Urban Economics 33, 127–137.

Krugman, P., 1991. Geography and Trade. MIT Press, Cambridge, MA.

Krugman, P., 1996. Confronting the mystery of urban hierarchy. Journal of the Japanese and International Economies 10, 399–418.

Lewis, J.B., 2000. Estimating Regression Models in which the Dependent Variable is Based on Estimates with Application to Testing Key's Racial Threat Hypothesis, mimeo. Princeton University.

Mills, E.S., Becker, C.M., 1986. Studies in Indian Urban Development. Oxford Univ. Press, Oxford.

Pratten, C., 1988. A survey of the economies of scale, in commission of the european communities: research on the "cost of non-Europe". Studies on the Economics of Integration, vol. 2. Office for Official Publications of the European Communities, Luxembourg, 11–165.

Ramsey, J.B., 1969. Tests for specification error in classical linear least squares analysis. Journal of the Royal Statistical Society. Series B 31, 350–371.

Reed, W.J., 2001. The Pareto, Zipf and other power laws. Economics Letters 74, 15–19.

Rosen, K.T., Resnick, M., 1980. The size distribution of cities: an examination of the Pareto law and primacy. Journal of Urban Economics 8, 165–186.

Rossi-Hansberg, E., Wright, M.L.J., 2004. Urban Structure and Growth, mimeo. Stanford University.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Zipf, G.K., 1949. Human Behaviour and the Principle of Least Effort. Addison-Wesley, Reading, MA.