# ZIPF'S LAW: MAIN ISSUES IN EMPIRICAL WORK

**Rafael GONZÀLEZ-VAL***

***Abstract*** *- One of the stylised facts in Urban Economics is Zipf's law, according to which city size distribution in many countries can be approximated by a Pareto distribution, whose exponent is equal to one. In this paper we point out the three main issues in any empirical work on city size distribution and Zipf's law: city definition, sample size, and the choice of the estimator. We review the more recent developments, especially those related to the relationship between the geographical unit chosen and sample size, and the features of the different methods to estimate the Pareto exponent. We illustrate the arguments, providing empirical examples using actual data from the city size distribution in the United States.*

***Keywords:*** CITY SIZE DISTRIBUTION; ZIPF'S LAW

***JEL Classification:*** R00, R12, C16

***Acknowledgements*** *- The author acknowledges financial support from the Spanish Ministerio de Educación y Ciencia (ECO2009-09332 and ECO2010-16934 projects) the DGA (ADETRE research group) FEDER and the Generalitat (2009SGR102).*

* Universitat de Barcelona & Institut d'Economia de Barcelona.

## INTRODUCTION

Since the seminal work by Auerbach in 1913 researchers from many fields (economics, statistics, physics and geography) have been fascinated with the striking empirical regularity that establishes a linear and stable relationship between city size and rank. Later this empirical regularity became known as Zipf's law (Zipf, 1949), although what Zipf's law establishes is just a particular case of that linear relationship where the second-largest city in a country is exactly half the size of the largest one, the third-largest city is a third the size of the largest, etc. Over the last 100 years there have been a lot of studies testing the validity of this law (see the surveys by Cheshire, 1999, and Nitsch, 2005) for many different countries; to mention only a few, there are studies for France (Guérin-Pace, 1995), Greece (Petrakos et al., 2000), China (Song and Zhang, 2002), Malaysia (Soo, 2007) and the United States (Ioannides and Overman, 2003; Black and Henderson, 2003).

There has been a revival of interest in city size distributions and Zipf's law in the last years from urban economists, especially after the New Economic Geography by Krugman. Starting from the wide empirical literature, some theoretical models have been proposed recently to explain the law, with different economic foundations: productivity or technology shocks (Duranton, 2007; Rossi-Hansberg and Wright, 2007) or local random amenity shocks (Gabaix, 1999). These models justify Zipf's law analytically, associate it directly with an equilibrium situation, and connect it to proportionate city growth (Gibrat's law), another well-known empirical regularity which postulates that the growth rates of cities tend to be independent of their initial sizes. In both the theoretical and empirical literature, Zipf's law is seen as a reflection of a steady-state situation.

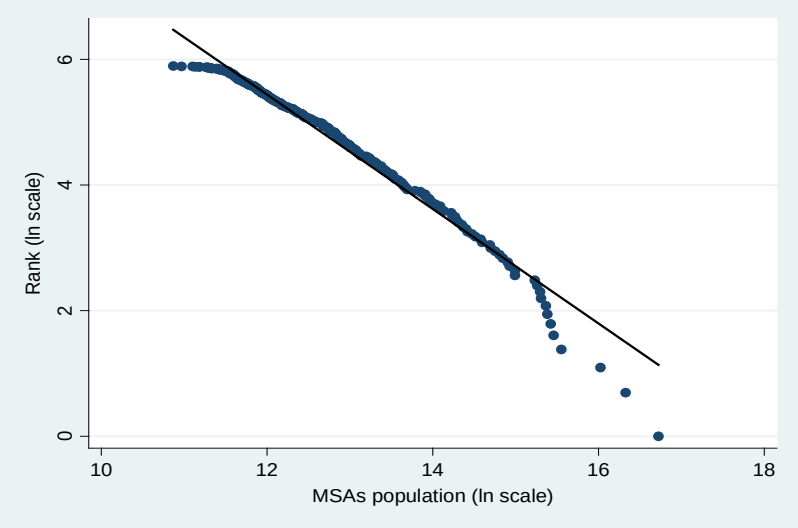
To obtain Zipf's law we must assume a particular statistical function for describing city size distribution, the Pareto distribution. Let $S$ be the city size (population); if it is distributed following a Pareto distribution, also known as power law, the density function is $p(S) = \frac{a\underline{S}^a}{S^{a+1}} \quad \forall S \geq \underline{S}$ and the cumulative density function $P(S)$ is $P(S) = 1 - \left(\frac{\underline{S}}{S}\right)^a \quad \forall S \geq \underline{S}$ (Eeckhout 2004) where $a > 0$ is the Pareto exponent and $\underline{S}$ is the population of the city at the truncation point. The relationship with the empirically observed rank $R$ (1 for the largest, 2 for the second largest, and so on) is $R = \underline{N} \cdot (1 - P(S)) = \underline{N} \cdot \left(\frac{\underline{S}}{S}\right)^a$, where $\underline{N}$ is the number of cities above the truncation point (minimum population threshold $\underline{S}$). Making $A = \underline{N} \cdot \underline{S}^a$ we obtain the simple expression $R = A \cdot S^{-a}$. Taking natural logarithms, we obtain the linear specification that is usually estimated:

$$\ln R = \ln A - a \ln S + u \tag{1}$$

where $u$ represents a standard random error ($E(u)=0$ and $Var(u)=\sigma^2$) and ln $A$ is a constant $\ln A = \ln \underline{N} + a \ln \underline{S}$. Zipf's law is an empirical regularity, which appears when the Pareto exponent of the distribution is equal to unity ($\hat{a}=1$) 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 (the law is also known as the rank–size rule, although they are not exactly the same; see Gabaix and Ioannides, 2004).

Moreover, the greater the coefficient, the more homogeneous the city sizes. A growing evolution over time means a process of convergence in city sizes. Also, the smaller the coefficient the less homogeneous the city sizes, and a decreasing evolution would mean a process of divergence. Therefore it is interesting to study not only the value of the exponent, but also its evolution over time, although changing boundaries make it difficult to carry out a long-term analysis with consistent boundaries. Pareto exponents are not the perfect instrument to analyse the degree of inequality within the distribution because they impose a specific size distribution (Pareto) although there is a statistical relationship between Zipf's law and the main concentration indices: Gini, Bonferroni, Amato and the Hirschman–Herfindahl index (Naldi, 2003).

**Figure 1. Rank-Size Plot (ln scale) US MSAs in 2010**




*Note: The slope of the line fitted by OLS is 0.911 (see Table 4).* $R^2 = 0.978$.

Eq. (1) can be represented as a graph, called Zipf plot in the specialised literature. Figure 1 shows the Zipf plot for the year 2000 using US Metropolitan Statistical Areas. Data are fitted by a power law, and its exponent is estimated by using the OLS estimator (more on this in the following). Eq. (1) provides a very good fit to the real behaviour of the distribution with an estimated Pareto exponent of 0.911. Similar graphs can be found in Krugman (1996) and Gabaix

(1999); they use data from metropolitan areas from the Statistical Abstract of the United States and conclude that for 1991 Pareto's exponent is exactly equal to 1.005, thus finding evidence supporting Zipf's law for this year in the US.

There are two excellent surveys that cover most of the issues related to Zipf's law, Cheshire (1999) and Gabaix and Ioannides (2004). Newman (2006) also reviews some of the empirical evidence from a more interdisciplinary point of view. This paper is simpler and shorter than these previous surveys and its aim is very specific, to point out the three main issues that every researcher faces in any empirical work on Zipf's law: the choice of the spatial units, the sample size, and the choice of the estimator.

Next section provides the explanation and consequences for these three issues. Subsections 1.1 and 1.2 expand the discussion in González-Val (2010) and González-Val (2011) about data selection. To illustrate the arguments, empirical examples using US data are provided in Section 2.

## 1. THE THREE MAIN ISSUES

There are three main issues in every study on city size distribution: city definition (Rosen and Resnick, 1980; Cheshire, 1999; Soo, 2005), sample size (Parr and Suzuki, 1973; Rosen and Resnick, 1980; Eeckhout, 2004) and the choice of the estimator (Gabaix and Ioannides, 2004; Nishiyama et al., 2008; Gabaix and Ibragimov, 2011).

### 1.1. City definition: What is a city?

Any study on city size distribution and Zipf's law faces the problem of what is meant by the term "city", as there are various ways to define a city. This first decision, the choice of the spatial unit, is not trivial. As Rosen and Resnick (1980) point out, "whether a study uses urban place,s legal cities or urban agglomerations may affect the value of the observed Pareto exponent and the closeness of the fit."

The two basic alternatives are the administratively defined cities (legal cities) and the metropolitan areas. Both units have advantages. As Glaeser and Shapiro (2003) indicate, metro areas represent urban agglomerations, covering huge areas that are meant to capture labour markets. Metropolitan areas are attractive because they are more natural economic units. Legal cities are political units that usually lie within metropolitan areas and their boundaries make no economic sense. But some factors, such as human capital spillovers, are thought to operate at a very local level (Eeckhout, 2004).

Which one is the best alternative? The answer depends basically on two aspects. First data availability; metro area definitions are not available in some countries (e.g. many European countries do not have official definitions for metropolitan areas), while data on legal cities are probably easier to get through official census and national statistical services. Moreover, if the objective is to carry out a dynamic analysis, both units can have problems of changing boundaries over time. Secondly and probably more important, is what Eeckhout (2004)

calls "the research objective." What do you want to study? The economic area of influence of labour markets and big infrastructure projects like airports exceeds the boundaries of single legal cities, while the geographical influence of factors, such as public services (schools public transportation etc.) and local externalities is more reduced. However, from a statistical point of view, there is a reason that recommends the use of cities (untruncated data), rather than metro areas; the next subsection deals with this issue.

### 1.2. Sample size: How many cities?

Usually sample size depends on data availability. However, in many cases research focuses only on upper-tail distribution. This approach is reasonable as the largest cities represent most of the population of a country and the behaviour of the upper-tail distribution can be different from that of the entire distribution (Levy, 2009). In any case, some kind of rule is necessary to decide the number of cities in the sample.

Cheshire (1999) summarises three possible criteria for sample size selection: a fixed number of cities, a size threshold (Rosen and Resnick, 1980) or a size above which the sample accounts for some given proportion of the country's population (Wheaton and Shishado, 1981; Black and Henderson, 2003). Against this background, Eeckhout (2004) demonstrates the statistical importance of considering the whole sample. In proposition 1 of his paper, he states that "if the underlying distribution is the lognormal distribution, then the estimate of the parameter $\hat{a}$ of the Pareto distribution is increasing in the truncation city size ($d\hat{a}/d\underline{S} > 0$) and decreasing in the truncated sample population ($d\hat{a}/d\underline{N} < 0$)." As we will see in Section 3, estimation results using subsamples support that proposition. This also implies that the Pareto exponent for spatial units constructed aggregating smaller units will be higher (and closer to Zipf's law).

Therefore, if any truncation point is imposed the estimates of the Pareto exponent may be biased. However when all cities are considered often nonlinear behaviours appear leaving the fulfilment of Zipf's law only for the largest cities. New statistical distributions have been proposed instead of the Pareto distribution to explain the behaviour of the entire distribution: lognormal distribution (Parr and Suzuki, 1973; Eeckhout, 2004), q-exponential distribution (Malacarne et al., 2001; Soo, 2007), the double Pareto lognormal distribution (Reed, 2002; Giesen et al., 2010), or even a new distribution function that switches between a lognormal and a power distribution (Ioannides and Skouras, 2009).

Moreover, the geographical unit chosen is also closely related to sample size. For example, if data come from metropolitan areas, you are imposing an implicit truncation point because in many countries, metro areas are defined according to some minimum population threshold (e.g. in the US, the central city needs to have 50 000 or more inhabitants and a total metropolitan population of at least 100 000). So, once the decision about the geographical unit is

made, the sample size (and the truncation point) may be determined as a consequence of the spatial definition of units.

### 1.3. Estimators of the Pareto exponent: How do I estimate $\hat{a}$ ?

Obviously, there are statistical tests to check whether the distribution is Pareto (Urzúa, 2000; Malevergne et al., 2011) without applying Eq. (1), but as Gabaix and Ioannides (2004) suggest, "estimate don't test." They argue that the main question should be how well a theory (Zipf's law) fits rather than whether or not it fits perfectly.

From Figure 1, it seems easy to estimate the Pareto exponent, because it is just the slope of the line fitted by OLS. This has been the method used in many works, until recent years (Nitsch, 2005, carried out a meta-analysis, with the results of a list of studies until the year 2002). However, it is not the only option; more methods have been proposed to try to solve some of the problems associated with OLS and accommodate the estimates of the exponent to nonlinear behaviours.

Table 1 summarises the main methods. The first one is the simple OLS estimator and the baseline equation is Eq. (1). The OLS estimate presents some problems (Nishiyama et al., 2008). The main one is that the Hill (Maximum Likelihood) estimator is more efficient, if the underlying stochastic process is really a Pareto distribution (Gabaix and Ioannides, 2004; Goldstein et al., 2004). Furthermore, as Gabaix and Ioannides (2004) point out, this "OLS regression underestimates the true standard error on the estimated coefficient", thus "taking the OLS estimates of the standard errors at face value will lead one to reject Zipf's law much too often." Finally, this procedure is strongly biased in small samples (Gabaix and Ibragimov, 2011). To correct this last pitfall, Gabaix and Ibragimov (2011) propose specifying Eq. (1), by subtracting 1/2 from the rank, to obtain an unbiased estimation of the exponent. Their numerical results demonstrate the advantage of this approach over the standard OLS estimation procedures, especially in small samples. However, again if the underlying stochastic process is really a Pareto distribution, the Hill estimator is more efficient.

The Hill estimator assumes the null hypothesis of the power law, so the procedure does not estimate Eq. (1); it is based on the maximisation of the log-likelihood function. The problem in this case is that, when the size distribution of cities does not follow a Pareto distribution, the Hill estimator may be biased (Soo, 2005). In particular, Gabaix and Ibragimov's preliminary results suggest that their specification is more robust than Hill's estimator under deviations from power laws.

The next two methods try to incorporate nonlinear behaviours basically augmenting Eq. (1) introducing new terms. Rosen and Resnick (1980) added $b(\ln S)^2$ to the specification; a $\hat{b}$ coefficient significantly different from zero is interpreted as a deviation from the Pareto distribution and Zipf's law. Fan and

Casetti (1994) introduced $c \cdot S \ln S$ in the equation; again, this term tries to capture departures from Zipf's law and how deviations depend on the size of the city. These specifications have fallen into disuse, because the significance of these coefficients may be spurious. Gabaix and Ioannides (2004) perform Monte Carlo simulations and find that, with the OLS regression in Rosen and Resnick's equation, one will often find a statistically significant coefficient $\hat{b}$, even if Zipf's law holds perfectly by construction.

**Table 1. Estimators of the Pareto Exponent**

| Method | Equation | Estimator |
|---|---|---|
| Simple OLS | $\ln R = \ln A - a \ln S + u$ | OLS |
| Gabaix and Ibragimov (2011) | $\ln\left(R - \frac{1}{2}\right) = \ln A - a \ln S + u$ | OLS |
| Hill (Maximum likelihood) | Log-likelihood function | $\hat{a} = \frac{N-1}{\sum_{i=1}^{N-1}\left(\ln S_i - \ln S_N\right)}$ |
| Rosen and Resnick (1980) | $\ln R = \ln A - a \ln S + b(\ln S)^2 + u$ | OLS |
| Fan and Casetti (1994) | $\ln R = \ln A - a \ln S + c \cdot S \ln S + u$ | OLS |
| Ioannides and Overman (2003) | $a(S) = 1 - 2 \cdot \frac{\mu(S)}{\sigma^2(S)} + \frac{\partial \sigma^2(S)/\sigma^2(S)}{\partial S/S}$ | Nonparametric |

Finally, Ioannides and Overman (2003) developed a nonparametric method to estimate the Pareto exponent, based on the statistical explanation of Zipf's law for cities, offered by Gabaix (1999). The nonparametric estimate of the exponent $a(S)$ is calculated, using nonparametric estimates of the mean growth rates $\mu(S)$ and of the variance of growth rates $\sigma^2(S)$. This allows us to test whether Gibrat's law holds. The drawback of this procedure is that it requires a big sample size, because when there are few observations sparsity in the data, it can make the estimates of the Zipf exponent fluctuate considerably.

## 2. AN EMPIRICAL EXERCISE: US CITY SIZE DISTRIBUTION

### 2.1. Data

The US city size distribution has been the focus of attention of many researchers: Dobkins and Ioannides (2000, 2001), Overman and Ioannides (2001), Black and Henderson (2003), Ioannides and Overman (2003), Eeckhout (2004) and González-Val (2010), among others. In this wide literature, different spatial units, time periods, and statistical and econometrics methods are considered.

The US Census Bureau offers information for many different geographical levels, so the choice is not only between legal cities and metro areas. There are studies using data from states (Soo, 2011), counties (Beeson et al., 2001; Michaels et al., 2012), minor civil divisions (Michaels et al., 2012), metropolitan areas (Ehrlich and Gyourko, 2000; Dobkins and Ioannides, 2000, 2001;

Black and Henderson, 2003; Ioannides and Overman, 2003), places (Eeckhout, 2004, 2009; Levy, 2009; Giesen et al., 2010; González-Val, 2010), urbanized areas (Garmestani et al., 2005; Garmestani et al., 2008) or the economic areas recently defined by Rozenfeld et al. (2011), using the city clustering algorithm. Berry and Okulicz-Kozaryn (2011) argue that the best units are the economic areas, defined by the Bureau of Economic Analysis.

For illustrative purposes, we will only focus on some of these units: states, counties, metropolitan areas, urbanized areas and places. Table 2 shows the descriptive statistics for the year 2000. The data source is the US Census Bureau (www.census.gov). To show the effect of sample size, we consider three different samples: all units, top 100 and top 250. Obviously, when the minimum population size at the truncation point ($\underline{S}$) increases, the smaller the sample size is, in any case.

**Table 2. Descriptive Statistics by Unit**

| Units | Sample Size ($N$) | % of Total US Population | Mean | Standard Deviation | Minimum ($\underline{S}$) | Maximum ($\underline{S}$) |
|---|---|---|---|---|---|---|
| States | All (50) | 99.78% | 5 615 899.48 | 6 186 487.80 | 493 782 | 33 871 648 |
| Counties | All (3 114) | 99.79% | 90 183.01 | 293 622.20 | 67 | 9 519 338 |
| | Top 100 | 42.39% | 1 193 056.72 | 1 111 938.54 | 556 678 | 9 519 338 |
| | Top 250 | 61.03% | 687 006.92 | 817 484.23 | 226 778 | 9 519 338 |
| Metropolitan Statistical Areas | All (362) | 82.64% | 642 486.02 | 1 485 743.37 | 52 457 | 18 323 002 |
| | Top 100 | 64.65% | 1 819 293.54 | 2 467 957.00 | 446 997 | 18 323 002 |
| | Top 250 | 78.23% | 880 572.32 | 1 736 721.06 | 145 666 | 18 323 002 |
| Urbanized Areas | All (463) | 66.40% | 403 591.39 | 1 237 112.01 | 665 | 17 799 861 |
| | Top 100 | 53.39% | 1 502 442.81 | 2 360 044.21 | 292 637 | 17 799 861 |
| | Top 250 | 62.15% | 699 611.67 | 1 627 212.92 | 95 766 | 17 799 861 |
| Places | All (25 358) | 74.17% | 8 231.54 | 68 390.23 | 1 | 8 008 278 |
| | Top 100 | 20.19% | 568 308.67 | 906 644.82 | 194 973 | 8 008 278 |
| | Top 250 | 27.38% | 308 237.96 | 610 371.80 | 99 216 | 8 008 278 |

*Note: Data in 2000. Total US population data are taken from the US Census reau. http://www.census.gov/population/censusdata/table-4.pdf.*

States are the primary legal subdivision of the United States; there are 50 states and they are very big units with similar sizes to some European countries.

Therefore, states do not seem to be the most appropriate approximation to urban agglomerations, although empirical studies on Zipf's law exist even at the country level (Rose, 2006; González-Val and Sanso-Navarro, 2010). There are good reasons to study state size distribution. As Soo (2011) argues, states cover the entire population of the country, whereas cities do not and states are the administrative level at which many policies vary. Recent works relate the ful-

fillment of Zipf's law in city size distribution, at the regional and national level (Gabaix, 1999; Giesen and Südekum, 2011).

Counties are the primary legal subdivision of most states. At first glance, counties do not seem to impose a truncation point, as there were 3, 114 counties in the year 2000 with populations between 67 and 9 519, 338. Moreover, counties cover the entire population and land area of the country as they are administrative subdivisions of states. Beeson et al. (2001) explain the additional advantages of counties, while Michaels et al. (2012) point out some of the disadvantages, including that counties often pool together urban centers with their surrounding countryside, clouding the distinction between urban and rural areas.

Metropolitan Statistical Areas (MSAs) are geographic entities, defined by the federal Office of Management and Budget, based on the concept of a core area with a large population nucleus (central city), plus adjacent communities having a high degree of economic and social integration with that core. To qualify as an MSA, the presence of a central city with 50, 000 or more inhabitants is required or the presence of an urbanized area and a total population of at least 100, 000 (75, 000 in New England). That is why, there were only 362 MSAs in the year 2000 and the minimum population in the sample is 52 457, always higher than the minimum population threshold. MSAs are multi-county units; the county or counties containing the largest city and surrounding densely settled territory are the central counties of the MSA.

**Table 3. Incorporated Places: Number of Cities and Descriptive Statistics by Year**

| 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 |

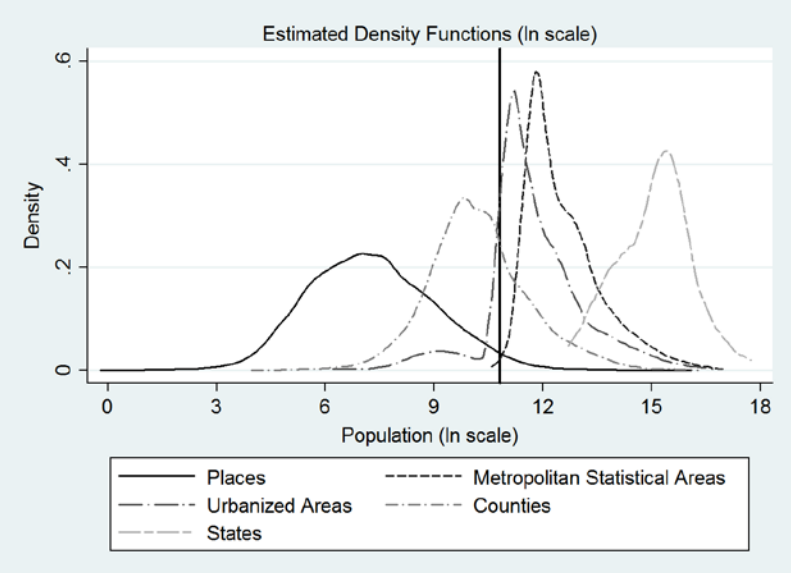
*Note: Excluding Alaska Hawaii and Puerto Rico.*
*Source: Table 1 in González-Val (2010).*

An urbanized area consists of a central place(s) and adjacent territory, with a general population density of at least 1, 000 people per square mile of land area, that together have a minimum residential population of at least 50, 000 people. An urbanized area comprises a central place and the urban fringe (Garmestani et al., 2005; Garmestani et al., 2008). As in the case of MSAs, ur-

banized areas are meant to capture economic areas, although they are smaller than MSAs.

Finally, places are concentrations of populations, either legally bounded as an incorporated place or identified as a Census Designated Place. Incorporated places are the legal cities, incorporated under state Law as cities towns (except in the states of New England, New York and Wisconsin), boroughs (except in Alaska and New York) or villages, which have legally established limits powers and functions. These places have been used recently in the empirical analyses of American city size distribution (Eeckhout, 2004, 2009; Levy, 2009; Giesen et al., 2010; González-Val, 2010) and their main advantage is that they do not impose any truncation point (populations range from 1 to 8, 008, 278 inhabitants). As an example of a long-term analysis of Zipf's law, we will also consider a sample of all incorporated places, without any size restriction for each decade of the 20th century. From a long-term perspective, units such as MSAs or urbanized areas are excluded, as they were introduced in the middle of the 20th century. The data are the same as those used by González-Val (2010); Table 3 shows the number of cities and descriptive statistics by year.

**Figure 2. Estimated Density Functions (ln scale) by Unit**




*Note: Data in 2000. The vertical black line represents the minimum population threshold of 50 000 inhabitants (10.82 in logarithmic scale).*

Figure 2 shows how the choice of the geographical definition of city is related to the sample size. The vertical black line represents the minimum threshold of 50, 000 inhabitants and we can observe how MSAs and urbanized areas distributions are to the right of that cut-off point, while most of the places and counties distributions are to the left. This minimum population threshold represents the main inconvenience of MSAs and urbanized areas (besides changing

boundaries, which is a common problem in long-term analysis for all units, except states), because it implies that only the largest cities (upper-tail distribution) are represented in the samples, leading to biased results in the estimates of the Pareto exponent (Eeckhout, 2004), while other units, such as places or counties, cover the entire distribution, including even the smallest units.

### 2.2. Results

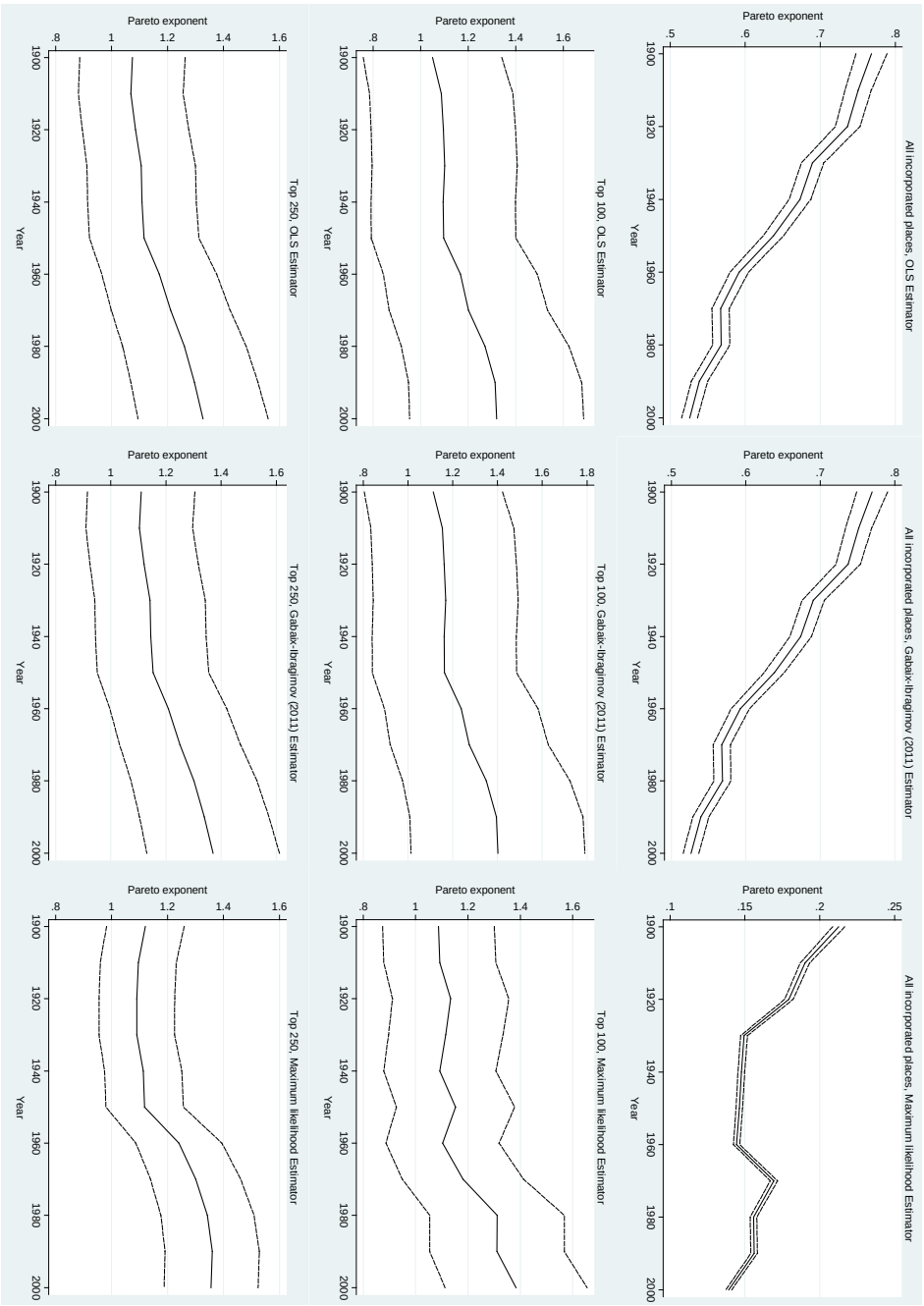
Table 4 shows the estimates of the Pareto exponent for the different geographical units, using the simple OLS regression, Gabaix and Ibragimov's (2011) Rank $-1/2$ estimator and the Hill (maximum likelihood) estimator. We consider three sample sizes: all units, top 100, and top 250. We can obtain several conclusions from these results.

**Table 4. Estimated Pareto Exponents by Unit**

| Units | Sample Size (*N*) | OLS | Gabaix-Ibragimov (2011) | Hill |
|---|---|---|---|---|
| States | All (50) | 0.797 | 0.857 | 0.513 |
| | | (0.159) | (0.171) | (0.073) |
| Counties | All (3 114) | 0.660 | 0.662 | 0.166 |
| | | (0.017) | (0.017) | (0.003) |
| | Top 100 | 1.763 | 1.871 | 1.730 |
| | | (0.249) | (0.265) | (0.173) |
| | Top 250 | 1.444 | 1.487 | 1.206 |
| | | (0.129) | (0.133) | (0.076) |
| Metropolitan Statistical Areas | All (362) | 0.911 | 0.930 | 0.582 |
| | | (0.068) | (0.069) | (0.031) |
| | Top 100 | 1.117 | 1.177 | 1.022 |
| | | (0.158) | (0.166) | (0.102) |
| | Top 250 | 0.974 | 1.001 | 0.877 |
| | | (0.087) | (0.090) | (0.055) |
| Urbanized Areas | All (463) | 0.630 | 0.641 | 0.191 |
| | | (0.041) | (0.042) | (0.009) |
| | Top 100 | 0.984 | 1.036 | 0.914 |
| | | (0.139) | (0.147) | (0.091) |
| | Top 250 | 0.897 | 0.923 | 0.851 |
| | | (0.080) | (0.083) | (0.054) |
| Places | All (25 358) | 0.526 | 0.526 | 0.137 |
| | | (0.005) | (0.005) | (0.001) |
| | Top 100 | 1.340 | 1.424 | 1.410 |
| | | (0.190) | (0.201) | (0.141) |
| | Top 250 | 1.352 | 1.397 | 1.377 |
| | | (0.121) | (0.125) | (0.087) |

*Notes: Data in 2000. Pareto exponents are estimated using OLS Gabaix and Ibragimov's (2011) Rank - 1/2 estimator and the Hill (maximum likelihood) estimator. Values in parenthesis are the standard errors; in the case of OLS and Gabaix and Ibragimov's estimators, they are calculated applying Gabaix and Ioannides's (2004) corrected standard errors:* $\text{GI s.e.} = \hat{a} \cdot (2/N)^{1/2}$ *, where N is the sample size.*

**Figure 3. Evolution of the Estimated Pareto Exponents (Incorporated Places) by Year**



First, the results using simple OLS regressions (Eq. 1) and Gabaix and Ibragimov's estimator are equal (or almost the same), when the sample size is large enough; see the results for all places (25, 358 observations), all counties (3 114) or all urbanized areas (463). The reason is that Gabaix and Ibragimov's estimator performs better in small samples but when the sample size is large, there are not significant differences. Therefore, results are different when only the upper-tail distribution is considered; this is the case for states and top samples. The gap between both estimators in those cases is the bias from simple OLS regressions.

Second, the values estimated using the Hill estimator are always much lower, probably indicating that the null hypothesis of the power law is not fulfilled. Remember that the Hill estimator is more efficient only if the underlying stochastic process is really a Pareto distribution, but when this assumption does not hold, the Hill estimator may be biased (Soo, 2005). The latter seems to be the case, given the much lower values estimated in most cases. However, for the top samples the differences with the other estimators are smaller, pointing to clearer power law behaviour in the upper-tail distribution.

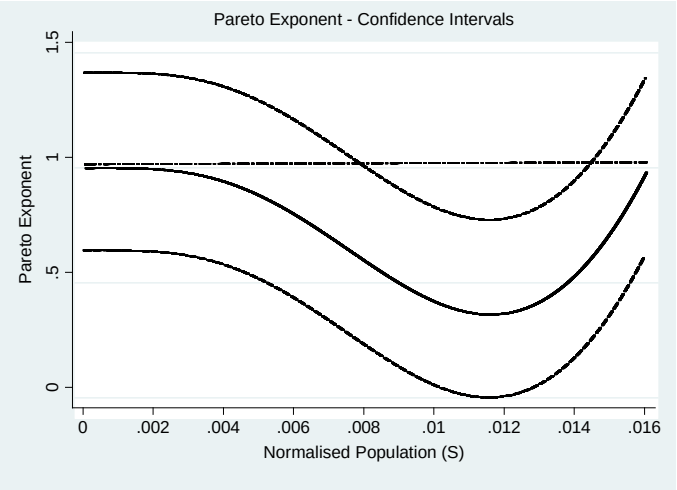
Third, the estimated values of the Pareto exponent increase with the truncation point ($\underline{N}$) for all the geographical units using any of the estimators, supporting Eeckhout's (2004) claim that the estimated value of Pareto's exponent depends negatively on the cut-off point. Therefore, when all units are considered (samples with no truncation point), the estimated coefficients are the lowest ones. Fourth Zipf's law only holds for Metropolitan Statistical Areas and top urbanized areas; in the rest of the cases, the estimates are significantly different from the value one. This indicates that city definition really matters in the fulfilment of Zipf's law (Rosen and Resnick, 1980).

Figure 3 displays an example of the evolution of the Pareto exponent over time, using the untruncated sample of incorporated places, summarised in Table 3. Again the three estimators are applied (simple OLS, regressions Gabaix and Ibragimov's estimator, and the Hill estimator). The same explanations about the differences between estimators apply (e.g. when all incorporated places are considered with no truncation point, results from simple OLS regressions and Gabaix and Ibragimov's estimator are the same, while Hill's estimates are much lower). The evolution of the Pareto exponent is similar with the three estimators: decreasing over time for all incorporated places and increasing for top samples. This would indicate that for the entire sample (including all the incorporated places for each year), a divergent behaviour was produced, while for the biggest incorporated places, the trend has been convergence: they have become closer in size. González-Val (2010) obtains the same conclusion, using Gini coefficients.

Figure 4 shows the estimates of the Pareto exponent, applying the innovative Ioannides and Overman's (2003) nonparametric methodology, considering a pool over the whole century, with all incorporated places in Table 3 (162, 698 observations); some observations with extreme values are excluded, see González-Val (2012). To calculate the nonparametric estimates of the condi-

tional mean variance of growth on city size and the derivates (see the expression in Table 1), we apply the LOcally WEighted Scatter plot Smoothing algorithm. Results are shown until city sizes with a normalised population of 0.016, because of one technical problem with this procedure: the sparsity of data at the upper tail of the distribution, which produces extreme values of the estimations (Ioannides and Overman, 2003). The dotted lines are bootstrapped: 95% confidence bands calculated, using 500 random samples with replacement.

**Figure 4. Nonparametric estimate for all the 20th century (Incorporated Places a pool of 162, 698 observations LOcally WEighted Scatter plot Smoothing (LOWESS) algorithm)**




*Note: This figure is obtained by applying the nonparametric procedure proposed by Ioannides and Overman (2003); this empirical strategy relies on the statistical foundation of Zipf's law offered by Gabaix (1999). The normalised population of city i is the population of city i divided by the contemporary total urban population. More information about the nonparametric estimations of the local Zipf exponent using data for all cities can be found in González-Val (2012). This graph is equivalent to Figure 3a in González-Val (2012).*

The exponent decreases with city size until reaching the normalised size of 0.012, when it begins to grow to reach a value close to one. For most of the distribution of city sizes the value one falls within the confidence bands, indicating that Zipf's law holds for most of the city sizes (especially because most of the observations are concentrated in the lower tail of the distribution on the left side of the graph).

## 3. CONCLUSIONS

Zipf's law is an appealing empirical regularity. One of its main attractions is that it is easy to check – the estimation of the Pareto distribution can be carried out simply by fitting a line to data on city size (population). Moreover, Zipf plots provide a graphical tool to observe the quality of the fit to the real behaviour of the distribution. However, even this simple empirical exercise

implies several choices that can affect the results. In this paper, we review the three main issues in any empirical work on city size distribution and Zipf's law: city definition, sample size and the choice of the estimator.

The choice of city definition depends on data availability and the research objective. The geographical unit chosen is also closely related to sample size and the sample size has a clear effect on the estimate of the Pareto exponent (Eeckhout, 2004). If any truncation point is imposed, the estimates of the Pareto exponent may be biased, but if all cities are considered, often nonlinear behaviours appear.

Finally, we carry out an empirical exercise with US city size distribution data to examine the features of several estimators of the Zipf exponent. To sum up, the Hill estimator is the best if the distribution actually follows a power law, but if you have doubts about the power law behaviour of your sample, the Gabaix and Ibragimov specification performs better than the others, especially for small samples.

## REFERENCES

Beeson P. E. D. N. DeJong and W. Troesken (2001). Population Growth in US Counties 1840-1990. *Regional Science and Urban Economics* 31: 669–699.

Berry B. J. L. and A. Okulicz-Kozaryn (2011). The city size distribution debate: Resolution for US urban regions and megalopolitan areas. *Cities forthcoming.*

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

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

Dobkins L. H. and Y. M. Ioannides (2000). Dynamic evolution of the US city size distribution. Included in Huriot J. M. and J. F. Thisse (Eds.) *The economics of cities. Cambridge: Cambridge University Press* 217–260.

Dobkins L. H. and Y. M. Ioannides (2001). Spatial interactions among US cities: 1900–1990. *Regional Science and Urban Economics* 31: 701–731.

Duranton G. (2007). Urban Evolutions: The Fast the Slow and the Still. *American Economic Review* 97(1): 197–221.

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

Eeckhout J. (2009). Gibrat's Law for (all) Cities: Reply. *American Economic Review* 99(4): 1676–1683.

Ehrlich S. and J. Gyourko (2000). Changes in the scale and size distribution of US metropolitan areas during the twentieth century. *Urban Studies* 37(7): 1063–1077.

Fan C. C. and E. Casetti (1994). The Spatial and Temporal Dynamics of US Regional Income Inequality 1950–1989. *Annals of Regional Science* 28: 177–196.

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

Gabaix X. and R. Ibragimov (2011). Rank-1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics* 29(1): 24–39.

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

Garmestani A. S., C. R. Allen and K. M. Bessey (2005). Time-series Analysis of Clusters in City Size Distributions. *Urban Studies* 42(9): 1507–1515.

Garmestani A. S., C. R. Allen and C. M. Gallagher (2008). Power laws discontinuities and regional city size distributions. *Journal of Economic Behavior & Organization* 68: 209–216.

Giesen K. and J. Südekum (2011). Zipf's law for cities in the regions and the country. *Journal of Economic Geography* 11(4): 667–686.

Giesen K., A. Zimmermann and J. Suedekum (2010). The size distribution across all cities – double Pareto lognormal strikes. *Journal of Urban Economics* 68: 129–137.

Glaeser E. L. and J. Shapiro (2003). Urban Growth in the 1990s: Is city living back? *Journal of Regional Science* 43(1): 139–165.

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.

González-Val R. (2010). The Evolution of the US City Size Distribution from a Long-run Perspective (1900–2000). *Journal of Regional Science* 50(5): 952–972.

González-Val R. (2011). Deviations from Zipf's law for American cities: An empirical examination. *Urban Studies* 48(5): 1017–1035.

González-Val R. (2012). A nonparametric estimation of the local Zipf exponent for all US cities. Forthcoming in *Environment and Planning B: Planning and Design.*

González-Val R. and M. Sanso-Navarro (2010). Gibrat's Law for countries. *Journal of Population Economics* 23(4): 1371–1389.

Guérin-Pace F. (1995). Rank-Size Distribution and the Process of Urban Growth. *Urban Studies* 32(3): 551–562.

Ioannides Y. M. and H. G. Overman (2003). Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics* 33: 127–137.

Ioannides Y. M. and S. Skouras (2009). Gibrat's Law for (All) Cities*:* A Rejoinder. *Discussion Papers Series Department of Economics* Tufts University 0740 Department of Economics Tufts University.

Krugman P. (1996). The Self-organizing economy. *Cambridge: Blackwell.*

Levy M. (2009). Gibrat's Law for (all) Cities: A Comment. *American Economic Review* 99(4): 1672–1675.

Malacarne L. C., R. S. Mendes and E. K. Lenzi (2001). q-exponential distribution in urban agglomeration. *Physical Review E* 65 017106.

Malevergne Y. V. Pisarenko and D. Sornette (2011). Gibrat's Law for cities: Uniformly most powerful unbiased test of the Pareto against the lognormal. In press in *Physical Review E.* arXiv:0909.1281v1 [physics.data-an].

Michaels G., F. Rauch and S. J. Redding (2012). Urbanization and Structural Transformation. Forthcoming in *Quarterly Journal of Economics.*

Naldi M. (2003). Concentration indices and Zipf's law. *Economics Letters* 78: 329–334.

Newman M. E. J. (2006). Power laws Pareto distributions and Zipf's law. *Contemporary Physics* 46: 323–351.

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.

Nitsch V. (2005). Zipf zipped. *Journal of Urban Economics* 57: 86–100.

Overman H. G. and Y. M. Ioannides (2001). Cross-Sectional evolution of the US City Size Distribution. *Journal of Urban Economics* 49: 543–566.

Parr J. B. and K. Suzuki (1973). Settlement populations and the lognormal distribution. *Urban Studies* 10: 335–352.

Petrakos G., P. Mardakis and H. Caraveli (2000). Recent Developments in the Greek System of Urban Centres. *Environment and Planning B: Planning and Design* 27(2): 169–181.

Reed W. J. (2002). On the rank-size distribution for human settlements. *Journal of Regional Science* 42(1): 1–17.

Rose A. K. (2006). Cities and countries. *Journal of Money Credit and Banking* 38(8): 2225–2245.

Rosen K. T. and M. Resnick (1980). The Size Distribution of Cities: An Examination of the Pareto Law and Primacy. *Journal of Urban Economics* 8: 165–186.

Rossi-Hansberg E. and M. L. J. Wright (2007). Urban structure and growth. *Review of Economic Studies* 74: 597–624.

Rozenfeld H. D., D. Rybski, X. Gabaix and H. A. Makse (2011). The Area and Population of Cities: New Insights from a Different Perspective on Cities. *American Economic Review* 101(5): 2205–2225.

Song S. and K. H. Zhang (2002). Urbanisation and City Size Distribution in China. *Urban Studies* 39(12): 2317–2327.

Soo K. T. (2005). Zipf's Law for cities: A cross-country investigation. *Regional Science and Urban Economics* 35: 239–263.

Soo K. T. (2007). Zipf's Law and Urban Growth in Malaysia. *Urban Studies* 44(1): 1–14.

Soo K. T. (2011). The size and growth of state populations in the United States. Working paper Lancaster University.

Urzúa C. M. (2000). A simple and efficient test for Zipf's law. *Economics Letters* 66: 257–260.

Wheaton W. C. and H. Shishado (1981). Urban concentration agglomeration economies and the level of economic development. *Economic Development and Cultural Change* 30: 17–30.

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

## LA LOI DE ZIPF : LES QUESTIONS FONDAMENTALES DES TRAVAUX EMPIRIQUES

***Résumé -*** *La loi de Zipf figure parmi les régularités les plus singulières dans le domaine de l'économie urbaine. Selon cette loi, la distribution rang-taille des villes suit, dans un grand nombre de pays, une distribution de Pareto avec un coefficient de hiérarchisation égal à 1. Cet article examine trois questions importantes dans les travaux empiriques relatifs à la loi de Zipf : la définition de la ville et de l'unité géographique retenue, la taille de l'échantillon et le choix de la méthode d'estimation associée. A titre d'illustration, une application est proposée à partir de données récentes sur des villes américaines.*

***Mots-clés :*** LOI DE ZIPF ; DISTRIBUTION RANG-TAILLE DES VILLES