US city size distribution and space

González-Val, Rafael

Universidad de Zaragoza, Institut d'Economia de Barcelona (IEB)

14 December 2018

US city size distribution and space

Rafael González-Val

Universidad de Zaragoza & Institut d'Economia de Barcelona (IEB)

***Abstract*:** This paper focuses on the spatial city size distribution in the United States. We propose a new distance-based approach to analyse the influence of distance on the city size distribution parameter, by considering the Pareto distribution and using data from different definitions of US cities in 2010. Considering all possible combinations of cities within a 300-mile radius, our results indicate that the Pareto distribution cannot be rejected in most cases regardless of city size. Placebo regressions validate our results, thereby confirming the significant effect of geography on the Pareto exponent.

***Keywords*:** space, city size distribution, distance-based approach, Pareto distribution, Zipf's law.

***JEL*:** C12, C14, R11, R12.

Rafael González-Val, Departamento de Análisis Económico, Facultad de Economía y Empresa, Universidad de Zaragoza, Gran Vía 2, Zaragoza 50005, Spain; Institut d'Economia de Barcelona (IEB), Facultat d'Economia i Empresa, Universitat de Barcelona, c/ John M. Keynes, 1-11, Barcelona 08034, Spain. Email: rafaelg@unizar.es

1. Introduction

In 1913, Auerbach found a striking empirical regularity that establishes a linear and stable relationship between city size and rank, which has fascinated researchers from many fields (e.g., economics, statistics, physics, and geography) since then. In statistical terms, this relationship means that city size distribution can be well fit with a Pareto distribution, which is also known as a power law. Some decades later, this empirical regularity became known as Zipf's law (Zipf, 1949), although Zipf's law is simply a particular case of that linear relationship where the parameter of the Pareto distribution is equal to one, which means that, when ordered from largest to smallest, the size of the second-largest city in a country is half that of the first, the size of the third is a third of the first, etc. Over the years, numerous studies have tested the validity of this law for many different countries (see the surveys by Cheshire, 1999; Nitsch, 2005; and, more recently, Cottineau, 2017).

Although interest in city size distributions and Zipf's law has fluctuated over time, in the last few decades there has been a revival of interest among urban economists, especially since Krugman (1996a) highlight the "mystery of urban hierarchy." In a fundamental contribution, Krugman (1996b) use data from metropolitan areas from the Statistical Abstract of the United States (135 cities) and conclude that in 1991 the Pareto's exponent was exactly equal to 1.005. This finding provides evidence supporting Zipf's law at that time in the United States (US). Zipf's law provides a simple but accurate representation of city size distribution and, therefore, some theoretical models with different economic foundations have been proposed to explain the law: 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 that postulates that the growth rates of cities tend to be independent of their initial sizes). In the theoretical literature, Zipf's law was viewed as a reflection of a steady-state situation.

However, things changed after publication of Eeckhout (2004). Traditionally due to data limitations, most of the studies have considered only the largest cities. However, Eeckhout (2004) demonstrated the statistical importance of considering both large and small cities. Truncated samples lead to biased results, and city definition (administrative cities versus metro areas) also plays a key role in the final results (Rosen and Resnick, 1980). But in a larger blow to Zipf's law, Eeckhout (2004) concludes that city size distribution is actually lognormal rather than Pareto. Since then, most studies have considered un-truncated data (Giesen et al., 2010; González-Val et al., 2015; Ioannides and Skouras, 2013), but the lognormal distribution soon was replaced by other more convoluted distributions that provide a better fit to the actual data: the q-exponential distribution (Malacarne et al., 2001; Soo, 2007), the double Pareto lognormal distribution (Giesen et al., 2010; Giesen and Suedekum, 2014; Reed, 2002), or the distribution function by Ioannides and Skouras (2013) that switches between a lognormal and a power distribution.

Most of these new distributions combine linear and nonlinear functions, separating the body of the distribution from the upper-tail behaviour. The reason is that the largest cities encompass most of the population of a country, and the behaviour of the upper-tail distribution can be different from that of the entire distribution. In fact, the largest cities follow a Pareto distribution in many cases (Levy, 2009). As Ioannides and Skouras (2013) pointed out, "*most cities* obey a lognormal; but the upper-tail and therefore *most of the population* obeys a Pareto law."

Therefore, if the Pareto distribution is still valid but only for the upper-tail distribution (i.e. the largest cities), it is possible to reconcile the traditional literature focused on small data sets of large cities with the last empirical studies considering all cities with no size restrictions and the theoretical models considering Zipf's law as the benchmark for the distribution of city sizes. However, this solution is unsatisfactory for two main reasons.

First, urban theoretical models should attempt to explain city sizes and urban systems without imposing any size restriction. It is true that the Pareto distribution provides a simple theoretical specification to include in an analytical framework, but if models are restricted to studying only the largest cities at the upper-tail of the distribution, where Zipf's law holds, we are excluding the majority of cities, which actually are small and medium size, from the analysis. It is not easy to justify from a theoretical or empirical point of view the exclusion of most cities, particularly when there is empirical evidence indicating that the lower tail of the distribution, the smallest cities, are also Pareto-distributed (Giesen et al., 2010; Giesen and Suedekum, 2014; Luckstead and Devadoss, 2017; Reed, 2001, 2002).

Second, a Pareto distribution can be fit to a wide range of phenomena: the distribution of the number of times that different words appear in a book (Zipf, 1949), the losses caused by floods (Pisarenko, 1998), and the intensity of wars or forest fires (Roberts and Turcotte, 1998), for example. However, the city size distribution case is different because there is a spatial dependence among the elements of the distribution; cities are connected through migratory flows. An essential assumption in urban models to obtaining spatial equilibrium is free migration across cities. Therefore, there is a relationship between the population of one city and the populations of nearby cities. Nevertheless, the upper-tail of the distribution contains large cities that typically are

very far away from one another. For instance, the bilateral physical distance between New York, the largest city in the US, and Los Angeles, the second largest city, is more than 2,400 miles. If we consider the 10 largest cities in the US in 2010, the average physical distance between these cities is greater than 1,000 miles. So, on average, there is a significant distance between the largest cities. Is it therefore possible that migration could be significant between these cities?[1]

Rauch (2014) answers this question using the 2000 US census to collect data pertaining to the distances that people move. He creates bins of size 100 kilometres (approximately 62 miles) and concludes that the large majority of people (over 68% of observations), fall into the bin with a distance between 0 and 100 km. This finding suggesting that the majority of US citizens live near their birthplace. Rauch (2014) also finds that the relationship between the number of people and the distance between home and place of birth decreases with distance by estimating a standard gravity equation.

Therefore, there can be migrations between the largest cities even if they are far from one another. However, these migrations are not significant because most people do not move so far. Therefore, it is not clear whether we can use a spatial equilibrium model to explain the distant largest cities as a whole, and what means that the Pareto distribution (and Zipf's law), which represents the steady city size distribution in many theoretical models (Duranton, 2007; Gabaix, 1999; Rossi-Hansberg and Wright, 2007), holds for the largest cities because they are almost independent elements. This situation implies that the largest cities are the centres of different urban systems. There are different theories that can explain a hierarchical system of cities with a multiplicity of equilibria, from the classical theory of the central place by Christaller and Lösch and von Thünen's model to more recent models that update these theories, including modern

[1] We focus on migrations because it is obvious that there cannot be significant commuting across such wide distances; commuting typically occurs within metropolitan areas from surrounding cities to a central place.

agglomeration economies (for instance, Fujita et al., 1999; Hsu, 2012). However, the empirical literature on city size distribution typically omits this spatial issue. As a result, interpretation of results has been reduced to identify the Pareto upper-tail, irrespective of whether there is any meaningful relationship between the largest cities. A few exceptions are Dobkins and Ioannides (2001) and Ioannides and Overman (2004).

Other authors also argue for the need to focus on the regional level rather than the overall city size distribution for the whole country (although both can be related). Gabaix (1999) shows that if urban growth in all regions follows Gibrat's law we should observe the Zipfian upper-tail distribution on both the regional and national levels. Giesen and Südekum (2011) test this hypothesis for the German case, finding that Zipf's law is not only satisfied for Germany's national urban hierarchy but also in single German regions. Lalanne (2014) studies the hierarchical structure of the Canadian urban system; splitting the Canadian territory into two parts (east and west) allows her to identify different dynamics that were not observable when studying the country as a whole. Finally, Hsu et al. (2014) analyse the size distribution of US Core Based Statistical Areas using subsets of cities. These authors find that spatial partitions of cities based on geographical proximity are significantly more consistent with the Pareto distribution than are random partitions.

In this study, we develop a new methodology to analyse how city size distribution changes over space. We consider all of the possible combinations of cities within a 300-mile radius. Section 2 presents the database that we use. In Section 3, we introduce a new distance-based approach to study the influence of distance on the city size distribution parameter, and finally we check the significance of that relationship with some robustness checks in Section 4, including placebo regressions. In section 5 we discuss the results and section 6 concludes the paper.

2. Data

There are many definitions of cities, but the two common alternatives in the literature are the metropolitan areas and the administratively defined cities (legal cities). Here we consider three different city definitions: places, urban areas, and core-based statistical areas. Table 1 lists the descriptive statistics. Our data derive from the 2010 US decennial census. Geographical coordinates (latitude and longitude) necessary to compute the bilateral distances between cities were obtained from the 2010 Census US Gazetteer files.[2] This same data set is used by González-Val (2018a) to study the spatial distribution of US cities; so our exposition here follows closely both the geographic terms and concepts of the US Census Bureau and González-Val's (2018a) data description.

The generic denomination 'places' has included all incorporated and unincorporated places since the 2000 census. According to the US Census Bureau guidelines,[3] the generic term 'incorporated place' designates a type of governmental unit incorporated under state law, "established to provide governmental functions for a concentration of people." An incorporated place usually is a city, town, village, or borough, but can have other legal descriptions. On the other hand, there are 'unincorporated places' (which were renamed Census Designated Places in the 1980 census), which designate "settled concentrations of population that are identifiable by name but are not legally incorporated under the laws of the state in which they are located." Thus, the difference between incorporated and unincorporated places is merely political and/or administrative in most cases. Last years these places have been used in empirical studies of American city size distribution (Eeckhout, 2004; Giesen et al.,

[2] Although there are several definitions of cities in the US, the Census US Gazetteer files only provide coordinates for places, urban areas, and core-based statistical areas. Therefore, the use of any other definition of city would imply the use of non-official geographical coordinates.
[3] See https://www.census.gov/geo/reference/gtc/gtc_place.html.

2010; González-Val, 2010; Levy, 2009), and their primary advantage is that this city definition does not impose any truncation point (populations range from 1 to 8,175,133 inhabitants).

'Urban area' is the generic term for urbanized areas and urban clusters. As the US Census Bureau indicates,[4] "urbanized areas consist of a densely developed area that contains 50,000 or more people", while "urban clusters consist of a densely developed area that has a least 2,500 people but fewer than 50,000 people." Therefore, a minimum size restriction of 2,500 inhabitants is imposed (see Table 1). The US Census Bureau classifies all territory and population located within an urbanized area or urban cluster as urban and all areas outside as rural. Previous empirical studies based on this definition of urban areas include Garmestani et al. (2005) and Garmestani et al. (2008). Moreover, urban areas are used as the cores for which core-based statistical areas (CBSAs) are defined.

Finally, 'Core-based statistical areas' are defined by the US Census Bureau[5] as "the county or counties or equivalent entities associated with at least one core (urbanized area or urban cluster) of at least 10,000 population, plus adjacent counties having a high degree of social and economic integration with the core as measured through commuting ties with the counties associated with the core." The term core-based statistical areas includes both metropolitan and micropolitan statistical areas. The difference between them is the classification of the core as urbanized area or urban cluster. Following the US Census Bureau definitions, "metropolitan statistical areas are CBSAs associated with at least one urbanized area that has a population of at least

[4] Visit https://www.census.gov/geo/reference/webatlas/uas.html to see more information and examples of urban areas.

[5] US Census Bureau definitions for CBSAs, metropolitan and micropolitan statistical can be found at https://www.census.gov/geo/reference/gtc/gtc_cbsa.html.

50,000" and "micropolitan statistical areas are CBSAs associated with at least one urban cluster that has a population of at least 10,000 but less than 50,000" people.

Note that our city definitions are nested; most places are included in urban areas, and most urban areas and places are located inside CBSAs. For research purposes, any of these spatial units have pros and cons. The three samples include most of the population of the country (73.3% of the total US population lives in places, 81.9% is located in urban areas and 93.9% is included in CBSAs). Places are administratively defined cities (legal cities), and their boundaries make no economic sense. However, some factors, such as human capital spillovers, are believed to operate at a very local level (Eeckhout, 2004). Urban areas represent urban agglomerations from which rural locations are excluded. Moreover, CBSAs are more natural economic units; they cover huge areas that are meant to capture labour markets. Core-based statistical areas have economic meaning because they include the core area with a population nucleus together with adjacent communities with a high degree of economic and social integration with that core. Nevertheless, Eeckhout (2004) demonstrated the statistical importance of considering the whole sample. This author recommends the use of places (un-truncated data) rather than metro areas (urban areas or CBSAs) because if any truncation point is imposed the estimates of the Pareto exponent may be biased.

3. The spatial city size distribution

We study how city size distribution changes over distance. However, this exercise is not a spatial econometrics one. City size distribution can be estimated using spatial econometrics techniques to account for spatial dependence. Le Gallo and Chasco (2008) consider Spanish urban areas from 1900–2001 to estimate Zipf's law using a spatial SUR model. Our approach is different; space is introduced in our methodology through the selection of geographical samples of cities based on distances.

Therefore, our first step is to define the geographical samples of neighbouring cities. Different criteria can be used to select the samples. For instance, Hsu et al. (2014) consider a fixed number of samples (regions) using geographical (travel distance between cities) and economic (trade linkages) criteria. Berry and Okulicz-Kozaryn (2012) use a labour market criterion, based on commuting time to jobs located in urban cores. Therefore, depending on the criterion, one obtains a concrete set of subsystems with particular groups of neighbouring cities. Because there are many alternative criteria (based on economic, social, or geographical factors) that could give rise to different groups of cities, in this paper we follow an agnostic view: we consider all the possible combinations of cities within a 300-mile radius based on physical geographic distances. The choice of this threshold is based on a conservative criterion; although Rauch (2014) concludes that most of people in the US (over 68% of observations) live near their place of birth (within 100 km) and therefore the extent of spatial interactions between cities is reduced to this short distance, we consider a higher threshold of 300 miles, which is roughly one third of the median distance between all pairs of cities (848 miles for places and 857 for urban areas, to be precise).[6]

Bilateral distances between all cities are calculated using the haversine distance measure.[7] Then, circles of radius $r = 15,20,...,300$ miles are drawn around the geographic centroid of each city's coordinates, starting from a minimum distance of 15 miles, adding 5 miles each time;[8] in the case of CBSAs we start the procedure at 50 miles because they are large spatial units encompassing huge areas and therefore for short distances there are very few units. We obtain 58 different geographical samples for each city for places and urban areas and 51 different geographical samples in the

[6] The threshold only indicates the distance at which to stop the procedure, but estimated results for each particular distance are not sensitive to the threshold selection.
[7] The haversine formula determines the great-circle distance between two points on the surface of the Earth given their longitudes and latitudes, taking into account the mean radius of the Earth.
[8] We repeated the analysis adding 1 mile each time for a few cities, and the results were very similar.

case of CBSAs. We repeat this exercise for all cities, considering both places and urban areas. We recover 1,666,804 (28,738x58), 208,336 (3,592x58), and 47,379 (929x51) geographical samples for places, urban areas, and CBSAs, respectively. Note that, within these geographical samples, we consider all cities with no size restriction. Obviously, the number of cities included within the circles also increases as distance increases; in Section 4.2 we explicitly analyse the relationship between geographical distance and sample size. Finally, in some cases samples are repeated (different circles include exactly the same cities) or are single-city samples. We will deal with these issues later.

Once we defined the geographical samples, we examined the behaviour of city size distribution from this spatial perspective. As noted in the Introduction section, the Pareto distribution is the benchmark in both the theoretical and empirical literature on city size distribution. Let S denote the city size (measured by the population); if S is distributed according to a power law, also known as a Pareto distribution[9], the density function is $p(S) = \frac{a-1}{\underline{S}}\left(\frac{S}{\underline{S}}\right)^{-a} \quad \forall S \geq \underline{S}$ and the complementary cumulative density function $P(S)$ is $P(S) = \left(\frac{S}{\underline{S}}\right)^{-a+1} \quad \forall S \geq \underline{S}$, where $a > 0$ is the Pareto exponent (or the scaling parameter) and $\underline{S}$ is the population of the city at the truncation point. The Pareto distribution is the typical distribution without a characteristic scale; urban systems are complex systems for which we cannot determine the characteristic scales in many cases (Chen and Zhou, 2008). Therefore, some authors favour quantitative analysis based on scaling instead of quantitative analysis based on characteristic scales, and the solution to an equation of scaling relation is always a power law.

[9] According to Newman (2006), 'Zipf's law' and 'Pareto distribution' are effectively synonymous with 'power law distribution.'

It is easy to obtain the expression $R = A \cdot S^{-a}$, which relates the empirically observed rank R (1 for the largest city, 2 for the second largest, and so on) to the city size. This expression has been used extensively in urban economics to study city size distribution (Cheshire, 1999; Gabaix and Ioannides, 2004).

First, we tested whether this distribution provides an acceptable fit to our geographical samples of cities. For each geographical sample, we used the statistical test for goodness-of-fit proposed by Clauset et al. (2009),[10] recently used by González-Val (2018b) to analyse the evolution of the European urban system from 1300 to 1800. The test is based on a measurement of the 'distance' between the empirical distribution of the data and the hypothesised Pareto distribution. This distance is compared with the distance measurements for comparable synthetic data sets drawn from the hypothesised Pareto distribution, and we defined the p-value as the fraction of the synthetic distances that are larger than the empirical distance. This semi-parametric bootstrap approach is based on the iterative calculation of the Kolmogorov-Smirnov (KS) statistic for 100 bootstrap data set replications.[11] The Pareto exponent is estimated for each geographical sample of cities using the maximum likelihood (ML) estimator, and then the KS statistic is computed for the data and the fitted model.[12] Single-city samples are excluded. The test samples from the observed data and checks how often the resulting synthetic

[10] As a robustness check, we also used the statistical test proposed by Gabaix (2009) and Gabaix and Ibragimov (2011) to study the validity of the Pareto distribution; this test is based on a modification of the Rank-$1/2$ OLS regression. This test has been specifically developed to work with small samples because it reduces the small-sample bias, but our results revealed that the number of rejections of the null of an exact power law significantly increased with the number of cities in the sample. Therefore, the results of Gabaix's (2009) test for urban areas and CBSAs are quite similar to those obtained with Clauset et al.'s (2009) test. However, the results for places using large sample sizes are different because Gabaix's test detects a larger number of rejections of the Pareto distribution than Clauset et al.'s test. These results are available upon request.

[11] The procedure is highly computationally intensive. We computed the test with 300 replications for a few cities, and the results were similar.

[12] Actually, the procedure by Clauset et al. (2009) is specifically designed to select an optimal truncation point. To select the lower bound, the Pareto exponent is estimated for each sample size using the ML estimator, computing the KS statistic for each sample size. The truncation point that is finally selected corresponds to the value of the threshold for which the KS statistic is the smallest. However, in this paper we do not truncate our data. Therefore, the value of the threshold is set to the minimum population in the sample in all cases, considering all the available observations in each geographical sample.

distribution fit the actual data as poorly as the ML-estimated power law. Therefore, the null hypothesis is the power law behaviour of the original sample. Nevertheless, this test has an unusual interpretation because, regardless of the true distribution from which our data were drawn, we can always fit a power law. Clauset et al. (2009) recommend the conservative choice that the power law is ruled out if the p-value is below 0.1, that is, if there is a probability of 1 in 10 or less that we would obtain data merely by chance that agree as poorly with the model as the data that we have. Therefore, this procedure only allows us to conclude whether the power law achieves a plausible fit to the data.

Figure 1 shows the result of the Pareto test by distance. For each distance, the graphs represent the fraction of p-values less than 0.1[13] divided by the total number of tests carried out at that distance.[14] Regarding places (Figure 1(a)), the percentage of rejections of the Pareto distribution clearly increases with distance, but it is always below 40%, even for the longest distance considered. One explanation for the increasing number of rejections is as follows: the power law can be replaced by another type of distribution function such as a lognormal distribution when we consider un-truncated data and when distance and sample size increase.[15]

The results for urban areas and CBSAs are similar (Figures 1(b) and 1(c)). For small sample sizes at short distances, the percentage of power law rejections is high but lower than 50%. As distance increases, the rejection rate decreases to a rather constant value lower than 10%. This situation suggests that the Pareto distribution is a plausible

[13] We use the 0.1 reference value for the p-value, as Clauset et al. (2009) recommend. Other significance levels (1% and 5%) yield similar results.

[14] By construction, as we start to build up the geographical samples from each city the number of tests by distance should coincide with the number of cities in the sample. However, in some specific cases with very low sample sizes the log-likelihood cannot be computed and therefore the test cannot be carried out. Single-city samples are also excluded. Thus, the number of tests by distance is not constant, although the differences are small. The number of tests carried out by distance ranges from 27,886–28,755 in the case of places, from 2,088–3,591 for urban areas and from 796–929 for CBSAs.

[15] Power laws imply scaling in cities. A power law can often be identified among a certain range of scales, but a power law must eventually breaks if the scales of measurements are too large or too small (Bak, 1996; Chen, 2011; Chen and Zhou, 2008; Williams, 1997). Therefore, it is easier to fit a Pareto distribution to city size data with a scale-free range compared with using un-truncated data.

approximation for the real behaviour of the data in our geographical samples in all cases, for any distance, and for the three definitions of city we adopt. Recall that we do not impose any size restriction; therefore, nearby cities are Pareto-distributed regardless of the size of the cities included in the samples. Most of the possible combinations of neighbouring cities, for which economic interactions and migratory flows are significant, are Pareto-distributed.

Once we conclude that the Pareto distribution is an acceptable description of city sizes, we proceed to estimate the Pareto exponent. Although previously we have estimated the parameter by ML to run the goodness-of-fit test, now we apply Gabaix and Ibragimov's Rank-$1/2$ estimator. The reason for this choice is that this estimator performs better with small samples. However, when the sample size is large differences between estimators are reduced (González-Val, 2012). Moreover, Gabaix and Ibragimov (2011) suggest that their estimator produces more robust results than the ML estimator when data deviate from a power law distribution.

Taking natural logarithms from the expression $R = A \cdot S^{-a}$, we obtain the linear specification that is typically estimated:

$$\ln R = b - a \ln S + \xi , \qquad (1)$$

where ξ is the error term and b and a are parameters that characterise the distribution. Gabaix and Ibragimov (2011) propose specifying Equation (1) by subtracting $1/2$ from the rank to obtain an unbiased estimation of a:

$$\ln\left(R - \frac{1}{2} \right) = b - a \ln S + \varepsilon . \qquad (2)$$

The larger the coefficient $\hat{a}$, the more homogeneous are the city sizes. Similarly, a small coefficient (less than 1) indicates a heavy-tailed distribution. Zipf's law is an

empirical regularity that appears when the Pareto exponent of the distribution is equal to unity ($a = 1$).

Equation (2) is estimated iteratively by OLS for all of our geographical samples by distance starting from every city. In other words, we obtain 58 different estimates (51 for CBSAs) of the Pareto exponent for each city. After running all of the regressions, we obtain 1,665,962 Pareto exponent-distance pairs for places, 204,959 in the case of urban areas, and 45,912 for CBSAs. Single-city samples are excluded.[16] Next, to summarise all these point-estimates, we conduct a nonparametric estimation of the relationship between distance and the estimated Pareto exponents using local polynomial smoothing. The local polynomial smoother fits the Pareto exponent to a polynomial form of distance via locally weighted least squares, and a Gaussian kernel function is used to calculate the locally weighted polynomial regression.[17] Figure 2 shows the results, including the 95% confidence intervals. Our results are similar for the three city definitions: as distance increases, the Pareto exponent decreases. The decreasing Pareto exponent converges to the value estimated for the entire sample of cities, which is represented by the horizontal line in Figure 2. A possible explanation for this convergence is that, as distance increases, so does the number of cities within the samples. This situation decreases the coefficient (Eeckhout, 2004). In Section 4.2, we discuss the placebo regressions that we run to test whether sample size is the only factor driving our results. Finally, the estimated coefficients of urban areas and CBSAs tend to be higher than those of places because of the different definition of cities (González-Val, 2012). Empirical research has established that the city size data are typically well

[16] The number of regressions does not coincide exactly with the number of cities multiplied by the 58 different distances considered (51 for CBSAs) because in some cases there is only one city in the sample at the start of the procedure with small distances. Therefore, the regression is skipped until there is more than one city in the geographical sample.

[17] We use the lpolyci command in STATA with the following options: local mean smoothing, a Gaussian kernel function, and a bandwidth determined using Silverman's rule-of-thumb.

described by a power law with an exponent between 0.8 and 1.2 (Gabaix, 2009). In the case of urban areas and CBSAs, the average value of the estimated exponent is between 0.8 and 1.2 for all distances beyond 30 and 75 miles, respectively. Moreover, for short distances (50–75 miles for urban areas and 75–80 miles in the case of CBSAs), an exponent of 1 falls within the confidence bands. Therefore, we cannot reject Zipf's law for those geographical samples at those distances.

4. Robustness checks

In this section, we carry out some robustness checks. Previous results have indicated that the Pareto distribution is an acceptable approximation for the real behaviour of the data in our geographical samples, for any distance and for the three city definitions. Moreover, nearby cities are Pareto-distributed regardless of the size of the cities included in the samples because we do not impose any size restrictions.

4.1 Repeated estimations

In some cases, some of our geographical samples may be repeated. Recall that we draw circles of different radii from 0 to 300 miles starting from each city to consider all of the possible combinations of cities. Therefore, if the core cities of two different circles are close the geographical samples may be similar or even identical. Many repeated observations could be driving these results. To check whether this situation was a problem, we repeated the analysis considering only geographical samples with a core city of more than 100,000 inhabitants. The largest places tend to be dispersed geographically; if we consider only the geographical samples with a large core city, we should avoid replicated samples.

Figure 3 shows the results of the Pareto goodness-of-fit test. Figures 3(a), 3(b), and 3(c) display a similar evolution of the percentage of rejections with distance to that shown in Figures 1(a), 1(b), and 1(c) when all of the geographical samples are

considered.[18] The only difference is that now the percentage of rejections in the case of places is slightly higher, especially for the largest distances, when it reaches 51%. Nevertheless, we can still argue that the Pareto distribution is a plausible fit to the city size distribution for places for most distances. For urban areas and CBSAs, the percentage of rejections remains below 10% for most distances.

The sample selection of this robustness check reduces the number of point estimates of the Pareto exponent; now we obtain 16,237 Pareto exponent-distance pairs for places, 17,034 for urban areas, and 18,530 in the case of CBSAs. Figure 4 shows the nonparametric relationship between the Pareto exponent and distance, which is still decreasing in the three cases. In the case of urban areas, the Pareto exponent strongly decreases from 0 to 50 miles and then starts to slowly increase as it approaches the estimated value for the entire sample (the horizontal line in Figure 4) from below.

4.2 Placebo regressions

We consider geographical samples that represent all the possible combinations of cities within a 300-mile radius. Each geographical sample includes a particular number of cities; we have 1,666,804, 208,336, and 47,379 sample sizes for places, urban areas, and CBSAs, respectively. The surface area πr^2 of a circle is a quadratic function of its radius r. Therefore, the number of cities asymptotically will be a quadratic function of r.As the radius (i.e., distance) increases, the number of cities included in the circles naturally also increases.

It may be that our results are only driven by sample size, especially because the decreasing relationship between the Pareto exponent and sample size is already known (Eeckhout, 2004). To investigate this issue, we ran placebo regressions. We had previously constructed 58 different geographical samples starting from each city (51 for

[18] Now the number of tests carried out by distance ranges from 276–280 in the case of places, from 143–298 for urban areas, and from 309–370 for CBSAs.

CBSAs). Now, we construct the same number of samples starting from each city. But instead of including nearby cities, we draw exactly the same number of random cities without replacement from the whole city size distribution, regardless of the physical bilateral distances. Single-city samples are excluded again. Then, using the Gabaix and Ibragimov (2011) specification (Equation (2)), we estimate the Pareto exponent for all these random samples of cities. Note that sample size is the same in random and geographical samples, but they only share one common element: the initial core city. Finally, we compute the difference between the previously estimated Pareto exponent from the geographical samples and the placebo Pareto exponent obtained from random samples. Therefore, for each city we obtain 58 values of the difference between the Pareto exponents estimated using geographical and random samples (51 in the case of CBSAs).[19] Alternatively, from the sample size view, for each number of cities we carry out an average number of 267, 291, and 233 replications in the case of places, urban areas, and CBSAs, respectively.

The results are summarised by conducting a nonparametric estimation of the relationship between distance and the difference between the Pareto exponents estimated using geographical and random samples using local polynomial smoothing. Figure 5 shows the results, including the 95% confidence bands. Note that this time the *x*-axis represents sample size rather than distance. For small sample sizes, the difference between Pareto exponents estimated using geographical and random samples is positive but decreases with sample size. In the case of urban areas, the difference is not significant for sample sizes smaller than 50 cities. Nevertheless, as sample size increases, the difference stabilises around a positive value that is significantly different from 0 for each of the three city definitions.

[19] Therefore, the numbers of values is the same than those used previously to obtain Figure 2: 1,665,962 values for places, 204,959 for urban areas, and 45,912 for CBSAs.

The interpretation of a significant positive difference between the Pareto exponents estimated using geographical and random samples is that geography has a significant effect on the value of the Pareto exponent. This effect is not just the consequence of a larger or smaller sample size: Pareto exponents estimated using geographical samples of nearby cities are (on average) higher than those obtained with random samples of cities. This finding indicates that neighbouring cities are more homogeneous in city sizes than random samples of cities. Using data from the US, Hsu et al. (2014) also find significant differences in the results obtained from spatial partitions of cities and random partitions.

5. Discussion

The spatial distribution of population has deep economic and social implications. Economists, statisticians, physicists, and geographers have all pointed to the Pareto distribution as a benchmark distribution. In recent years, after an enriching debate, studies from the mainstream literature have been updated to a new paradigm that states that, although most of the city size distribution is nonlinear, the Pareto distribution (and Zipf's law) holds *for the largest cities* (Levy, 2009; Giesen et al., 2010; Ioannides and Skouras, 2013).

This paper questions this statement. Large cities are typically far away from one another; it is not clear whether we can use theoretical spatial equilibrium models to explain the largest cities as part of an entire city size distribution, and what means that the Pareto distribution (and Zipf's law) holds for these largest cities because they are almost independent elements. Rather than focusing on city size, as most studies do, we analyse the validity of the Pareto distribution from a spatial perspective, and we propose a new distance-based approach. This new methodology enables us to confirm that:

(1) Using all possible combinations of cities within a 300-mile radius, our results indicate that the Pareto distribution cannot be rejected in most cases regardless of city size and city definition. Therefore, the Pareto distribution fits the city size distribution well for cities of all sizes as long as they are nearby. Thus, we emphasise that the proper statistical function of city size distribution is a matter of distance rather than size.

(2) Eeckhout (2004) concluded that, when all US cities are included with no size restriction, city size distribution is actually lognormal rather than Pareto. This assertion may be right for the whole city size distribution,[20] but, as we argue in the Introduction section, the city size distribution for all cities includes elements without any spatial relationship. Therefore, finding support in urban theory for this analysis is difficult. Our results show that the Pareto distribution cannot be either discarded or confined to the upper-tail analysis; it is valid for cities of all sizes as long as they are close, which implies that there should be a meaningful spatial relationship between cities (as theoretical models assume).

(3) Zipf's law only emerges for urban areas and CBSAs at a very particular range of distances (50–75 miles for urban areas and 75–80 miles in the case of CBSAs). For the rest of distances, an exponent with a value of 1 falls outside the confidence bands. We can accordingly reject Zipf's law for most distances. Regarding places, the estimated Pareto exponents are always lower than 1. Therefore, some evidence supporting Zipf's law can only be found for the aggregate geographical units but not for places that are the lowest spatial unit considered. The literature highlights city definition (Rosen and Resnick, 1980; Cheshire, 1999; Soo, 2005) as a crucial issue, along with sample size and the choice of the estimator. Our spatial perspective adds a new factor influencing the value of the estimated Pareto exponent that has not been considered before in the

[20] Although some authors find that other nonlinear distributions fit city size data better than a lognormal distribution (Reed, 2002; Giesen et al., 2010; González-Val et al., 2015; Ioannides and Skouras, 2013; Giesen and Suedekum, 2014).

literature: distance. Nevertheless, because Zipf's law cannot be rejected for only a small range of distances, the validity of this law may be called into question. Do the results from this spatial approach imply that Zipf's law is a ghost statistic regularity or even that the law has become obsolete? The key point is whether the ranges of distances for which the law is valid have any economic or spatial meaning rather than whether the law holds for a large set of distances. Unfortunately, this question remains open because evidence of the spatial limits of urban systems is not conclusive. Only a few studies have explored this issue (Hsu et al., 2014; González-Val, 2018a) because, as Pumain (2006) points out, systems of cities are difficult to isolate as scientific objects of study.

(4) We run some robustness checks, including placebo regressions, and we show that there is a significant effect of geography on the Pareto exponent for the three city definitions: Pareto exponents estimated using geographical samples of nearby cities are (on average) higher than those obtained with random samples of cities. This finding indicates that neighbouring cities, which share economic and trade interactions, commuting, and migratory flows, are more homogeneous in city sizes than random samples of cities.

These findings also imply important characteristics for urban hierarchies and the spatial organisation of cities. First, the regular hierarchical differentiation of urban systems is typically summarised by a Pareto-like or lognormal distribution of city size (Pumain, 2006); the Pareto distribution suggests complex systems of cities, and the lognormal distribution indicates simple systems of cities. Our results support the Pareto distribution for geographical samples of nearby cities, thereby confirming complex systems of cities. According to Chen (2011), this complexity can be external (at the macro level) and/or internal (at the micro level).

Second, this type of complex systems of cities involves a hierarchy that is statistically self-similar and hence fractal (Batty, 2006). Therefore, urban systems are a kind of hierarchy with a cascade structure, similar to other hierarchies observed in nature, such as the hierarchy of rivers and the energy distributions of earthquakes (Chen and Zhou, 2008). These hierarchies can be described with a set of exponential laws from which we can derive a set of power laws indicating hierarchical scaling in cities. As Chen (2016) demonstrates, all types of Zipf models can be transformed into the corresponding hierarchies with a cascade structure.

6. Conclusion

This paper uses data from three different definitions of US cities in 2010 (places, urban areas, and CBSAs) to introduce a new distance-based approach with the aim of analysing the influence of distance on the city size distribution Pareto exponent using all possible combinations of cities within a 300-mile radius.

Our results lend support to the Pareto distribution, which cannot be rejected in most cases regardless of city size and city definition. Our findings have deep implications for urban hierarchies and the spatial organisation of cities and raise new questions about the spatial limits of urban systems. These questions, in our opinion, deserve more attention from spatial researchers.

Acknowledgements

Financial support was provided by the Spanish Ministerio de Economía y Competitividad (ECO2017-82246-P and ECO2016-75941-R projects), the DGA (ADETRE research group) and FEDER. The assistance of Miriam Marcén with the Stata code is greatly appreciated. The author benefited from the helpful suggestions from David Cuberes and Fernando Sanz-Gracia. All remaining errors are the author's

alone. This paper was previously circulated under the title “"City Size Distribution and Space”, Working Papers 2017/18, Institut d'Economia de Barcelona (IEB).

References

Bak, P. 1996. *How Nature Works: the Science of Self-organized Criticality*. New York: Springer.

Batty, M. 2006. “Cities and City Systems.” In: *Hierarchy in Natural and Social Sciences*, Methodos Series, Vol. 3, D. Pumain ed., Springer: Dordrecht, 143–168.

Berry, B. J. L., and A. Okulicz-Kozaryn. 2012. “The city size distribution debate: Resolution for US urban regions and megalopolitan areas.” *Cities* 29: S17–S23.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. “Power-law distributions in empirical data.” *SIAM Review* 51(4): 661–703.

Chen, Y. 2011. “Modeling Fractal Structure of City-Size Distributions Using Correlation Functions.” *PLoS ONE* 6(9): e24791.

Chen, Y. 2016. “The evolution of Zipf’s law indicative of city development.” *Physica A-Statistical Mechanics and its Applications* 443: 555–567.

Chen, Y., and Y. Zhou. 2008. “Scaling laws and indications of self-organized criticality in urban systems.” *Chaos, Solitons & Fractals* 35(1): 85–98.

Cheshire, P. 1999. “Trends in sizes and structure of urban areas.” In: *Handbook of Regional and Urban Economics*, Vol. 3, edited by P. Cheshire and E. S. Mills, 1339–1373. Amsterdam: Elsevier Science.

Cottineau, C. 2017. “MetaZipf. A dynamic meta-analysis of city size distributions.” *PLOS ONE* 12(8): e0183919.

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.

Fujita, M., P. Krugman, and T. Mori. 1999. “On the evolution of hierarchical urban systems.” *European Economic Review* 43: 209–251.

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

Gabaix, X. 2009. “Power laws in economics and finance.” *Annual Review of Economics* 1: 255–294.

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.” In: *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., and J. Südekum. 2014. “City Age and City Size.” *European Economic Review* 71: 193–208.

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.

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. 2012. "Zipf's law: Main issues in empirical work." *Région et Développement* n° 36-2012, 147–164.

González-Val, R. 2017. "City Size Distribution and Space." Working Papers 2017/18, Institut d'Economia de Barcelona (IEB).

González-Val, R. 2018a. "The spatial distribution of US cities." Cities, forthcoming.

González-Val, R. 2018b. "Historical urban growth in Europe (1300–1800)." Papers in Regional Science, forthcoming.

González-Val, R., A. Ramos, F. Sanz-Gracia, and M. Vera-Cabello. 2015. "Size distributions for all cities: which one is best?" *Papers in Regional Science* 94(1): 177–196.

Hsu, W.-T. 2012. "Central place theory and city size distribution." *The Economic Journal* 122: 903–932.

Hsu, W.-T., T. Mori, and T. E. Smith. 2014. "Spatial patterns and size distributions of cities." Discussion paper No. 882, Institute of Economic Research, Kyoto University.

Ioannides, Y. M., and H. G. Overman. 2004. "Spatial evolution of the US urban system." *Journal of Economic Geography* 4(2): 131–156.

Ioannides, Y. M., and S. Skouras. 2013. "US city size distribution: Robustly Pareto, but only in the tail." *Journal of Urban Economics* 73: 18–29.

Krugman, P. 1996a. “Confronting the Mystery of Urban Hierarchy.” *Journal of the Japanese and International Economies* 10: 399–418.

Krugman, P. 1996b. *The Self-organizing economy*. Cambridge: Blackwell.

Lalanne, A. 2014. “Zipf's Law and Canadian Urban Growth.” *Urban Studies* 51(8): 1725–1740.

Le Gallo, J., and C. Chasco. 2008. “Spatial analysis of urban growth in Spain, 1900-2001.” *Empirical Economics* 34: 59–80.

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

Luckstead, J., and S. Devadoss. 2017. “Pareto tails and lognormal body of US cities size distribution.” *Physica A: Statistical Mechanics and its Applications* 465: 573–578.

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

Newman, M. E. J. 2006. “Power laws, Pareto distributions and Zipf’s law.” *Contemporary Physics* 46: 323–351.

Nitsch, V. 2005. “Zipf zipped.” *Journal of Urban Economics* 57: 86–100.

Pisarenko, V. F. 1998. “Non-linear Growth of Cumulative Flood Losses with Time.” *Hydrological Processes* 12(3): 461–470.

Pumain, D. 2006. “Alternative Explanations of Hierarchical Differentiation in Urban Systems.” In: *Hierarchy in Natural and Social Sciences*, Methodos Series, Vol. 3, D. Pumain ed., Springer: Dordrecht, 169–222.

Rauch, F. 2014. “Cities as spatial clusters.” *Journal of Economic Geography* 14(4): 759–773.

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

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

Roberts, D. C., and D. L. Turcotte. 1998. “Fractality and Self-organized Criticality of Wars.” *Fractals* 6(4): 351–357.

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.

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.

Williams, G. P. 1997. *Chaos Theory Tamed*. Washington, D.C.: Joseph Henry Press.

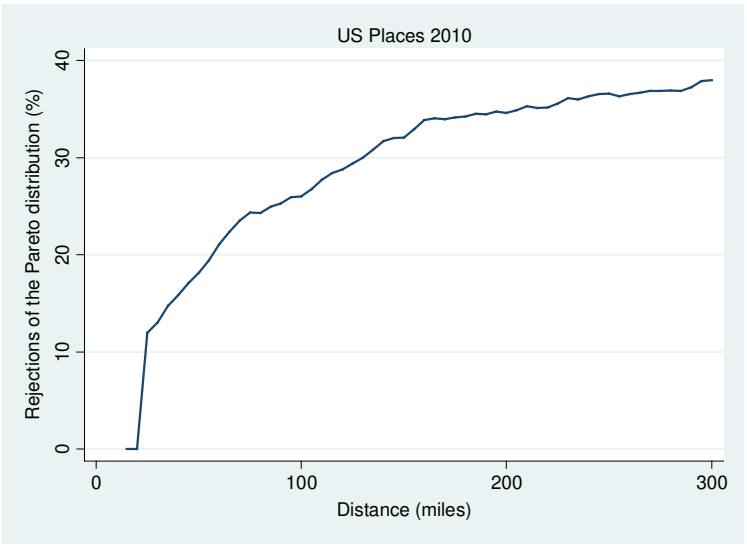
Zipf, G. 1949. *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Table 1. Descriptive statistics

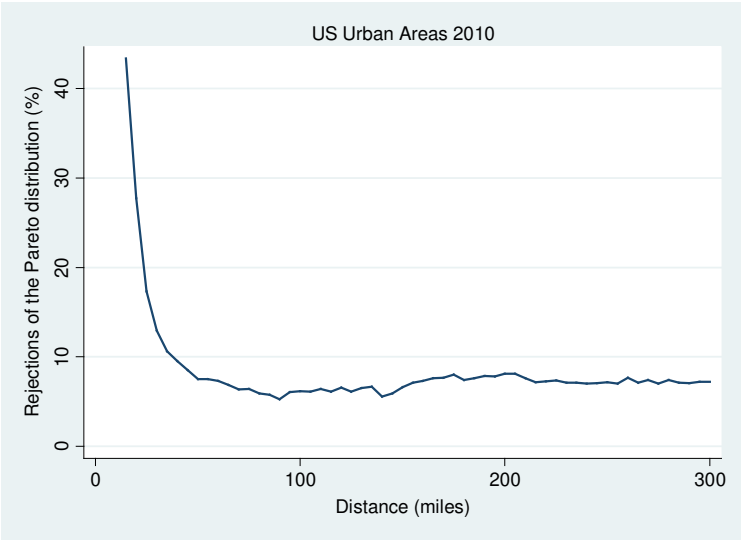
City definition	Cities	Mean size	Standard deviation	Minimum	Maximum	% of US population
Places	28,738	7,880.2	66,192.9	1	8,175,133	73.3%
Urban areas	3,592	70,363.7	495,447.5	2,500	18,351,295	81.9%
Core-based statistical areas	929	310,836.9	1,056,227.6	13,477	19,567,410	93.9%

Notes: Source: US Census 2010.

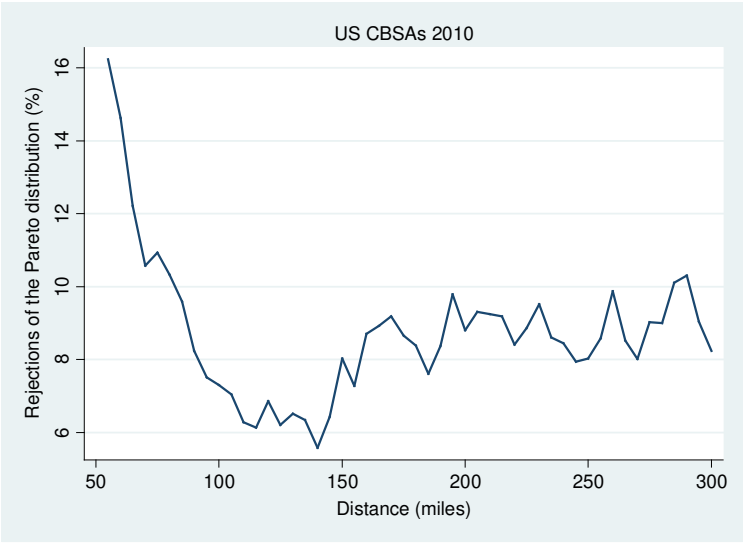
Figure 1. Pareto distribution test over space



(a) Places



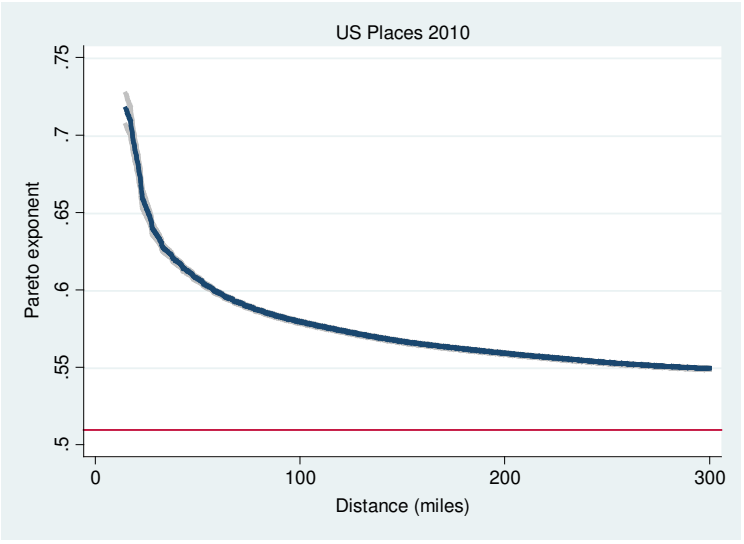
(b) Urban areas



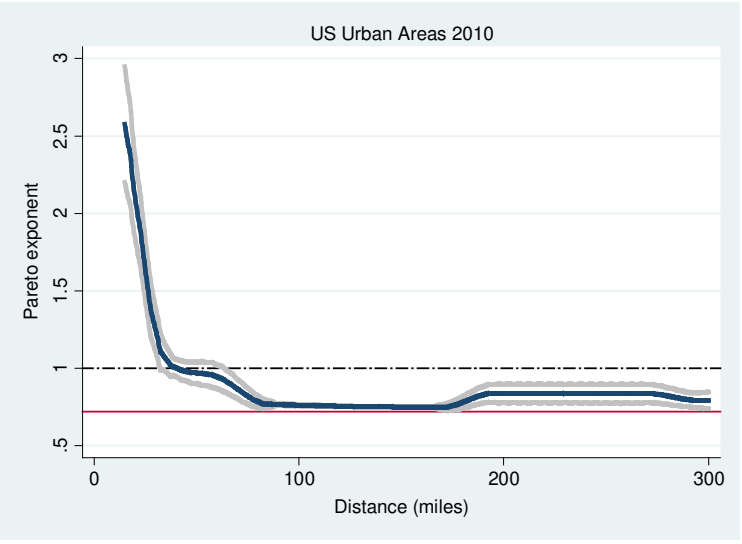
(c) CBSAs

Notes: Percentage of rejections of the goodness-of-fit test proposed by Clauset et al. (2009) at the 10% level.

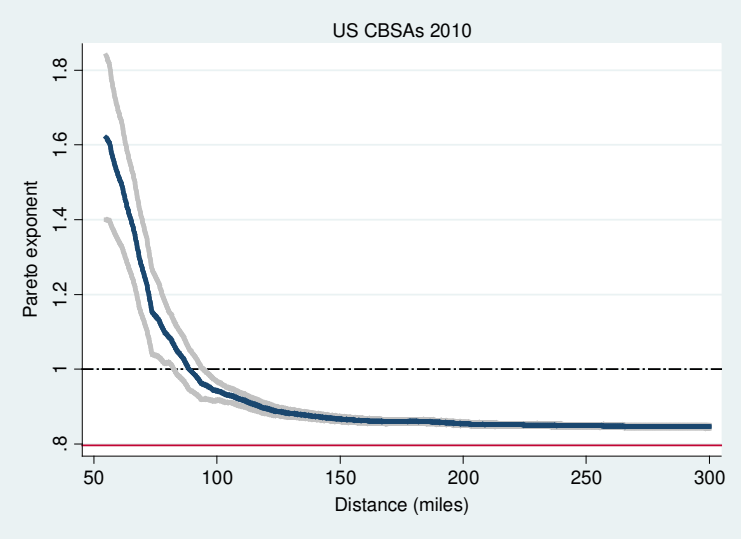
Figure 2. Pareto exponent by distance



(a) Places



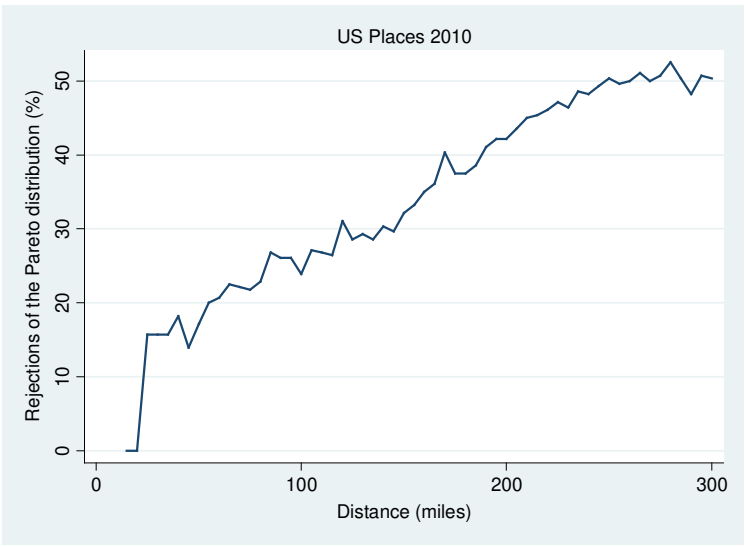
(b) Urban areas



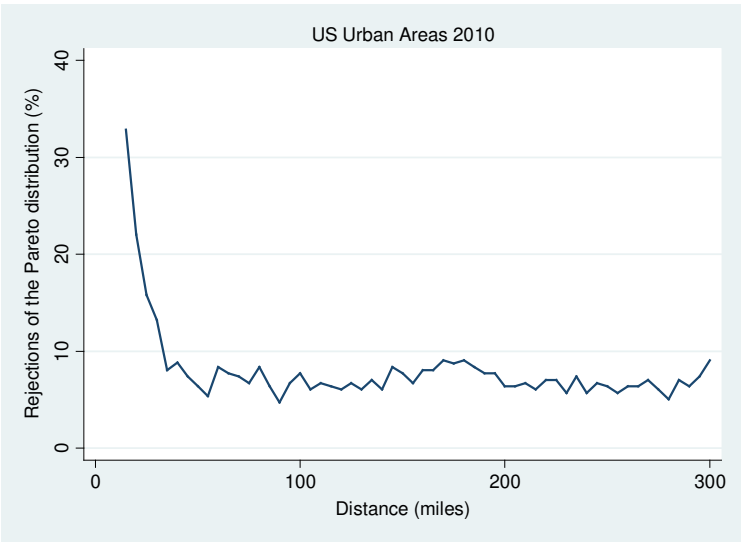
(c) CBSAs

Notes: Nonparametric relationship between distance and the estimated Pareto exponents including the 95% confidence intervals, based on 1,665,962 (Figure (a)), 204,959 (Figure (b)), and 45,912 (Figure (c)) Pareto exponent-distance pairs. The horizontal line indicates the estimated Pareto exponent for the entire sample of cities.

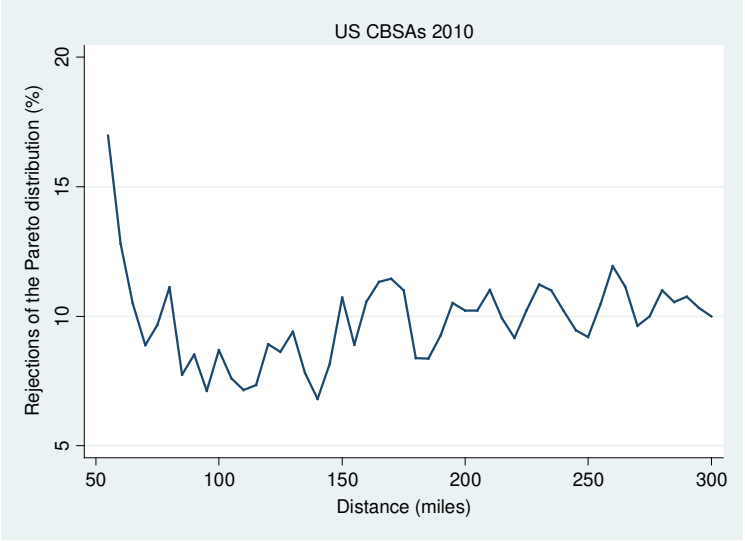
Figure 3. Pareto distribution test over space for geographical samples with a large core city (>100,000 inhabitants)



(a) Places



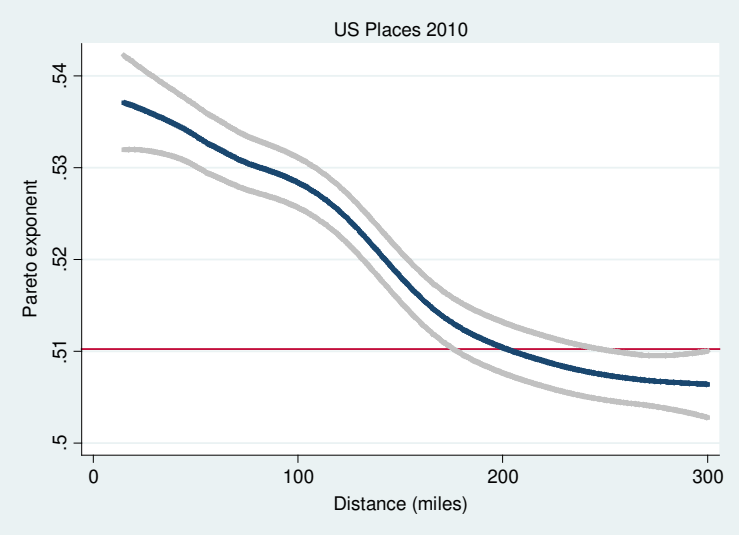
(b) Urban areas



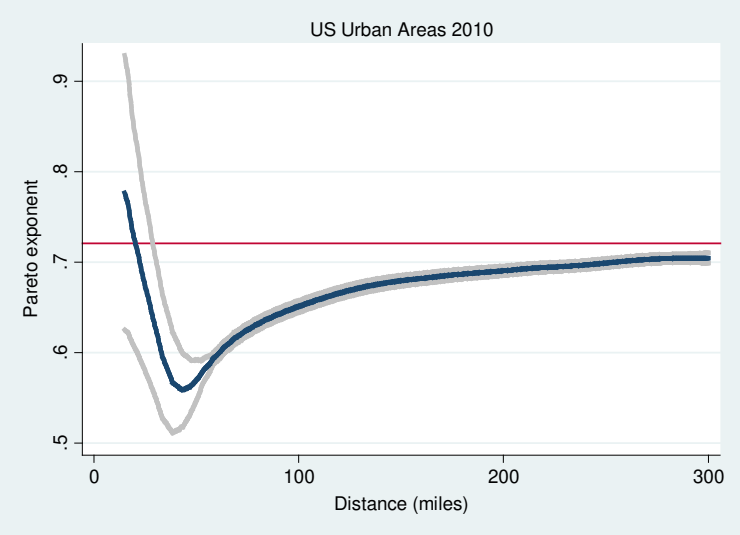
(c) CBSAs

Notes: Percentage of rejections of the goodness-of-fit test proposed by Clauset et al. (2009) at the 10% level.

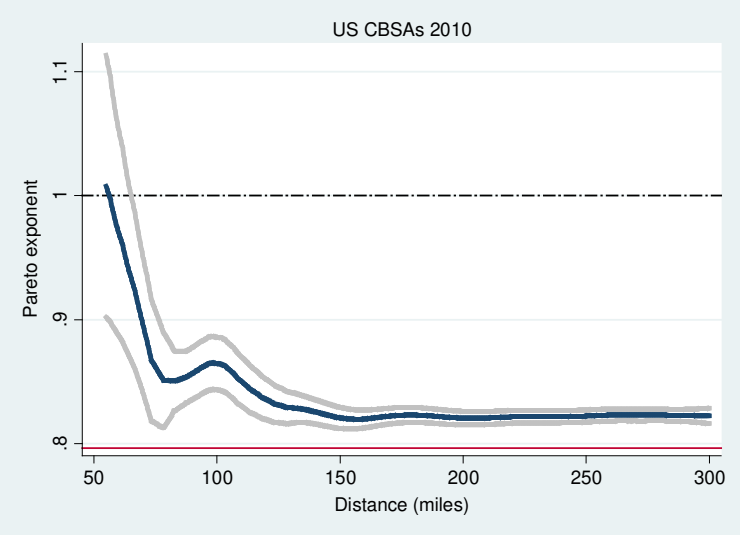
Figure 4. Pareto exponent by distance for geographical samples with a large core city (>100,000 inhabitants)



(a) Places



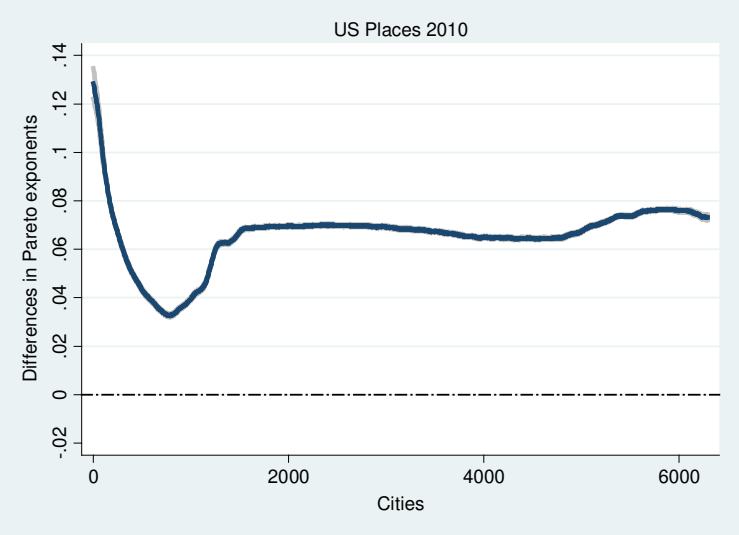
(b) Urban areas



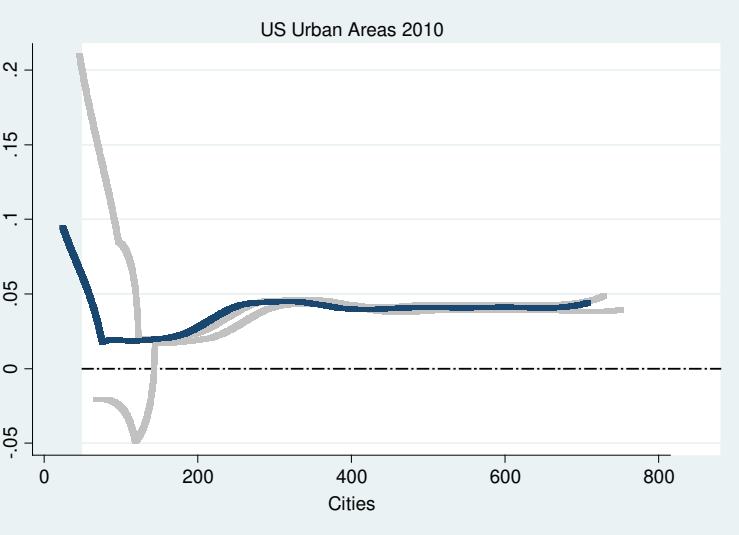
(c) CBSAs

Notes: Nonparametric relationship between distance and the estimated Pareto exponents including the 95% confidence intervals, based on 16,237 (Figure (a)), 17,034 (Figure (b)), and 18,530 (Figure (c)) Pareto exponent-distance pairs. The horizontal line indicates the estimated Pareto exponent for the entire sample of cities.

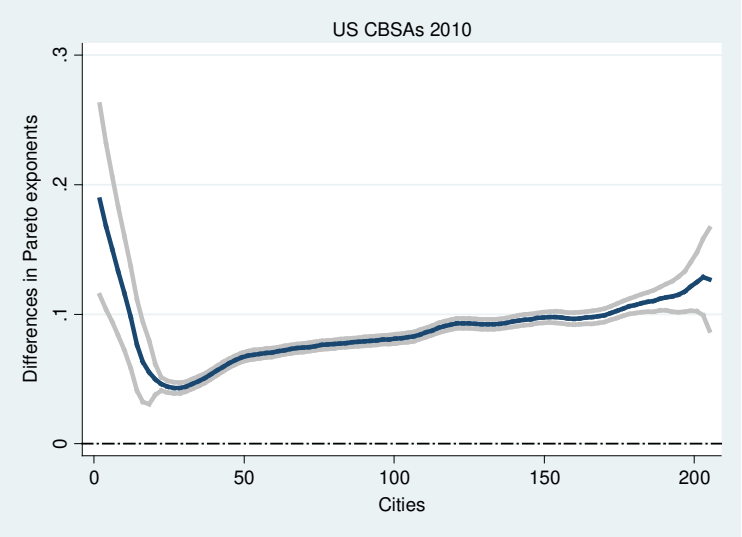
Figure 5. Placebo regressions: differences in Pareto exponents between geographical samples and random samples by sample size



(a) Places



(b) Urban areas



(c) CBSAs

Notes: Nonparametric relationship between distance and the difference between Pareto exponents estimated using geographical and random samples, including the 95% confidence intervals, based on 1,665,962 (Figure (a)), 204,959 (Figure (b)), and 45,912 (Figure (c)) observations.