Lognormal city size distribution and distance

González-Val, Rafael

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

April 2019

Lognormal city size distribution and distance

Rafael González-Val

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

***Abstract*:** This paper analyses whether the size distribution of nearby cities is lognormally distributed by using a distance-based approach. We use data from three different definitions of US cities in 2010, considering all possible combinations of cities within a 300-mile radius. The results indicate that support for the lognormal distribution decreases with distance, although the lognormal distribution cannot be rejected in most of the cases for distances below 100 miles.

***Keywords*:** space, city size distribution, distance-based approach, lognormal distribution.

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

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.

1. Introduction

The study of city populations has deep economic and social implications; for instance, the population of cities is related to the extension of local labour markets and the intensity of internal migrations. The lognormal distribution has been considered, for many years, to study city size along with the Pareto distribution. More recently, Eeckhout (2004) was the first to argue the importance of considering all cities without size restrictions and fitted the lognormal distribution to un-truncated US city size data. He also developed a theoretical model of local externalities with a lognormal distribution of city sizes in equilibrium.

Eeckhout's (2004) influential paper stimulated an academic debate about what distribution provides a better fit to actual un-truncated data, and the lognormal distribution soon was replaced by other more convoluted distributions. The current consensus is that the best city size distribution may be a mixed distribution, separating the lognormal body of the distribution from the upper Pareto tail (Reed, 2001; Ioannides and Skouras, 2013; Giesen et al., 2010). Thus, the lognormal distribution can still be useful to fit the body of the distribution, which includes most cities; however, as González-Val (2019a) argues, this mass of cities includes many elements without any spatial relationship because they are very far from one another. Adopting a spatial perspective, the purpose of this study is to clarify whether the size distribution of nearby cities is lognormally distributed by using a new distance-based approach.

2. Data

As González-Val (2019a, 2019b) did, we use three different definitions of US cities: places, urban areas, and core-based statistical areas. Our data come from the 2010 US decennial census, and the geographical coordinates (latitude and longitude) needed

to compute the bilateral distances between cities were obtained from the 2010 Census US Gazetteer files.

The US Census Bureau uses the generic term 'places' to include, since the 2000 census, all incorporated and unincorporated places. 'Incorporated places' are administratively defined cities and 'unincorporated places' (also known as Census Designated Places since 1980) designate a densely settled concentration of population that is not within an incorporated place, but is locally identified by a name. The number of places considered in this study is 28,738 and the populations range from 8,175,133 (New York City, NY) to 1 (Monowi village, NE).

According to the US Census Bureau, 'urban area' is the generic term for urbanized areas and urban clusters. Both consist of densely developed areas; urbanized areas contain 50,000 or more people, whereas urban clusters have at least 2,500 but fewer than 50,000 people. Therefore, the minimum population in the sample is the minimum population threshold (2,500) and the maximum population is 18,351,295 (New York-Newark, NY-NJ-CT). The total number of urban areas is 3,592.

Finally, 'core-based statistical areas' (CBSAs) are defined by the US Census Bureau as the county, counties, or equivalent entities associated with at least one core, an urban area (urbanized area or urban cluster) with a population of at least 10,000, plus adjacent counties with a high degree of social and economic integration with the core as measured through commuting ties. If the core is an urbanized area (urban cluster) the CBSA is called a 'metropolitan statistical area' ('micropolitan statistical area'). The number of CBSAs in the sample is 929 with populations from 19,567,410 (New York-Newark-Jersey City, NY-NJ-PA) to 13,477 (Ketchikan, AK).

The sample of places is un-truncated, whereas urban areas and CBSAs impose different minimum population thresholds. Moreover, these city definitions are nested;

most places are included in urban areas and most urban areas and places are located inside CBSAs.

3. Methodology

We apply the new distance-based approach of González-Val (2019a). In this new methodology, space is introduced through the selection of geographical samples (or sub-regions) of cities based on distances. This recursive procedure is based in the following steps:

1. Calculate the bilateral physical geographic distances between all cities.[1]
2. Define the limit of the geographical samples of neighbouring cities. Following an agnostic view, we consider all the possible combinations of cities within a 300-mile radius. As González-Val (2019a) argued, this choice of the threshold is based on a conservative criterion; Rauch (2014) found that most of the people in the US (over 68% of observations) live between 0 and 100 km from their birthplace.
3. Draw circles of radius $r = 15,20,...,300$ around the geographic centroid of each city's coordinates, starting from a minimum distance of 15 miles and adding 5 miles for each subsequent iteration. For CBSAs, the procedure starts at 50 miles because for short distances there are very few elements as CBSAs are large spatial units that cover huge areas.
4. Repeat this exercise for all cities.
5. Fit the lognormal distribution to each geographical sample and run a lognormality test. The standard test to check the lognormal behaviour of a sample is the Kolmogorov-Smirnov (KS) test, which was previously used with city sizes by Giesen et al. (2010), among others. The KS test's null hypothesis is

[1] We use the haversine distance.

that the two samples (the actual data and the fitted lognormal distribution) come from the same distribution.

4. Results

Figure 1 shows the result of the KS test by distance. For each distance, the graphs represent the percentage of p-values lower than 0.05 (the standard significance level) over the total number of tests carried out at that distance (single-city samples are excluded).[2] One well-known inconvenience of the KS test is its relatively low power: with very high sample sizes, it tends to systematically reject the null hypothesis unless the fit is almost perfect. Therefore, one might expect that the power of the test decreases with distance as the sample size increases, and our results indeed show that support for the lognormal distribution clearly decreases with distance for the three samples.

In the case of places and urban areas, for distances longer than 100 miles, the percentage of rejections soon increases to higher than 50%, and for the longest distances, the test rejects the lognormal distribution in most cases (almost 80% and 100% for places and urban areas, respectively). Nevertheless, for CBSAs, the percentage of rejections is significantly lower for all distances and only at the longest distances (higher than 250 miles) is the percentage higher than 50% (but lower than 60% in all cases). Thus, although rejections of the lognormal hypothesis depend crucially on the city definition (a classical issue in this literature; see Rosen and Resnick, 1980), for short distances (below 100 miles) the lognormal distribution is valid for the three city definitions.

Next, we estimate the lognormal distribution parameters. The ML estimators for the mean (μ) and standard deviation (σ) are, respectively, the mean and standard deviation of the logarithm of the data. This gives us 1,666,804 mean- and standard

[2] The number of tests carried out by distance ranges from 28,250–28,738 in the case of places, from 2,173–3,591 for urban areas, and from 796–929 for CBSAs.

deviation-distance pairs for places, 208,336 for urban areas, and 46,450 for CBSAs. To summarize all these values, we estimate the nonparametric relationship distance-mean and distance-standard deviation using a local polynomial smoothing. The panels in Figure 2 display the results, including the 95% confidence intervals. There are no important differences in the behaviour of the parameters between places, urban areas, and CBSAs. In the three cases, the mean decreases with distance, whereas the standard deviation increases with distance. It is important to note that these average values by distance soon converge to the mean and standard deviation for the whole sample, represented by the horizontal line.

5. Conclusion

We find that the lognormal distribution cannot be rejected in most of the cases for distances below 100 miles, considering three different US city definitions and all possible combinations of cities. Thus, the lognormal distribution is not only suitable to fit the body of the city size distribution containing many (spatially independent) middle-sized cities, but from a spatial perspective it can also fit the city size distribution for nearby cities of all sizes. Nevertheless, González-Val (2019a) shows that the Pareto distribution is valid for distances longer than 100 miles, thus outperforming the lognormal distribution for sub-regions covering large areas.

References

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

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. 2019a. “US city size distribution and space.” Spatial Economic Analysis, forthcoming.

González-Val, R. 2019b. “The spatial distribution of US cities.” Cities, forthcoming.

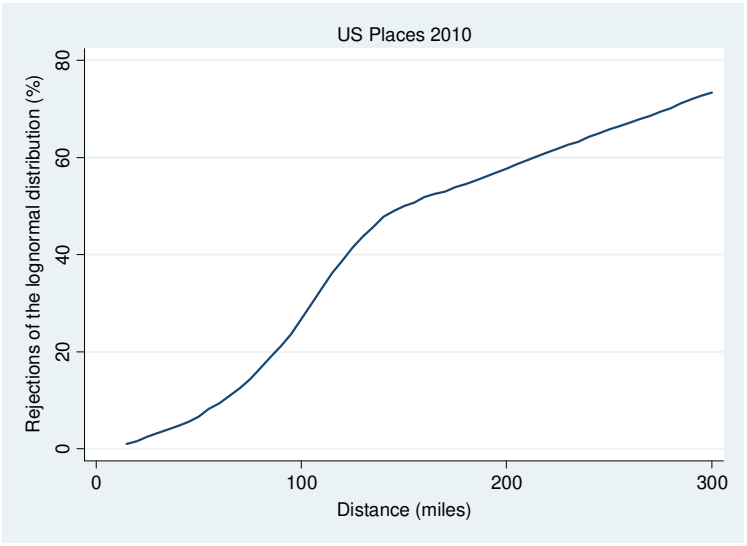
Ioannides, Y. M., and S. Skouras. 2013. “US city size distribution: Robustly Pareto, but only in the tail.” *Journal of Urban Economics* 73: 18–29.

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

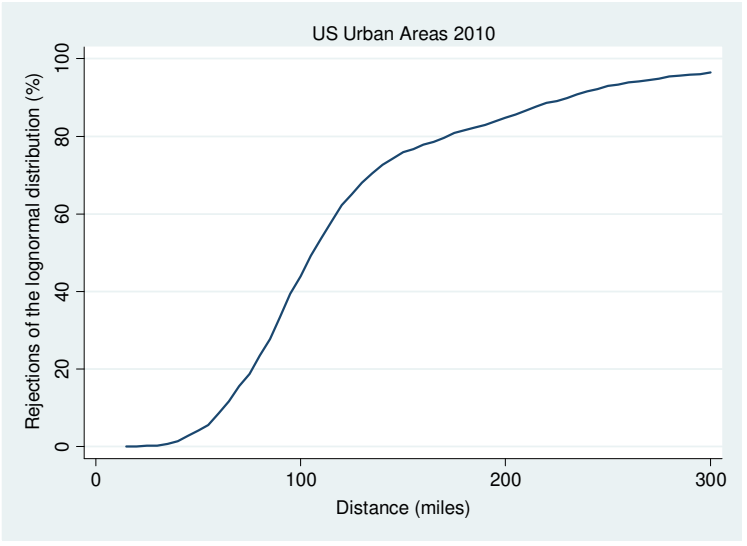
Reed, W. J. 2001. “The Pareto, Zipf and other power laws.” *Economics Letters* 74: 15–19.

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.

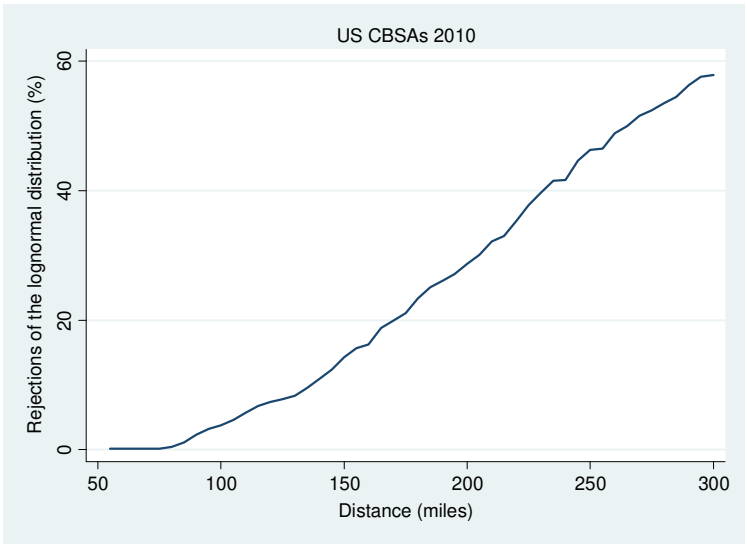
Figure 1. Lognormal distribution test over space



(a) Places



(b) Urban areas

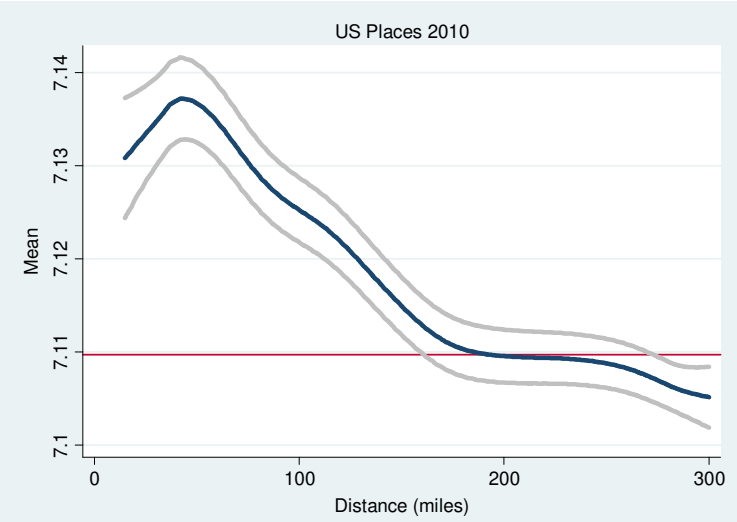


(c) CBSAs

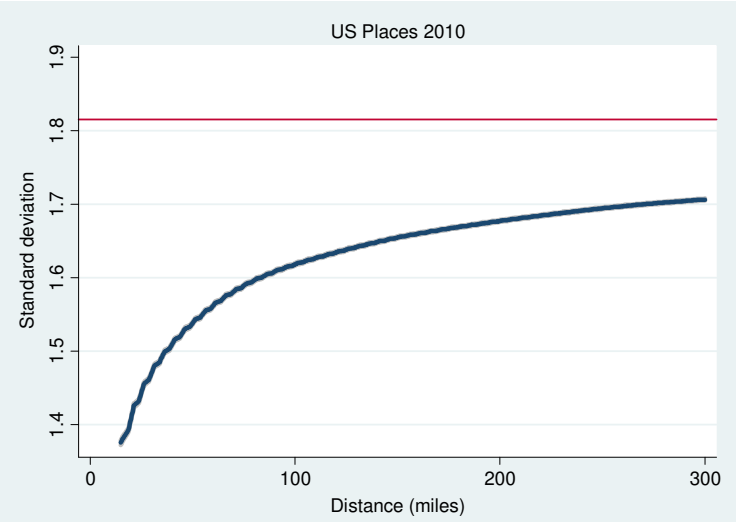
Notes: Percentage of rejections of the Kolmogorov-Smirnov test of the lognormal distribution at the 5% level.

Figure 2. Lognormal distribution over space: mean and standard deviation

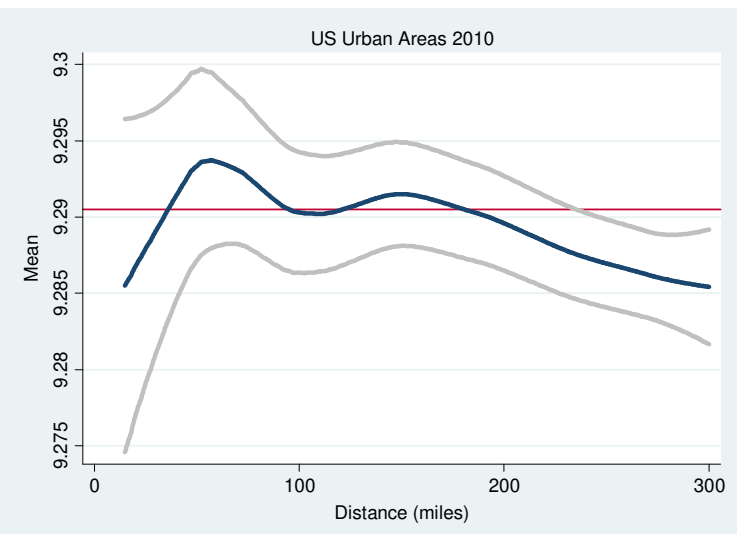
(A) Mean **(B) Standard deviation**



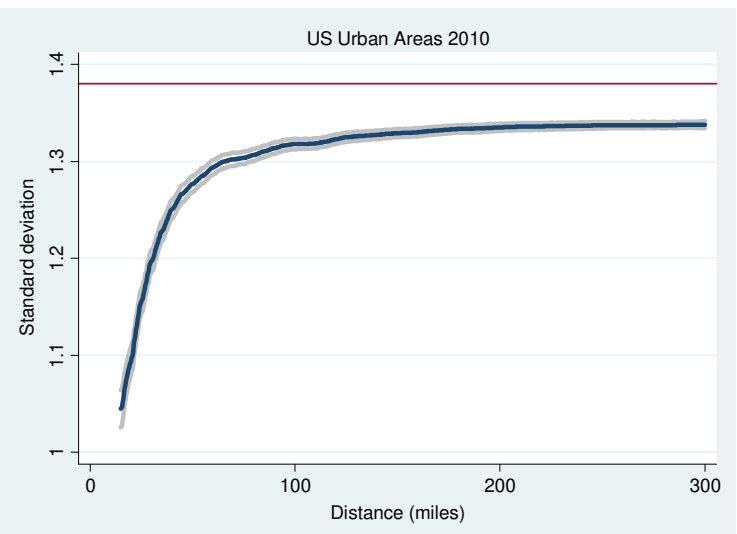
(a) Places



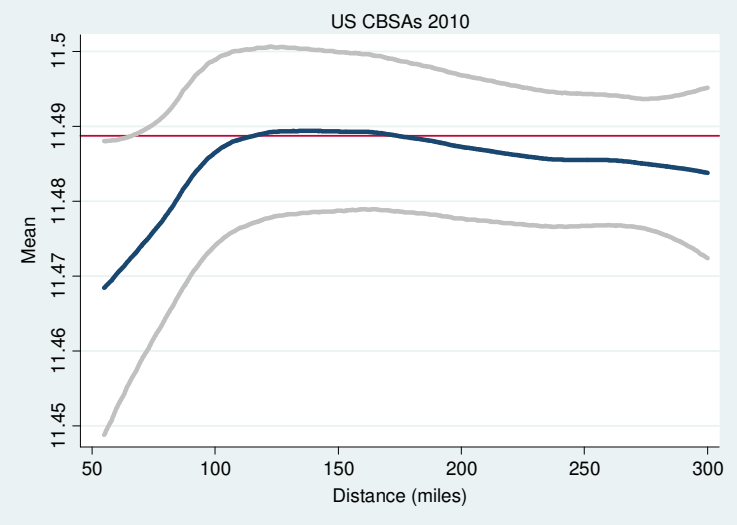
(b) Places



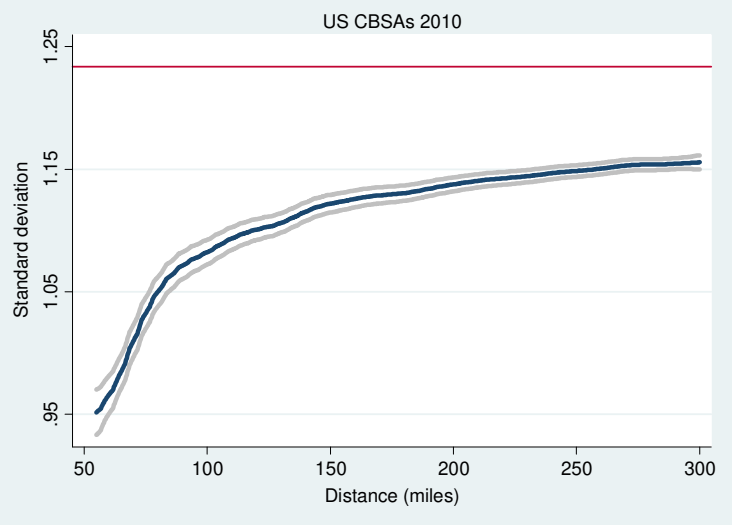
(c) Urban areas



(d) Urban areas



(e) CBSAs



(f) CBSAs

Notes:

Figures show the nonparametric relationship distance-mean and standard deviation-mean, respectively, including the 95% confidence intervals, based on 1,666,804 (Figures (a) and (b)), 208,336 (Figures (c) and (d)) and 46,450 (Figures (e) and (f)) observations. The horizontal lines represent the values for the whole sample of cities.