

DOES THE SIZE MATTER? Zipf's Law for Cities Revisited

Josep Roca Cladera & Blanca Arellano Ramos¹

1.- Summary

Several authors (Berry 1970, Krugman 1996 or Eaton and Eckstein 1997, among many others) have experienced amazement about the accurate functioning of the law of "least effort" established by Zipf (1949) in most places. Cities, ranked by population, seem to follow almost exactly a log/log function, in which the logarithm of the "mass" (population, density, number of employees, etc.) correlates almost perfectly with the logarithm of the order of that mass. This log/log function, advanced by Pareto in the nineteenth century, has seduced quite a number of researchers, for its presence, hypothetically, both in natural phenomena (earthquakes, meteorites, living species, ...) and in the ones which derive from society (language, or distribution of cities), which has led to investigate its theoretical basis (Simon 1955, Brakmar et al. 1999, Gabaix 1999).

While some authors (Rosen and Resnick 1980, Fan and Casetti 1994) have discussed the linear validity of Zipf's Law, introducing nonlinear models, technical literature has focused on the "upper tail" of the urban hierarchy, large cities or metropolitan areas, tend to silence the fact that the log / log function does not appear to be a general model. This paper attempts to show that when taking into account all the cases (ie, all populated localities in a particular territory), the log/log model seems to be only a special case of "the big". In fact it shows that a log/lin model tends to be more efficient, even with "folded tails." This has led to the hypothesis which was tested in this study, that the logarithm of the urban mass tends to have a "normal distribution", leading its cumulative distribution (and ordered by rank) to be spread in a logistical structure, in "S".

In this sense, the repeated observation of fulfillment of the Law of Zipf in the size of the cities would be just "the tip of the iceberg", in which small and medium cities also take their part, and where a "law" of a higher level appears.

The presented research questions if this "normal" appearance of the logarithm of the mass could be shaped in a simple and elegant form, and makes some experiments in this regard.

2.- Introduction

One issue that has attracted mostly urban specialists, particularly economists, consisted in understanding of the spatial hierarchy inherent in the size distribution of cities. It has been well known since decades that the distribution of large cities in many places of the world can

¹ Polytechnic University of Catalonia

be described through an exponential law, explained in its most popular form by Zipf (1949²), according to which the number of cities with a population greater than “P” is roughly proportional to P^{-a} , being “a” very close to 1. Studying the frequency of use of English words, in the relation of their complexity (the number of letters), Zipf came to formulate a "law", which he called "the law of minimal effort", according to which the frequency of a word, P_n , arranged in the order n would have a frequency equal to:

$$P_n \sim 1/n^a$$

This "law" has been used for many natural and artificial phenomena, from the frequency of earthquakes pursuant to its size, to the size of the cities.

The reason why the size distribution of cities follows the "law" of Zipf has intrigued not a few theorists (Simon, 1955, Henderson, 1974; Krugman, 1996; Brakmar et al., 1999, Gabaix, 1999³), which led Krugman (1999) to say:

“At this point, we do not dispose of an explanation of the astonishing regularity of the size distribution of cities. We have to admit that this fact presumes a real intellectual challenge.”⁴

The origin of the "law" described by Zipf seems to be found in the study of rent distribution carried out by Pareto (1896), and according to which occurs the known effect "80-20."⁵ Already in 1913, Auerbach (1913) proposed that the size distribution of cities could be very close to the Pareto distribution. So if we rank the cities from the largest (rank 1) to the smallest (rank N), the range of a city population P, $r(P)$, would be:

$$r(P) = AP^{-\alpha}$$

In logarithms:

$$\ln r(P) = \ln A - \alpha \ln p$$

Being $\alpha = 1$, as a particular case of the Pareto distribution, interpreted by Zipf.

The empirical work conducted over several decades (Berry, 1961, Berry & Horton, 1970, Rosen & Resnick, 1980, Carroll, 1982; Guérin-Pac, 1995, Eaton & Eckstein, 1997, Chesire, 1999; Dobkins & Ioannides, 2000) seems to conclude that, *in large cities*, the distribution of cities, generally, follows the Pareto distribution. Regarding to $\alpha = 1$, some authors, especially Krugman (1996), have strongly defended the validity of Zipf's thesis, while others, such as Alperovich (1993), have rejected it.

In Spain also, Lasuén (1967) first, and Lanaspá et al. (2004) more recently, have confirmed the validity of Pareto's exponential distribution. According to the last studies, that have

² Already in 1682, Alexandre Le Maître recognized the existence of a clear structure in the distribution of city size in France. But it was not until 1913 that Felix Auerbach formally established the structure of the mathematical relationship, Zipf would later generalize the exponential function with exponent -1.

³ Since Simon (1955) has come to discuss the link between Pareto distribution with the principle developed by Gibrat (1931), according to which there is no relationship between the rate of growth (of the cities in our case) and the initial size, for which reason no regular behavior can be deduced. See, among others, Gabaix (1999).

⁴ See Fujita, M., Krugman, P. & Venables, A.J. (1999), p. 223, the Spanish edition of 2000.

⁵ I.e. that 20% of the population gather 80% of wealth.

investigated the time series of the population of the major Spanish municipalities from 1900 to 1999, the degrees of adjustment of the developed logarithmic models are optimal, with levels of explanation (R^2) greater than 0.98, being *Pareto exponent* always statistically significant. Regarding to the Zipf's "law" this work concludes that the parameter α is, in all tested models, statistically different from one, which asserts that, in the Spanish case, there is no evidence in favor of that "law".

In the last years significant part of the literature, accepting the Pareto's principle, has been aimed to make an analysis on the *shape of that distribution*. Dobkins & Ioannides (2000) found that the coefficient α has decreased over the twentieth century for the cities of the USA. Similar results to those obtained by Lanaspá et al. (2004), which found regular drops of *Pareto's coefficient* from 1900 to 1970, as well as increases in the aforementioned period from that date, was interpreted as a demonstration of the changes in the Spanish urban structure⁶. Following Suarez-Villa (1988), these authors interpreted the evolution of the *Pareto's coefficient* as a *metropolitanization index*, confirming in the Spanish case, the hypothesis of Parr (1985) about the evolution of the exponent in the U form in the developed countries over time.

In the analysis of the form of size/range distribution of cities, some authors have proposed transformations in regard to the classical model of Pareto. The specialized literature (Lanaspá et al. 2004) pointed out that while Pareto's distributions fit reasonably well to the size distribution of cities, the possibility that the relation between the rank and size might be of a non-linear nature (Rosen and Resnick, 1980, Fan and Casetti, 1994) can be established in an complementary form.

Particularly it has been widely distributed in a quadratic transformation:

$$\ln r(P) = \ln A - \alpha \ln P + \beta \ln P^2$$

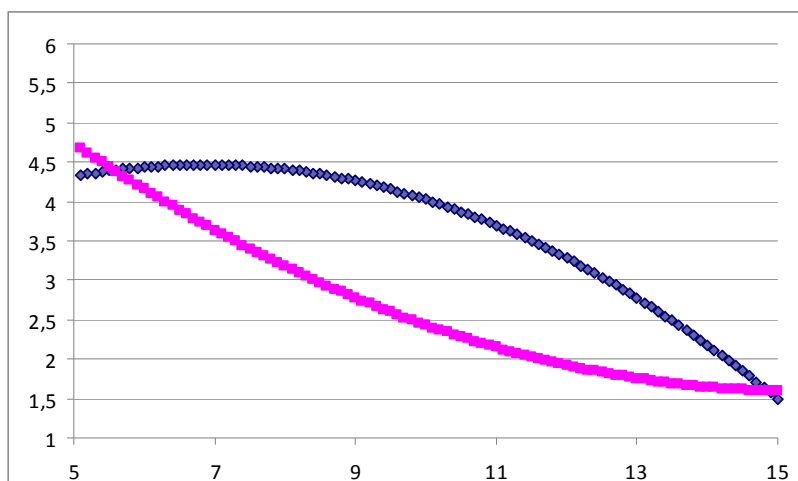
Where β helps to understand the observed curvature of the "tails" of the range/size distribution. The theoretical approach suggests, following Pareto, that the coefficient α should be positive, adjusting β , the degree of metropolitanization of structure of the urban system: increasing (ie with high "macrocephaly"), in case of having a positive value (purple in Figure n. 1) and decreasing (in structures tending to a major dispersion), if negative (blue in figure)⁷.

The empirical model of this nonlinear structure has been confirmed statistically for the major Spanish municipalities by Lanaspá et al. (2004), with R^2 surprisingly high (over 0.99), and a marked change in the curvature of the tails (β goes from having a positive sign to negative, from 1970).

6 For these authors, the Spanish urban structure undergoes a profound change in their evolution around the mid-seventies. Until that date the distribution was becoming less equal, so as to accentuate the differences between the sizes of cities, being higher in the upper part (larger cities) of the distribution. (...) In the mid-seventies until 1999, the landscape is altered and the concentration of population in major nucleus comes to its peak. The size distribution of cities as a whole becomes less unequal, in a way that the small and medium-sized agglomerations are the ones growing faster now (Lanaspá et al. 13-14).

7 For Lanaspá et al (2004), if $\beta = 0$ we would face ourselves with the Gibrat's Law (1931).

Figure n. 1: Quadratic Models



Most of the developed empirical work, however, seem to obey an implicit willingness to want to prove the validity of the Zipf's "law", or at least the less restrictive version of Pareto. The veiled warning that the log-log relationship is valid only for the "cities" often leads to an abstract definition, by the judgment of the authors of this work, of what the *city* is. For example, the work of Rosen & Resnick (1980) focuses on the top 50 cities in 44 countries. Krugman (1996) is limited to 130 U.S. major metropolitan areas. Dobkins & Loannides (2000) to all metropolitan areas, ignoring the fact of the smaller cities. Lanaspá et al. (2004) to the 100-300-700 largest municipalities in Spain. Almost all the literature has the same limitation. Berry & Horton (1970) refer to the threshold of 250,000 inhabitants as "the size of urban region (...) to establish the minimum threshold scale for economic and social viability in contemporary, metropolitanized America"⁸.

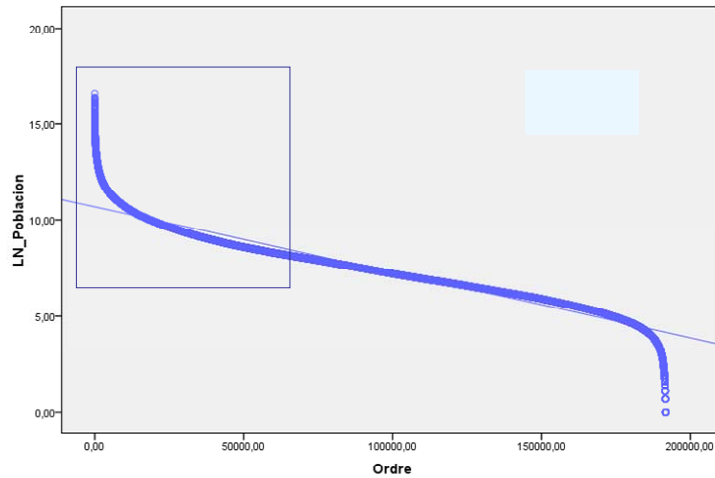
Rosen and Resnick (1980), in their work on the literature developed between 1950 and 1970, stress the importance of the definition of that "lower threshold size for cities", an aspect which has also been pointed out, more recently, by Dobkins & Ioannides (2000) and by Black & Henderson (2003). However, since 1980, it has not deepened significantly about the definition of "city", unlike the economic forces leading to the Pareto's distribution. The limitation of population thresholds, or administrative concepts of metropolitan area, are not of a concern in this investigation.

This work focuses on the discussion of the validity of the Zipf's "law" as well as the Pareto's distribution, when analyzing the entire urban system, not only its upper "tail". In this case very different forms of the relationship range/size emerge, in which the log-log relationship is only one part: limited to the "upper tail" of the total distribution. The Pareto's distribution, in this case, would limit to explain only the tip of the iceberg, requiring a more complete theory to explain whether there are regularities in the ratio range/size of cities, and its inherent causes.

This work proposes, following Eeckhout (2004), that the apparent compliance of the Pareto's distribution in large urban systems (eg > 1,000,000 inhabitants) actually reflects the biased view of the "upper tail" of a full territorial system, which apparently follows a log-normal distribution.

⁸ Berry & Horton, *op cit*, page 64.

Figure n. 2: A singularity of the "upper tail"?



As Eeckhout (2004) indicated: "At the very upper tail of the distribution, there is no dramatic difference between the density function of the log-normal and the Pareto. Now both the truncated log-normal and the Pareto density are downward sloping and similar (the Pareto is slightly more convex). As a result, both the Pareto and the truncated log-normal trace the data relatively closely"⁹. Apparently both laws could occur simultaneously: the log-log distribution of Pareto in the "upper tail" as a partial singularity of the normal logarithm of size. Based on a data from the U.S. Census 2000, Eeckhout (2004) argues that the overall distribution of American cities adopted a log-normal form rather than the Paretian, contrasting this hypothesis by applying the Kolmogorov-Smirnov test (KS) for normal distributions.

This proposal, criticized by Levy (2009)¹⁰, is currently a discussed topic in the specialized literature. González-Val et al. (2008) have found evidence of the log-normal distribution in all the urban units in Italy, Spain and USA, from 1900 until today, using a specific application of the Verification Wilcoxon's test of the null hypothesis of equality of distributions. Meanwhile Malevergne, et al. (2009) confirm the statistical validity of the Pareto's distribution to the first 1,000 USA cities, but suggests the log-normal distribution would be more efficient for the smaller cities.

The presented work shows empirical evidence for the above mentioned discussion, suggesting alternative ways of development. Especially it holds that log-normal law appears more evident when the actual cities are taken in consideration, rather than mere administrative units, proving to be an efficient tool for understanding the phenomenon of size distribution of urban systems.

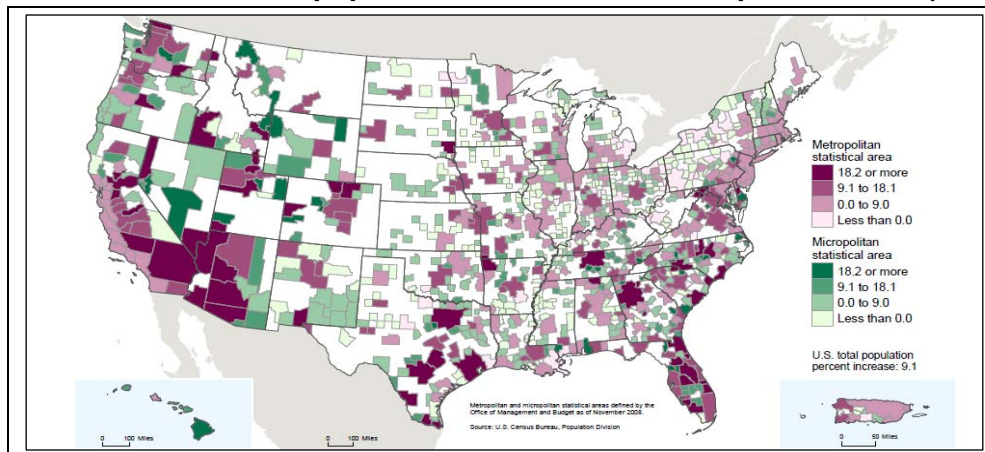
9 Eeckhout (2004), page 1432.

10 Levy (2009) argues that the top 0.6% of American cities, which comprises more than 23% of the population is separated sharply from the log-normal distribution, showing more congruency with the Pareto's log-log hypothesis. For Levy, although the bulk of the log-normal distribution follows the log-normal law, it can not be confirmed in the upper tail by applying the conventional 2-test. Levy points out that the fact that Eeckhout (2004) didn't reject the log-normal hypothesis, comes from the use of Lilienfords test (L test), which is dominated by the distribution center, rather than by their tails, "where the interesting action occurs".

3.- A first empirical analysis: the metropolitan and micropolitan areas in the USA

First of all, we will replicate the analysis of the size/range distribution of urban systems in the USA, without limit referring only to the upper tail of metropolitan areas (Metro). In order to do this, we included not only the Metro, but micropolitan areas (Micro)¹¹, as defined by the Census Bureau for 2000. This allows us to work not only with systems of more than 100,000 inhabitants (385 Metropolitan Areas), but with the 940 urban systems exceeding 10,000 inhabitants, according to data from 2009.

Figure n. 3: Evolution of the population of micro and metropolitan areas (2000-2009)



Source: U.S. Census Bureau

Figure N. 4 shows the result of applying the Pareto's distribution for urban systems of more than 500,000 inhabitants (102 Metro). It is visible that the log-log relationship seems to be confirmed ($R^2 = 0.975$), what could not be said for the Zipf's "law", as the Pareto's exponent is statistically different from 1 (-1.114).

Apparently the log-log model continues to work well ($R^2 = 0.974$), when considering the group of the urban systems of more than 10,000 inhabitants, even though the Zipf's "law" is still not confirmed ($\alpha = -0.795$). However, given figure n. 5 clearly demonstrates the concavity of the distribution, confirmed by the tested quadratic model ($R^2 = 0.997$, $\alpha = 0.913$, $\beta = -0.070$)¹². This model casts doubt on the validity of the Pareto's distribution, not so much because it is significantly more efficient than the log-log, but for the obvious change of sign suffered from the coefficient α ¹³. This change is due to a deeper reason than the co-linearity between the logarithm of the population and the square of that logarithm. The change is due to the real relationship underlying the study sample, and it is not so much the log-log relationship, but log-log. The logarithm of the population explains the residuals not explained

11 The U.S. Census differentiates the metropolitan areas from the micropolitan. The first one with a county or central city of 50,000 or more inhabitants, which adds an urban system (collectively) of more than 75,000 inhabitants. The metropolitan area is consolidated starting from the counties (or cities) that send more than a certain percentage of its residents to work at the heart of the urban system. For their part, micropolitan areas are delimited following a similar procedure, although the urban center can reach a minimum threshold of 10,000.

12 Other works, with previous data had given similar results for U.S. Metropolitan Areas. Rosen and Resnick also (1980) found, for 1980 data, the same downward concavity.

13 Same lower concave result is obtained if the studied sample was limited to Metropolitan Areas > 500,000 inhabitants.

by the square of the logarithm, and not *vice versa*, as should be expected if it were true Pareto distribution.

Figure n. 4: Metropolitan Areas USA (> 500.000 inhabitants, 2009)

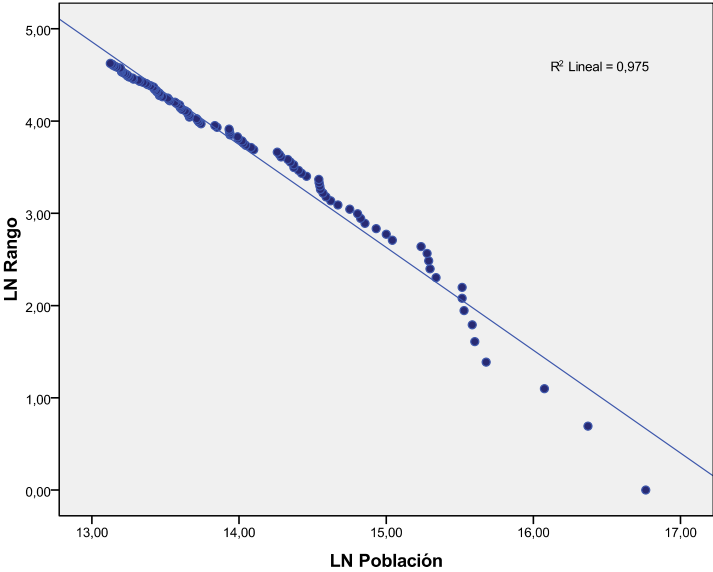
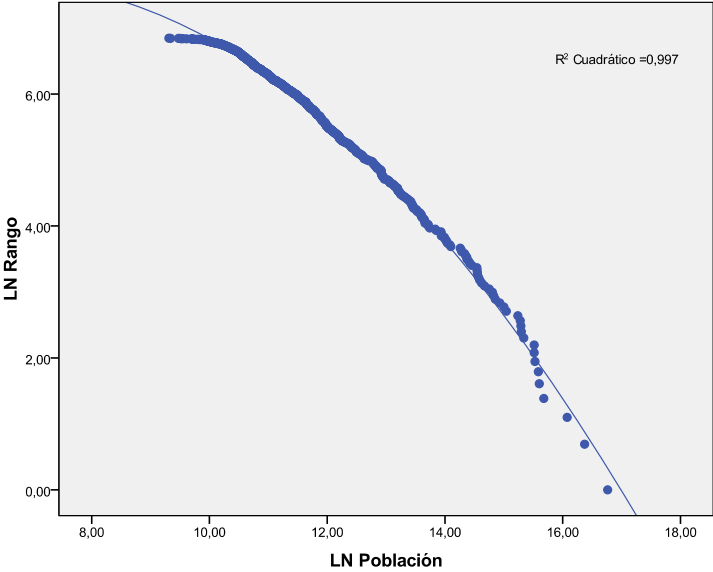


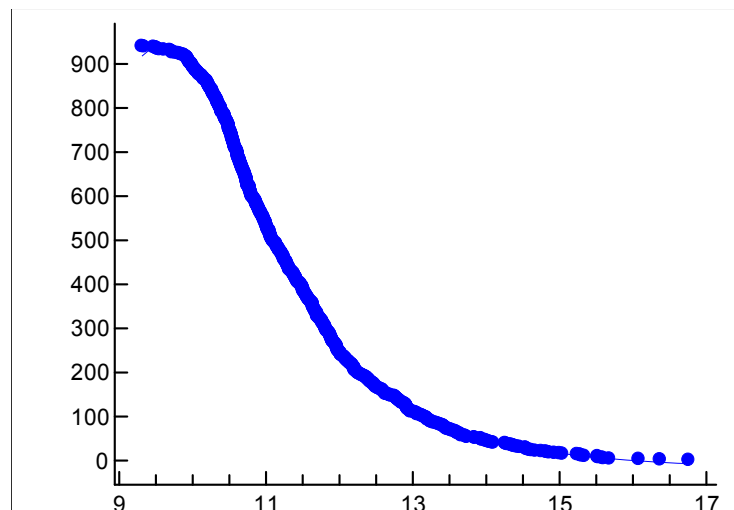
Figure n. 5: Micropolitan and Metropolitan systems USA (> 10,000 inhabitants, 2009)



The non-intuitive interpretation of the quadratic equation of the log, seems to suggest that something remains hidden in the range/size distribution of cities, beyond the interpretation given by Pareto.

This leads to analyze the range/size relationship from a non-Pareto perspective, in a lin/log form. In this case, distribution is evident in S form, which is contrasted with the good performance of the sigmoid-type models, as shown in Figure n. 6 ($R^2 = 0.999$).

Figure n. 6: Distribution of "Rational" Model¹⁴



S-shaped, sharply reminding the cumulative distribution of a normal distribution (cdf in the literature). This leads to the following questions:

- Do we face a normal distribution of the logarithm of the population?
- Is the Zipf's / Pareto's distribution then just the "upper tail" of the cumulative normal distribution?

The Kolmogorov-Smirnov test (KS) does not confirm the null hypothesis, which seems to allow confirming that the distribution of the logarithm of the Metro and Micropolitan Areas is moving away from the normal¹⁵. Moreover, the herein histogram analysis also doesn't ensure the normal structure of the logarithm of the population. Although the trend of the histogram clearly suggests a normal structure, the increased frequency of cities in the range of population (log) between 10 and 11, in relation to the range "center", between 11 and 12, leaves doubt that the distribution of the logarithm of the size obeys to a normal law.

However, the KS test, as noted by the doctrine, tends to return negative results in relatively large samples, which lead to the search for alternative resources, such as the Wilcoxon test (W), suggested by Lanaspá et al (2004), to demonstrate the equality between two distributions).

The implementation of the W test involves comparing the distribution of ranks obtained from the population size (arranged from lowest to highest) to the one that would correspond if it was normal. In order to do this: a) normal cumulative distribution (cdf) is estimated by maximum likelihood corresponding to the empirically observed population ($\mu = 11.489$, $\sigma = 1.222$), b) it is obtained by least squares the theoretical range would correspond to that cdf ($R^2 = 0.967$) and c) finally applying the W test to both distributions (sorted ranges 1, 2 ... 940 increasing, and the resulting ranges of the regression analysis obtained by the cdf). The results of applying the latter technique allows, unlike the KS test, the hypothesis of normality

14 The tested "Rational" distribution can be expressed by the following:
 $R(P) = ((A + (B * x)) / ((1 + (C * x)) + (D * (x^2))))$, where A, B, C and D constants.

15 A result of the test is a value $D = 0.097$, which corresponds to a p-value < 0.0001 , so the null hypothesis of equality with the normal distribution has to be rejected.

(p-value > alpha). We can say that the two distributions of the observed ranges are corresponding to the same pattern.

Figure n. 7: Histogram Population LN

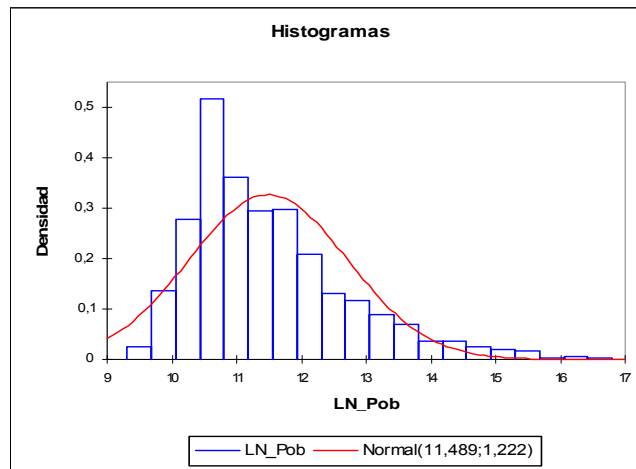


Table n.1 Comparison of distributions by the Wilcoxon method

V	210459
Esperanza	221135,000
Varianza (V)	69325822,500
p-valor (bilate)	0,200
alfa	0,05

The p-exact value could not be calculated
The p-value was calculated by approximation

However, neither the Wilcoxon test shows very reliable results, given its inherently ordinal nature.

It leads to conclusion, considering the study sample, that there is no final determination about the normal distribution of the logarithm of population of metropolitan and micropolitan areas in the USA¹⁶. However, the application of the W test suggests the possible existence of a hidden structure of the distribution of the logarithm of population tending to normal.

Table n. 2 summarizes the main results of different tests done to verify the normality of the logarithm of the population and the alternative hypothesis on the Pareto's law.

Table n. 2: Contrast Tests

Hypotesis	Log-Normal	Log-Normal	Log-Normal	Log-Log	Log-Log	Log-Normal
Test	1 (W)	2 (W)	3 (W)	4 (W)	5 (W)	6 (W)
Contrast	O_I Pred Cdf ¹	O_I Pred cdf ¹	O_I_Pred cdf ²	O_Pred pareto ³	O_Pred pareto ⁴	Prob. Acum. ⁵
Size	940	102	102	102	940	940
p-value	0,200	< 0,0001	0,526	0,276	<0,0001	0,730
Alfa	0,05	0,05	0,05	0,05	0,05	0,05
Result	Positive	Negative	Positive	Positive	Negative	Positive

¹ This shows predicted growth of range (for Metro and Micropolitan Areas 940) from the cdf ² Predicted growth of range for the 102 metropolitan areas (over 500,000) starting from the cdf. ³ Predicted descending range from the log-log model (Pareto) for the 102 metropolitan areas. ⁴ Predicted descending range from the log-log model for all the 940 areas. ⁵ Cumulative probability of the empirical distribution

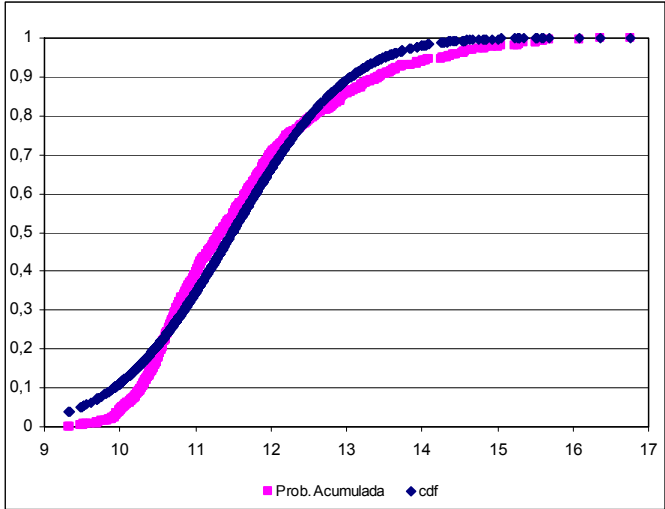
¹⁶ The lack of a conclusive demonstration may be due to the very structure of the used information: metro and micropolitan areas. First of all, the population of urban systems with less than 10,000 inhabitants, which are very numerous in the USA, is missing. It should be noted that, referring to places (> 25,000 places across the country), Eeckhout (2004) demonstrated the validity of the lognormal distribution.

It is visible in the table below that while the test 1 W permits the hypothesis of normal structure of the logarithm of the population for the whole sample (the 940 metropolitan and micropolitan areas), when applied to the upper tail, the 102 metropolitan areas (2 W), this hypothesis can not be confirmed. On contrary, the Pareto distribution is confirmed, (log-log, 4 W) for the upper tail, but not for the whole population (5 W). Our work allows us to compare the conclusions developed by Malevergne, et al. (2009) for corroboration of the advanced log-normal distribution by Eeckhout (2004) for all U.S. cities, but not for the upper tail, a segment in which, however, would be applied the log-log distribution.

The results appears consistent with the established idea that the Pareto and log-normal distributions exhibit qualitative differences in their correspondent upper tails. Log-normal density tends to zero in the upper tail, faster than any paretian density, which should help to distinguish them clearly.

Nevertheless, the assumption of normality of the logarithm of the population remains strong. W Model 6 confirms the null hypothesis regarding the identity of the normal cumulative distributions (cdf) and empirical cumulative distributions (see Figure n. 8)¹⁷.

Figure n. 8: Normal and empirical cumulative distributions



Meanwhile, the W test 3 suggests that if the growing range is adjusted by a regression model with the cdf as the dependent variable only for the upper tail, confirms the identity of both distributions. The upper tail of the empirical distribution would imply, therefore, a distribution based on normality. However, this result is not consistent with the one obtained for the whole sample.

4.- The distribution of cities in Spain

Secondly, we will replicate the study for Spanish cities and urban systems. The used data are the population (relative to 2005) of the 8,109 municipalities, as well as the one relative (up to 2009) to 1,316 urban systems defined according to the methodology described below. The

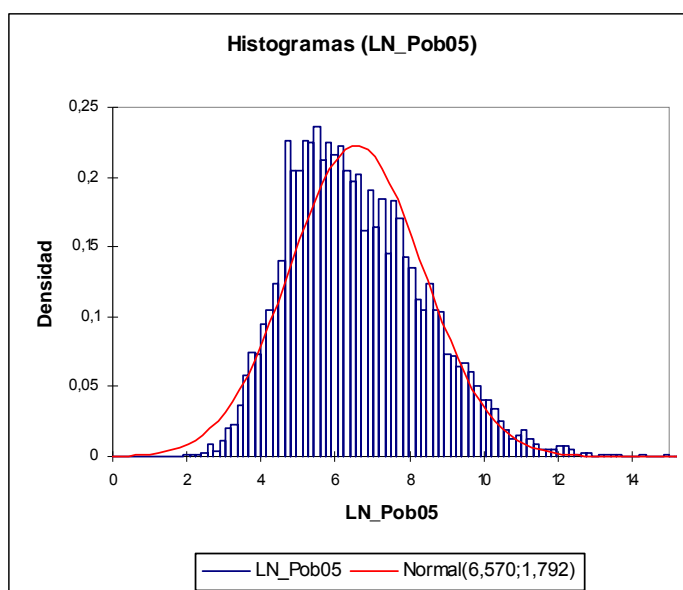
¹⁷ The regression model for both variables returns an R2 = 0.979

reason for using urban systems, complementing the municipalities, is to test whether the actual structures (urban systems) improve their performance in relation to historically inherited administrative structures (municipalities).

4.1 .- The Spanish municipalities

All the tests suitable for checking normality of the logarithm of population give us, as in the case of U.S. metropolitan areas and micropolitan, a negative result which does not confirm, prima facie, the log-normal hypothesis that partially suggests the frequency histogram of the population (Figure n. 9)¹⁸.

Figure n. 9: Histogram of the LN of the population of Spanish municipalities



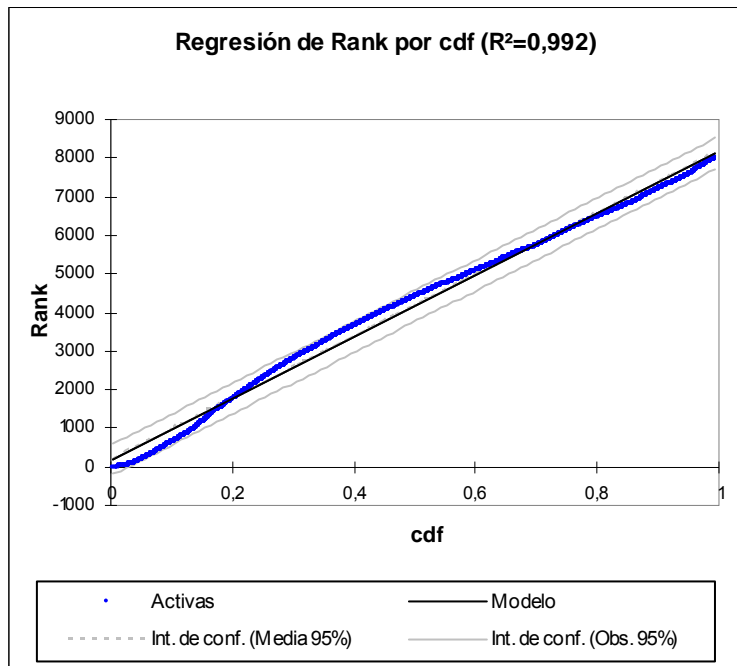
Despite the high explanatory power ($R^2 = 0.992$, fig. n. 10) of the regression model, with the growing range as the dependent variable, and cdf as an independent variable, the KS test can not verify the identity between the distribution of that range and the value predicted by the regression model resulting from the cumulative normal, as shown in the table n. 3 (test 1 KS). Identity can be hypothesized under W test of ordinal nature (W test 2).

Table n.3: Tests of contrast for the population of municipalities (2005)

Hypotesis	Log-Normal	Log-Normal	Log-Log	Log-Log	Log-Normal	Log-Normal
Test	1 KS	2 W	3 KS	4 W	5 W	6 W
Contrast	O_I pred. cdf	O_I pred. cdf	O_pred. pareto	O_pred. pareto	O_I_N / cdf	O_pred.Pareto
size	8.109	8.109	58	58	8.109	58
p-value	< 0,0001	0,355	0,919	0,264	< 0,0001	0,927
alfa	0,05	0,05	0,05	0,05	0,05	0,05
Result	Negative	Positive	Positive	Positive	Negative	Positive

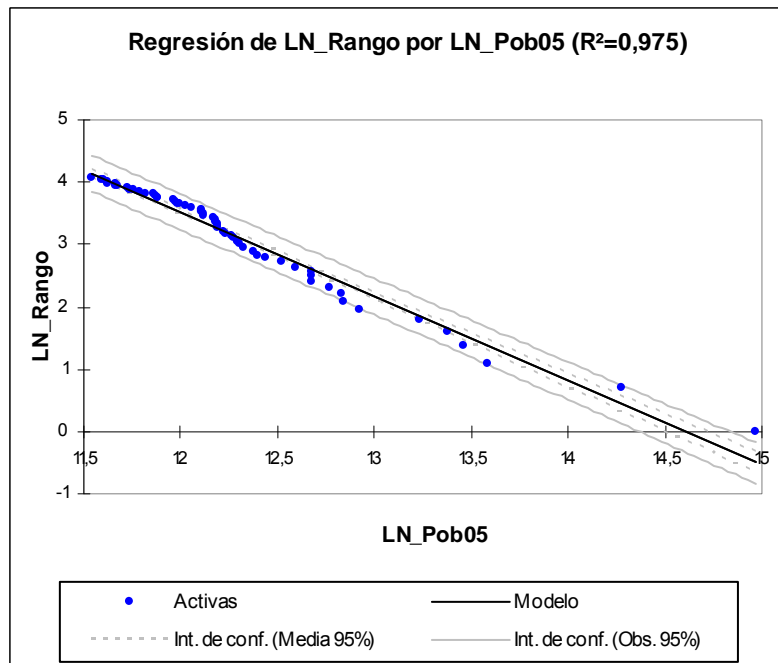
18 Again, the histogram shows a relative lack of symmetry in the distribution of the logarithm of the population. Municipalities with a logarithm of its population between 5 and 6 are significantly more abundant than those between 6 and 7, having doubts about the normal character of the distribution, despite its formal appearance.

Figure n. 10: Model based on the cdf (entire sample)



Again, for the upper tail (in this case the municipalities with more than 100,000 inhabitants), both the Kolmogorov-Smirnov (KS test 3) and the Wilcoxon test (test 4 W) confirm the adequacy of the log-log function (which achieves a benefit of adjustment to 0.975, figure n. 11). Pareto's Law is thus confirmed for the larger Spanish municipalities. On contrary, Zipf's "law", is shown to be far from the hypothesis -1 (- 1.355)¹⁹ due to the regression coefficient.

Figure n. 11. The log-log model of the Spanish municipalities (100,000 inhabitants)



¹⁹ The differences observed between the coefficient α in the models of the urban areas of the USA and the Spanish municipalities, particularly calls for attention. The steeper slope of the Spanish model suggests a greater macrocephaly in relation to the distribution of the population in the USA.

As the case of micro and metropolitan areas in the USA have shown, the log-log model fails when trying to explain the distribution of the whole sample. The R^2 of 0.900, is considerably removing from the adjustment shown by the log-normal model, therefore proving that the Pareto's Law quickly stops to be efficient for the explanation of the distribution of the whole urban structure, opposed to what happens with the cdf result for the hypothesis of normality.

Finally, it is important to note that if a regression model of descending order for the 58 municipalities in the upper tail with the cdf is replicated, a relatively high goodness of fit ($R^2 = 0.952$) is obtained, although lower than the one achieved by the log- log distribution. The fact, however, that both the Kolmogorov-Smirnov test and the Wilcoxon (6 W) are positive, suggests a non-deficient behavior of the cdf, if a specific model for the upper tail is set.

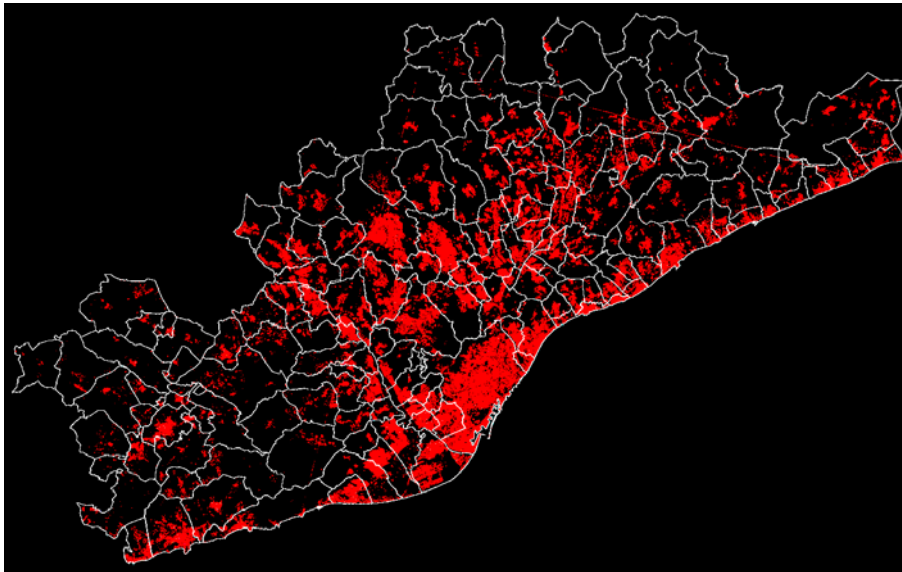
Exercise made on Spanish municipalities, therefore, suggests the validity of the normal hypothesis of the logarithm of the population. However, the upper tail still shows a structure that tends to escape from the log-normal distribution, suggesting the existence of singularities in the largest municipalities. Singularities that can only be explained by the Pareto's log-log hypothesis.

4.2. Urban systems.

In order to outline the previous analysis with *actual* cities, not mere administrative units (municipalities), a task to establish the limits of urban systems beyond the limits set by the administration was made.

For example, saying that "Barcelona" corresponds to the municipality of Barcelona (100 km², 1,593,080 inhabitants in 2005) is almost as absurd as saying that "London" is the City of London, the only British local entity that carries the name of UK capital. Figure 12 shows how the urban reality of Barcelona overflows by far its municipal boundary (in white).

Figure n. 12: "Artificial" land in the Metropolitan Region of Barcelona



Source: CPSV

The difficulty, however, is to obtain a reliable methodology for the delimitation of real cities. The objective of this work is not to deepen the discussion of alternative forms of actual realization of those cities, but to point out the proposal of non identification of these cities with their metropolitan areas. Metropolitan areas are characterized by incorporation of different urban realities, physically continuous or not, characterized by maintaining strong ties of interaction. Nevertheless, the metropolis exists beyond the "city". In other words, they are "cities of cities." For this reason we think that their use is improper, because they are neither municipalities nor smaller administrative entities²⁰.

For this work we chose a delimitation methodology used by Roca et al. (2009) in the work that concerns the delimitation of proto-systems and urban consolidated systems, based on the application of the technique of the *interaction value* (Roca & Moix, 2005). This methodology can be summarized by the following elements:

- From the matrix of work/residence flows on the municipal base (8.019 x 8.109), the i/j matrix of "values of interaction" is calculated through the following equation:

$$IV_{ij} = \frac{F_{ij}^2}{REP_i \cdot WP_j} + \frac{F_{ji}^2}{REP_j \cdot WP_i}$$

Where IV_{ij} is the interaction value between the municipalities i and j , F_{ij} and F_{ji} are the existing the flows from i to j and from j to i , respectively, REP_i and REP_j is the resident employed population of both municipalities, and WP_i and WP_j are the locally based workplaces within municipalities i and j .

- Later the municipalities in proto-systems are joined as a function of their maximum interaction value, so that proto-systems can finish only if all the municipalities have their maximum interaction value with other municipalities in the same proto-system and if the group is physically continuous.
- Finally proto-systems are consolidated in urban systems when self-containment²¹ is equal or exceeds 50%, as the authors understand that can only be called "cities" those urban systems capable of retaining at least 50% of the employed resident population²².

This allows the identification of 1,531 proto-systems, of which 218 do not satisfy the self containment condition (fixed at 50%), leading to a final delimitation of 1,316 consolidated proto-systems, which for the purposes of this study will be considered as a *real urban systems*. Figure n. 13 presents the results of delimitation.

Although the histogram (figure n. 14) shows a distribution that is strongly getting close to normal, the standard test of parametric nature, again, does not confirm that the distribution of the logarithm of the population responds fully to a normal structure, which demands seeking alternative validation mechanisms.

20 Such as "places" used in the recent literature devoted to the discussion of the laws of Pareto and Gibrat.

21 Self-contained means the percentage of the resident employed population working in the same municipality (or proto-system).

22 That 50% is the only condition imposed on urban systems. Therefore no administrative status of the minimum threshold of population or WP is imposed.

Figure n. 13: Urban systems delimited by the interaction value

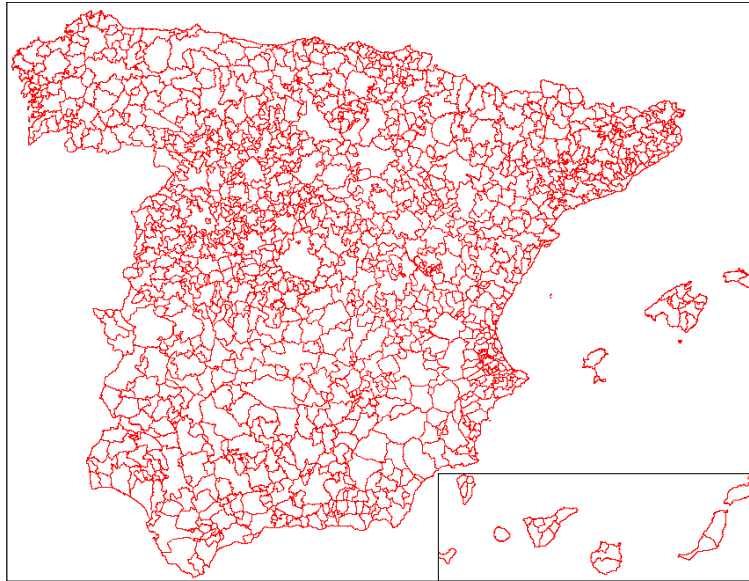
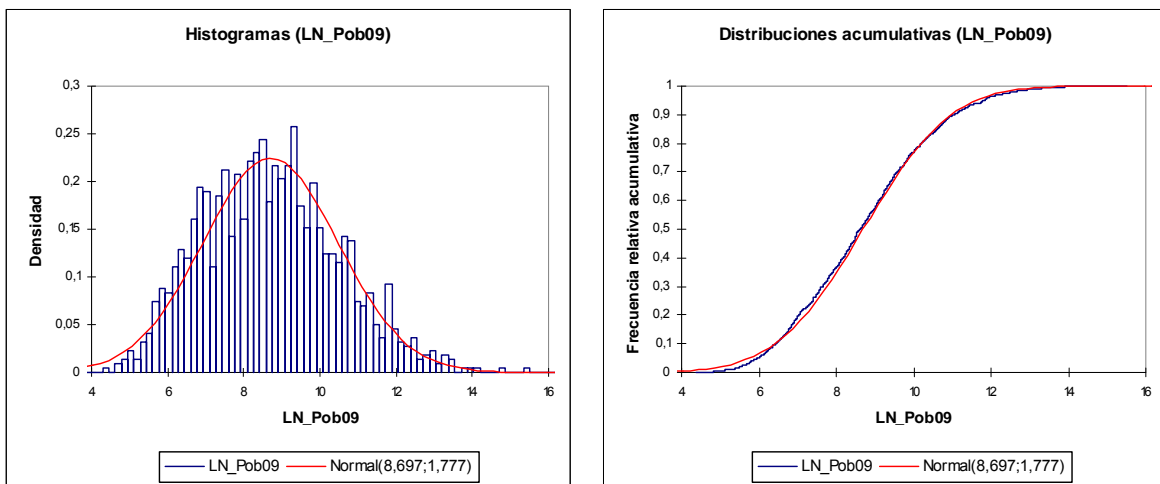


Figure n. 14: Histogram



The adjustment of a regression (figure 15) model with the ascending order (O_I) as the depending variable, and the cdf of the logarithm of the population, as an independent variable, permits to achieve a spectacular R^2 : 0.9984, which again permits the hypothesis that population shows a log-normal structure.

However, as it is clear from Table 4, the Wilcoxon's test (1 W) does not, unlike what happened in the case of Spanish municipalities and micro areas and metropolitan USA, ensure the identity of both distributions (the increasing order and the resulting predicted order regression model with cumulative normal density calculated from the logarithm of the population as an explanatory variable). This negative result could be interpreted as a proof of the non-normality of the logarithm of the population. However, from the following we can not conclude such a thing, but on the contrary, there are serious doubts about the validity of the Wilcoxon's test for corroboration of the identity of distributions, given its inherent ordinal nature.

Figure n. 15: Regression Model with the cdf as an independent variable

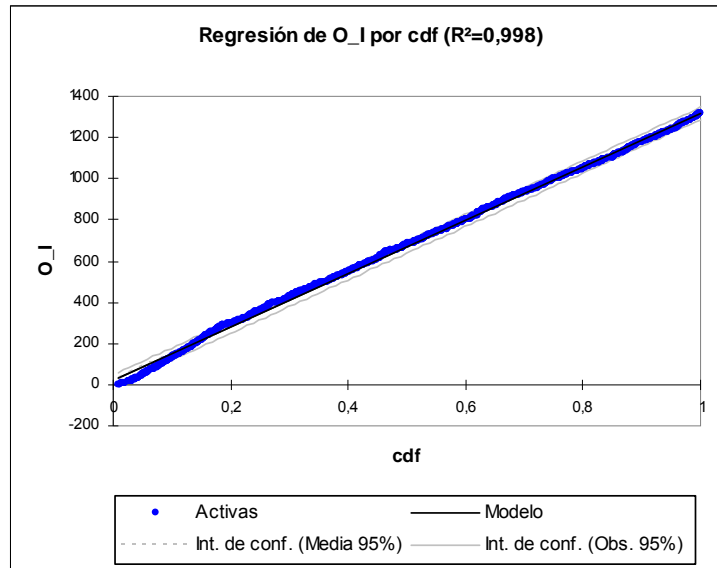


Table n.4: Contrast tests for the population of Spanish urban systems (2009)

Hypothesis	Log-Normal	Log-Normal	Log-Normal	Log-Normal	Log-Log	Log-Log
Test	1 W	2 KS	3 KS	4 MW	5 KS	6 KS
Contrast	O_I pred. cdf	Normal Distribution	O_I pred. cdf	O_I pred. cdf	Orden pred. Pareto	O_I_N pred. cdf
size	1.316	1.316	1.316	1.316	88	88
p-value	0,005	0,130	0,478	0,999	0,987	0,986
alfa	0,05	0,05	0,05	0,05	0,05	0,05
result	Negative	Positive	Positive	Positive	Positive	Positive

The application of Kolmogorov-Smirnov non-parametric test, to contrast the adjustment to a normal distribution of the logarithm of the population (2 KS), obtains a positive result (p-value = 0.1304), similar to KS test for verifying the identity of the distribution for the ascending (O_I) and the resulting regression model with dependent variable cdf (3 KS). As the p-value calculated (0.478) is greater than the significance of level alpha = 0.05, the null hypothesis of identity between the two distributions can be accepted. The contradiction of results between W and KS tests questions the methodology based on the comparison merely ordinal. In the same vein, the Mann-Whitney / sided test (see model 4 MW) gives positive results for the comparison between the crescent and the resultant of the regression model with cdf.

The tests, therefore, confirm the log-normal hypothesis of population of urban systems, with a strength not previously achieved for municipalities or metro and micropolitan areas. *The population structure of real cities seems to adjust to a log-normal distribution.*

The contrast of the Pareto's law is concentrated in the 88 urban systems of more than 100,000 inhabitants. The logarithm of the population get a $R^2 = 0.991$, with log-rank (descending order), confirming once again the excellent performance of log-log model in the upper tail. Meanwhile, the non-parametric KS (5 KS), as well as the Wilcoxon's, confirms the identity of the distribution of log-rank and the predicted value of the resulting regression model with the logarithm of the population as an independent variable.

However, the model of the upper tail, with the cdf as independent variable and the standard reverse order (O_I_N), surprisingly, reaches an even higher level of explanation ($R^2 =$

0.993), as well as the confirmation by tests KS (6 KS) and W, of the correspondence between the two distributions.

5.- Conclusions

The completion of the previous studies lead us, for micro and metropolitan areas of the USA, to the following conclusions:

1. The structure of the population of all U.S. urban areas over 10,000 population seems to correspond to a log-normal distribution, as proposed by Eeckhout (2004) for the group of USA cities.
2. This conclusion does not seem to be applicable to the upper tail, a segment in which, a log-log distribution should be applied, as suggested by Malevergne, et al. (2009).

Identical results seem to be observed in the study of the structure of the population of Spanish municipalities. For the whole sample the log-normal model is confirmed. However, the upper tail continues to show signs of weaknesses, showing a clear supremacy of the log-log model.

Nevertheless, the conclusions above are not definitive, given the nature of the data studied in the two previous series. None of them has considered the *real cities*. In the case of the USA, since the metro and micropolitan areas could obey to real cities, urban systems with less than 10,000 inhabitants (abundant outside the metropolitan surroundings) could not be considered. In the case of Spain, since the municipalities respond to administrative entities, they do not represent a true reflection of the reality of the country's urban structure.

Alternatively, the methodology developed by Roca et al. (2009), concerning the delimitation of urban systems through the interaction value system, was used. And the result seems to confirm the hypothesis that when we are confronted with real cities, the distribution of population responds precisely to a log-normal structure. The improvement of virtually all used indicators (see Table n. 5) suggests the need for improved empirical work using not only the entire city environment, but also the consideration of real urban systems.

Table n. 5

Pattern	R2 OIN-cdf ¹	KS Norm ²	KS OIN-cdf ³	KS OI-pred_cdf ⁴	MW OIN-cdf ⁵	MW OI-pred_cdf ⁶
counties	0,992	< 0,0001	< 0,0001	< 0,0001	0,00244041	0,99116671
Urban systems	0,998	0,13038	0,47755912	0,47774589	0,00462283	0,99872660

¹ Model of regression between the normalized reverse order (ISO) and the normal cumulative density (cdf) of the logarithm of the population. ² Contrast of Kolmogorov-Smirnov normality. ³ Comparison of the identity of the distributions of OIN and the cdf by the Kolmogorov-Smirnov test (KS). ⁴ Comparison by the KS test of the identity of the growing range distributions (OI) and the prediction of the same, from a regression model with the cdf as independent variable. ⁵ Verification, by the Mann-Whitney test (MW) of identity between the ISO and the cdf. ⁶ Verification by the MW test of the identity of the growing range and and the same prediction from a regression model with the cdf as independent variable.

Bibliography

- Alperovich, G. (1993): "An Explanatory Model of the City-Size Distribution: Evidence from Cross-country Data", *Urban Studies*, 30, 1591–1601.
- Berry, B.J.L. (1961): "City Size Distributions and Economic Development", *Economic Development and Cultural Change*, 9, 593–587.
- Black, D. & Henderson, J.V. (2003): "Urban Evolution in the USA", *Journal of Economic Geography*, 3, 343-372.
- Berry, B.J.L. & Horton, F.E. (1970): *Geographic perspectives on urban systems*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey.
- Carroll, G. (1982): "National city size distributions: what do we know after 67 years of research?", *Progress in Human Geography*, 6, 1-43.
- Cheshire, P.(1999): "Trends in Sizes and Structure of Urban Areas", en *Handbook of Regional and Urban Economics* (P. Cheshire, and E. S. Mills, eds.), Elsevier Science, B. V., Amsterdam.
- Dobkins, L.H. & Ioannides, Y.M. (2000): "Dynamic Evolution of the US City Size Distribution", in *The Economics of Cities* (J.-F. Thisse and J.-M. Huriot, eds.), Cambridge University Press, Cambridge (2000).
- Eeckhout, J. (2004): "Gibrat's law for (all) cities", *American Economic Review* 94, 1429-1451.
- Eeckhout, J. (2009): "Gibrat's law for (all) cities: Reply", *American Economic Review*, 99:4, 1676–1683.
- Krugman, P. (1999): "El tamaño de las ciudades" en *The Spatial Economy* (Fujita, M., Krugman, P. & Venables, A.J. eds), Massachusetts Institute of Technology.
- Gabaix, X. (1999): "Zipf's Law for Cities: An Explanation", *Quarterly Journal of Economics*, CXIV, 739–767.
- Gibrat, R. (1931): *Les inégalités économiques*. Paris, Librairie du Recueil Sirey.
- González-Val, R., Lanaspá, L., Sanz, F. (2008): *Nueva Evidencia sobre la Ley de Gibrat en Ciudades*. Universidad de Zaragoza.
- Krugman, P.R. (1996): *The Self-Organizing Economy*, Blackwell Publishers, Oxford.
- Lasuén, J. R., Lorca, A. y Oria, J. (1967): "City-Size Distributions and Economic Growth", *Ekistics*, vol. 24, págs. 221-226.
- Lanaspá, L., Perdiguero, A.M., Sanz, F. (2004): "La distribución Del tamaño de las ciudades en España", *Revista de Economía Aplicada*, 34, vol. XII, 5-16.
- Levy, M. (2009): "Gibrat's law for (All) cities, A Comment", *American Economic Review*.
- Malevergne, Y., Pisarenko, V. & Sornette, D. (2009): "Gibrat's Law for Cities: Uniformly Most Powerful Unbiased Test of the Pareto Against the Lognormal". *Swiss Finance Institute Research Paper* N. 09-40, September 2009. <http://ssrn.com/abstract=1479481>.
- Pareto, V. (1896): *Cours d'Economie Politique*. Geneva, Droz.
- Parr, J. (1985): "A Note on the Size Distribution of Cities over Time", *Journal of Urban Economics*, 18, 199-212.
- Roca, J., Marmolejo, C. & Moix, M. (2009): "Urban Structure and Polycentrism: Towards a Redefinition of the Sub-centre Concept", *Urban Studies*, 46, 2841–2868.
- Roca, J. & Moix, M. (2005): "The interaction value: its scope and limits as an instrument for delimiting urban systems", *Regional Studies* 39, 357-373
- Rosen, K.T. & Resnick, M. (1980), "The size distribution of cities: an examination of the Pareto law and primacy", *Journal of Urban Economics*, 8:165-186.
- Suárez-Villa, L. (1988): "Metropolitan Evolution, Sectoral Economic Change, and the City Size Distribution", *Urban Studies*, 25, 1-20.
- Zipf, G. (1949): *Human Behavior and the Principle of Least Effort*, New York, Addison-Wesley