The Spanish spatial city size distribution

González-Val, Rafael

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

16 June 2020

The Spanish spatial city size distribution

Rafael González-Val

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

***Abstract*:** This paper analyses the Spanish city size distribution from a new perspective, focusing on the role played by distance. Using un-truncated data from all cities in 1900 and 2011, we study the spatial distribution of cities and how the city size distribution varies with distance. First, K-densities are estimated to identify different spatial patterns depending on city size, with significant patterns of dispersion found for medium-sized and large cities. Second, using a distance-based approach that considers all possible combinations of cities within a 200-km radius, we analyse the influence of distance on the city size distribution parameters, considering both the Pareto and lognormal distributions. The results validate the Pareto distribution in most of the cases regardless of city size, and the lognormal distribution at short distances.

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

***JEL*:** C12, C14, 018, R11, R12.

1. Introduction

A classical topic in the regional science literature with strong socio-economic implications is city size distribution. The spatial distribution of population in cities is related to the extension of local labour markets and the intensity of internal migratory flows, defining the resulting economic landscape, and has direct consequences on the spatial distribution of income and on imbalances between territories. Since the beginning of the twentieth century, many studies have analysed the empirical regularity known as Zipf's law (Zipf, 1949) for many different countries (see the surveys by Cheshire, 1999; Nitsch, 2005; Soo, 2005; Cottineau, 2017). Zipf's law establishes that the second-largest city in a country is approximately one-half the size of the largest one, the third-largest city is one-third the size of the largest one, etc. (i.e., there is a linear relationship between rank and size). In that case, the population distribution across cities can be fitted by a Pareto distribution with an exponent equal to one.

Zipf's law provides a simple (and, apparently, accurate) representation of empirical city size distributions, which has facilitated its wide diffusion among urban economists, statistical physicists, and urban geographers. To give theoretical support to this empirical regularity, some authors have proposed theoretical models with different economic foundations to explain the law: productivity or technology shocks (Duranton, 2007; Rossi-Hansberg and Wright, 2007) or local random amenity shocks (Gabaix, 1999). These models link Zipf's law to an equilibrium situation: thus, Zipf's law has become the benchmark for steady-state city size distributions.

In the last decades, the empirical city size distribution literature has focused on some technical issues: the definition of 'city', sample size, estimation procedure, and the most accurate benchmark statistical distribution. Therefore, sophisticated city definitions have been developed (such as the United States (US) economic areas defined by using the city clustering algorithm by Rozenfeld et al., 2011), large sample sizes without the imposing of any population size restriction have been considered (Eeckhout, 2004), several methods are now available to estimate the distributional parameters (e.g., OLS (Gabaix and Ibragimov, 2011) and maximum likelihood (Gabaix and Ioannides, 2004; Goldstein et al., 2004; Clauset et al., 2009), as well as specific methods for estimating the population threshold of the distribution (Fazio and Modica, 2015)), and convoluted distributions have been fitted to population data rather than the classical Pareto and lognormal distributions (e.g., the double Pareto lognormal distribution

(Reed, 2002; Giesen et al., 2010; Giesen and Suedekum, 2014) and the distribution function of Ioannides and Skouras (2013), which switches between a lognormal and a power distribution).

In summary, the current mainstream of this literature is that when all cities are considered with no size restriction, the behaviour of the largest cities at the upper-tail of the distribution can be linear but different from that of the entire distribution. That is why most of the new distributions mix linear and nonlinear functions, separating the body of the distribution from the upper-tail. In the words of Ioannides and Skouras (2013), "*most cities* obey a lognormal; but the upper-tail and therefore *most of the population* obeys a Pareto law." Figure 1 illustrates this point using data from all Spanish cities (i.e., municipalities) in 1900 and 2011.[1] The data, plotted as a complementary cumulative distribution function (CCDF), are fitted by a Pareto distribution (the solid line) and a lognormal distribution (the blue dotted line) estimated by maximum likelihood. Nonlinear and concave behaviour is observed, especially in 2011, so the lognormal distribution provides a good fit for most of the distribution. However, important deviations between empirical data and the fitted lognormal can be found for the medium-sized and largest cities. Actually, the largest cities' behaviour seems almost linear; thus, a power law can be fitted to the upper-tail distribution. The population threshold defining the upper-tail is set using Clauset et al.'s (2009) methodology. The graphs also show important changes in Spanish city size distribution: The 2011 distribution is more non-linear, and the number of cities greater than the population threshold is lower than it was in 1900.

However, as González-Val (2019a) highlighted, a fundamental issue often omitted in this debate is the spatial perspective.[2] He argued that one specific feature of a city size distribution is the spatial dependencies between the elements of the distribution, as cities are connected through migratory flows: There is a relationship between the population of one city and the populations of nearby cities. In spatial equilibrium models, this is captured by the free migration assumption. However, large cities at the Pareto upper-tail are usually far away from one another. Table 1 shows the bilateral physical distances between the 10 largest cities in Spain in 1900 (Panel A) and 2011 (Panel B). Madrid is the largest city in both periods, and S_{MAD}/S (the quotient of

[1] Information about city definitions and sources is given in Section 2.
[2] Two recent papers that study spatial dependencies between cities, although looking at city growth, not at the entire distribution of cities, are Beltrán et al. (2017) and Cuberes et al. (2019).

Madrid's population divided by city *i*'s population) reports how close these top-10 cities are to Zipf's law. Although there is a difference of over 100 years between both panels, there are small changes in the ranking, as large city sizes are quite persistent. Nevertheless, the important point is that the average physical distance between these cities is quite long: 419.7 and 417.4 km in 1900 and 2011, respectively. Therefore, on average, there is a persistent great distance between the largest cities, so migrations between these places are difficult because most people do not move so far.

Internal mobility within European countries is low (Cheshire and Magrini (2006) estimated that the population mobility in the US is 15 times higher than that in Europe),[3] but migration movements in Spain are even lower: Bentolila (1997) showed that migration between Spanish regions has fallen significantly since the 1970s despite of large and widening regional unemployment rate differentials. According to the 2011 population census, only 15% of residents in Spanish municipalities were born in a different NUTS II region, while 44% lived in the same city where they had been born and 27% came from other cities within the same region. Some large cities in particular receive a high number of people from other regions (for instance, 28% in Madrid and 20% and 19% in Barcelona and Valencia, respectively), though the 15% average value holds for cities with more than 25,000 inhabitants. This pattern of low internal migrations was similar to that in the previous censuses (Romero Valiente, 2003); from 1981 to 2001, the percentage of people living in the city of their birth was always higher than 50%, and the share of residents coming from other places within the same NUTS II region oscillated between 25% and 30%.

As a consequence, it is not clear what it means to say that the Pareto distribution (and Zipf's law) holds for the largest cities (as Figure 1 shows) because they are almost independent elements in terms of long-run migratory flows. This situation implies that the largest cities within the same country are, indeed, the centres of different urban systems (González-Val, 2019a). Therefore, from this point of view, the city size distribution (and Zipf's law) should be addressed from a local scope, rather than considering all cities in a pool independently of the geographical distances between them, as most papers traditionally do in this literature. A related issue is the

[3]Mobility in the US is much higher than it is in Europe, but Dao et al. (2017) showed that interstate mobility in the US has been weakening since the early 1990s.

geographical spatial limit of urban systems, which is an open research question with scarce evidence (Pumain, 2006; Hsu et al., 2014; González-Val, 2019b).

The aim of this paper is twofold. First, as in González-Val (2019b), we study the spatial distribution of Spanish cities in 1900 and 2011 by considering space as continuous using the methods of Duranton and Overman (2005), obtaining evidence supporting a dispersion pattern of the medium-sized and large cities in both years, which would indicate the existence of several urban subsystems. This pattern would be consistent with the central place theory. Second, we follow the distance-based approach proposed by González-Val (2019a) to analyse how the Spanish city size distribution changes over space, considering all the possible combinations of cities within a 200-km radius and obtaining support for the Pareto distribution for nearby cities regardless of their sizes, while the lognormal distribution is found to be valid only for short distances. Thus, although we question the traditional approach in the city size distribution literature that considers all cities within a country or a large area, our results support the validity of the Pareto distribution (the traditional benchmark distribution) for local geographical samples. Moreover, placebo regressions confirm the significant effect of geography on the Pareto exponent when we compare the results obtained by using geographical and random samples of cities.

As far as we know, this analysis focused on distance has been applied only to the US city size distribution. In this paper, we analyse the Spanish city size distribution; this method can provide new insights that help us understand the local city size distribution in an urban context of old cities and low population mobility. As a European country, Spanish has an urban system very different from that of the US for several reasons. For instance, European cities date from ancient times, while US cities are relatively young (the first census by the US Census Bureau dates from 1790). Therefore, in the US, there has been a great entry of new cities (Dobkins and Ioannides, 2000) in the last two centuries, while the number of cities in Spain has remained quite stable over time. Moreover, as mentioned above, population mobility in the US is higher than that in Europe (Cheshire and Magrini, 2006).

Our paper is related to the study of the Spanish city size distribution by Le Gallo and Chasco (2008). They considered Spanish urban areas from 1900 to 2001 to estimate Zipf's law by using a spatial SUR model. However, our approach here is not a spatial econometrics exercise; we introduce space into our methodology through the selection

of geographical samples of cities based on distances. Other papers have also tried to take into account space in the study of city size distribution; Giesen and Südekum (2011) showed that Zipf's law holds for German cities as well as for German regions, while Lalanne (2014) studied the hierarchical structure of the Canadian urban system, splitting the Canadian territory into two parts (east and west). However, our approach does not rely on political boundaries of regions, as our geographical samples are defined considering continuous space.

This paper is organized as follows. Section 2 presents the data that we use. Section 3 contains the analysis of the spatial distribution of Spanish cities using the methods of Duranton and Overman (2005). In Section 4, we use a distance-based approach to study the influence of distance on the parameters of the city size distribution. Finally, Section 5 concludes.

2. Data

We use un-truncated size data from all Spanish cities. Primary data come from the 1900 and 2011 decennial censuses carried out by the Spanish National Statistics Institute (*Instituto Nacional de Estadística*, www.ine.es).[4] The 1900 census is the first census of the twentieth century, while the 2011 census is the most recent population census. Our data cover more than one century, allowing us to extract conclusions about the long-term evolution of population distribution. Nevertheless, such a long time period involves a potential issue, as city boundaries change over time. To deal with this problem, we use Goerlich et al.'s (2006, 2015) consistent definition of Spanish cities: Based on geographical information system (GIS) tools, they used the 2001 census as the base reference for homogenizing the population series, estimating municipal populations backward (until 1900) and forward (until 2011, although in the last census boundaries change in only eight cities).[5] Therefore, our spatial units are consistent, allowing for comparisons over time.

The geographical unit of reference is the municipality. Municipalities are the smallest spatial units (local governments); thus, they are the administratively defined 'legal' cities. They are the lowest spatial subdivision in Spain; in terms of the European Union's standard classification of European regions (*Nomenclature des unités*

[4] The censuses were conducted in years ending in zero, between 1900 and 1970, and one, from 1981 onward.

[5] This data set is available for download at: https://www.fbbva.es/bd/cambios-la-estructura-localizacion-la-poblacion-series-homogeneas-1900-2011/.

territoriales statistiques), municipalities are the LAU 2/NUTS 5 regions, comprising the country's total land area, and, therefore, the entire population. These spatial units have been used previously in studies of the Spanish city size distribution: González-Val et al. (2014, 2017) and Lanaspa et al. (2003, 2004).

Aggregate spatial units are not considered because wider areas imply much lower sample sizes. Metro areas indeed represent urban agglomerations, covering huge areas that are meant to capture labour markets. However, according to the Urban Audit 2018 project by the European Union, there are only 132 Larger Urban Zones (LUZ) in Spain. The number of NUTS 2 and 3 regions is even smaller: 19 and 59 units, respectively. There is also a subregional division between municipalities and the NUTS 3 regions, grouping nearby municipalities, the so-called '*comarcas*'. However, only a few regions have official definitions for *comarcas* (Aragón, Asturias, Catalonia, and Galicia) and, therefore, population data at this geographical level does not cover the whole country. Finally, although recent research has provided new definitions of urban areas based on a machine learning algorithm that groups buildings within portions of space of sufficient density (Arribas-Bel et al., 2020), this data set is available for only one year (2017).

Panel A in Table 2 shows the sample's descriptive statistics. The total number of cities is 7,959; as our method is based on physical distances, municipalities in islands (Balearic and Canary islands) and Ceuta and Melilla, located on the African coast, are excluded from the analysis. While there are small changes in the number of municipalities between censuses, we consider the same cities in both periods. The minimum values indicate that all municipalities—even the smallest units—are included, without size restrictions. Although their urban character is debatable, Eeckhout (2004) suggested considering the whole distribution because if any truncation point is imposed, the estimates of the Pareto exponent may be biased. The table also shows, in Panel B, the parameter estimates for both the Pareto and lognormal distributions using the population from all cities; in 2011 the estimated Pareto exponent is far from Zipf's law, which is a common result when un-truncated data is used, while in 1900 the parameter estimate is closer to one. This indicates that city sizes were more homogeneous in 1900 than they were in 2011 (the standard deviation in 2011 is over four times that of 1900, see Panel A). Graphically, this means that city size distribution better accommodates a power law linear fit in 1900 than in 2011, as Figure 1 shows.

3. The spatial distribution of Spanish cities

Table 1 provides some anecdotal evidence on the great bilateral distances between the 10 largest cities in Spain. To make this point clearer, the maps in Figure 2 represent the geographical distribution of all municipalities for different population thresholds in 1900 and 2011. It can be seen that, apart from a couple of cities in the east of the country (Murcia and Alicante), none of the other large Spanish cities (the darkest areas) are close to each other—a pattern observed in both time periods. From a spatial perspective, the population distribution is quite uneven, as the centre of the country is composed almost exclusively of small municipalities, except for Madrid and its surrounding area.

Moreover, the maps show the increasing inequality in city sizes, as many cities fall below the 5,000 population threshold over the century. This is especially the case in Galicia, the northwestern region of Spain. Besides the largest cities, the only municipalities that were able to gain population from 1900 to 2011 are those located around the main capital cities (Madrid, Barcelona, Valencia, etc.). This latter evidence is consistent with the traditional central place theory of Christaller (1933) and Lösch (1940): The tradeoff between scale economies and transportation costs gives rise to a pattern of central places, each serving the surrounding towns. Christaller argued that central places (i.e., cities) form a hierarchy, with the city at the top producing the entire range of urban products and lower order cities producing successively fewer products. To this theory, Lösch added the characteristic hexagonal shape of market areas, which minimizes transportation costs for a given density of central places. Although the central place theory was designed to explain towns serving as a rural market, its fundamentals hold for current business districts within metropolitan areas (Fujita et al., 1999).

Nevertheless, to confirm any spatial pattern, we carry out a systematic analysis of the spatial distribution of cities by considering space as continuous (González-Val, 2019b). We define four groups of city sizes in terms of population: lower than 3,000, 3,000–10,000, 10,000–15,000, and greater than 15,000 inhabitants in 1900, whereas in 2011 the population thresholds are lower than 5,000, 5,000–25,000, 25,000–50,000, and greater than 50,000 inhabitants.[6] The criterion used to define the thresholds of the

[6] As a robustness check, we changed the groups using different population sizes. The results did not qualitatively change.

different groups is that the number of cities in each of the categories should be similar in both years, especially for groups containing the large cities.[7] Then we study how the cities of similar size are distributed in space, following the methods of Duranton and Overman (2005, 2008). Although this methodology is often used to study the spatial distribution of firms, González-Val (2019b) applied it to analysing the spatial distribution of US cities.

First, we calculate the bilateral distance between all cities in a group. We define d_{ij} as the distance between cities i and j. Given n cities, the estimator of the density of bilateral distances (called K-density) at any point (distance) d is

$$\hat{K}_m(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{ij}}{h} \right),$$

where f is the Gaussian kernel function with bandwidth (smoothing parameter) h. To simplify the analysis, we consider only the range of distances between zero and 350 km. This threshold is the median distance between all pairs of cities (349.96 km, to be precise).[8]

Second, to distinguish the spatial location of cities from a random pattern, we must construct counterfactuals by first drawing locations from the overall cohort of cities and then calculating the set of bilateral distances. We assume that the set of all existing "sites" (S), i.e., all the cities in the distribution, represents the set of all possible locations for any city of a particular size. This means that, for instance, Barcelona could be located in any other place in Spain where a city exists. We run 2,000 simulations for each group of cities. In each simulation, the density of distances between pairs of cities is calculated as if the same number of cities within the group were randomly allocated across the set S of all possible locations (7,959 options). Sampling is done without replacement. Thus, for any of the four groups of cities A with n cities, we generate our counterfactuals $\tilde{A}_m$ for $m = 1,2,...,2000$ by sampling n elements without replacement from S, so that each simulation is equivalent to a random redistribution of cities across all the possible sites.

[7] Relative thresholds yield similar results. Actually, the most numerous group in both years defined by the lowest population threshold roughly corresponds to smaller-than-average cities.

[8] This is the median distance between all pairs of cities (including the closest cities but also the farthest ones). For instance, in the case of Madrid, which is located approximately in the centre of the country, the median distance between Madrid and all the other municipalities is 263 km.

To analyse the statistical significance of the localization pattern of cities, we compare the actual kernel density estimates to the simulated counterfactuals. To that end, global confidence bands are built using the simulated counterfactual distributions, following the method of Duranton and Overman (2005, 2008). Let $\overline{K}(d)$ be the upper global confidence band for a category of cities of a given size. This band is hit by 5% of our simulations between 0 and 350 km. In the same way, $\underline{K}(d)$ denotes the lower global confidence band, which is hit by 5% of the randomly generated K-densities that are not localized. Therefore, when the estimated K-densities lie within the global confidence bands for distance d, $\underline{K}(d) < \hat{K}(d) < \overline{K}(d)$, the spatial location of cities is not significantly different from randomness. Deviations from randomness involve a dispersion pattern if, graphically, the estimated K-densities fall below the lower global confidence band for at least one distance d, that is, when $\hat{K}(d) < \underline{K}(d)$. Analogously, when, graphically, $\hat{K}(d) > \overline{K}(d)$ and the K-densities lie above the upper global confidence band for any distance d, a localization pattern can be observed. For cities, the set S of all existing sites increases over distance (cities are distributed to cover all the country), which implies that, as our figures show, global bands are always increasing (the greater the distance, the higher the density of cities if they were randomly distributed).

Figures 3 and 4 show the results for the years 1900 and 2011, respectively. In all categories, significant deviations from randomness are observed. Figures 3(a) and 4(a) show the results for the smallest municipalities (fewer than 3,000 inhabitants in 1900 and fewer than 5,000 inhabitants in 2011), which are also the largest groups of cities; the estimated K-densities fall above the bands in both cases, pointing to a clear concentration (localization) pattern for almost all distances. Regarding medium-sized and large cities (the rest of the groups), the geographical pattern is similar across categories (Figures 3(b), 3(c), and 3(d) in 1900, and 4(b), 4(c), and 4(d) in 2011), indicating that beyond a distance threshold (around 100–175 km in 1900 and 75–100 km in 2011), a dispersion pattern emerges in all cases.

Nevertheless, there are two differences between results for medium-sized and large cities in those years. First, in 1900, for small distances, the K-density estimates lie between the confidence bands (see Figures 3(b), 3(c), and 3(d)), thus indicating a

random distribution of these cities for short distances. However, in 2011, for distances between zero and 50 km, we observe localization (see Figures 4(b), 4(c), and 4(d)), and then, around that threshold, the K-density crosses the confidence interval from above to below and, thus, the spatial distribution pattern changes from localization to dispersion. This means that in 2011 some large cities are located close to each other, although the rest are farther apart, pointing to different centres of urban systems, while we do not observe that in 1900. This finding is related to the recent suburbanization of the largest cities (Garcia-López et al., 2015), with people concentrating in cities surrounding the largest cities to avoid some of the congestion costs; see the maps in Figure 2.

The second difference is that, once the dispersion pattern emerges, in 1900 the estimated K-densities remain quite stable, while in 2011 the density of cities continuously decreases until a turning point at around 200 km (150 km for the top largest cities), when the density starts to increase (although the dispersion pattern still holds).

These geographical patterns support a hierarchical system of the Spanish cities in which the central city of each subsystem is far away from others in both periods. Overall, we could set 75–100 km as the general boundary of urban subsystems (on average) because, according to our results, city pairs beyond this distance are driven by dispersion and, hence, belong to different urban subsystems (there are some deviations; for instance, in 1900 the estimated density of cities for the group with 10,000–15,000 inhabitants crosses the confidence bands at a longer distance). Moreover, in 2011 pairs within this distance are likely to be driven by localization. A more conservative threshold would be a distance of 200–250 km, for which the estimated K-densities show a dispersion pattern in all cases in both years. In 2011 the density of large cities even recovers the initial values at that distance. Interestingly, the form of the density of large cities in 2011 is a U-shaped curve, suggesting a kind of agglomeration shadow that large cities cast on nearby big cities (25,000–50,000 and more than 50,000 inhabitants) until a distance of 250 km. The turning point at which the density of large cities starts to increase is around 200 km. Note that this pattern is observed only in 2011; in 1900 K-densities for large cities simply stabilize around a constant value.

If we focus on the results for the year 2011, as compared to the results found by González-Val (2019b) for US cities in 2010, Spanish cities show more pronounced spatial patterns, for both the small (localization) and medium-sized and large cities

(dispersion). This would indicate that Spanish urban subsystems are more clearly delineated. Why? The low Spanish internal migration rates may indicate that Spain is much more regionally segregated than the US; thus, each region in Spain can be viewed as a separate urban system, in contrast to the US. In that case, the population in each region may agglomerate towards the region's capital city (not coincidentally, capital cities are among the largest cities shown in Table 1).

4. The spatial city size distribution

Now, we follow the distance-based approach of González-Val (2019a, 2019c) to analyse the influence of distance on the city size distribution parameters. In this method, space is introduced through the selection of geographical samples of cities based on distances. This recursive procedure is based on the following steps:

1. Calculate the bilateral physical geographic distances between all cities, using the haversine distance measure.[9]

2. Define the limit of the geographical samples of neighbouring cities. The previous section provides evidence for the significant dispersion of Spanish large cities, which points to a hierarchical urban system. Although we have identified 75–100 km as the general limit of urban subsystems, in this analysis we consider all possible combinations of cities within a 200-km radius, a conservative (and higher) threshold, as previously we found that up until spatial distances of around 200 km, the density of large cities is low, especially in 2011. Henceforth, we take that distance as the spatial limit in our study of how city size distribution changes over space.[10]

3. Draw circles of radius r =10, 15,…, 200 around the geographic centroid of each city's coordinates, starting from a minimum distance of 10 km and adding 5 km for each subsequent iteration. This means that we obtain 39 geographical samples for each city.

4. Repeat this exercise for all cities. This provides 310,401 (7,959×39) geographical samples. Note that within these geographical samples, we consider

[9] 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.

[10] The particular results for each distance do not depend on the spatial limit considered. This limit only determines the point at which the recursive procedure stops. Results for distances longer than 200 km are available from the author upon request.

all cities with no size restriction. Moreover, as our city definitions are consistent over time, we have the exact same geographical samples in both years, with the same elements.

5. Fit the Pareto and the lognormal distribution to each geographical sample and run a goodness-of-fit test.

The first distribution we consider is the Pareto distribution. Let S denote the city size (measured by population); if this is Pareto-distributed, the expression $R = A \cdot S^{-a}$ relates the empirically observed rank R (1 for the largest city, 2 for the second-largest, and so on) to the city size, where $a > 0$ is the Pareto exponent.

Our first step is to test whether this distribution provides an acceptable fit to our geographical samples of Spanish cities. For each geographical sample, we conduct the statistical test for goodness-of-fit proposed by Clauset et al. (2009), based on the measurement of the "distance" between the empirical distribution of the data and the hypothesized Pareto distribution. This "distance" is compared to the distance measurements for comparable synthetic data sets drawn from the hypothesized Pareto distribution, and the p-value is defined 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; then the KS statistic is computed for the data and the fitted model.[12] The test samples from the observed data and checks how often the resulting synthetic distribution fit the actual data as poorly as the ML-estimated power law. Thus, 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) recommended 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, merely by chance, data that agree as poorly with the model as the data that we have. Therefore, this procedure allows us only to conclude whether the

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

[12] The procedure by Clauset et al. (2009) is designed to choose an optimal truncation point. However, in this paper we do not truncate our data in any case, so 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.

power law achieves a plausible fit to the data. This test was previously applied to European and US city size data previously by González-Val (2019a, 2019d).

Figures 5(a) and 5(c) show the results of the Pareto test by distance in 1900 and 2011, respectively. Here (and in the following graphs), the x-axis measures the distance to the initial city in each geographical sample. For each distance, the graphs represent the percentage of p-values lower than 0.1 over the total number of tests carried out at that distance.[13,14,15] The percentage of rejections of the Pareto distribution clearly increases with distance in 2011, but is always below 30%, even for the longest distance considered. The results for 1900 are a bit different, as rejections increase with distance only for short (15–60 km) and long (greater than 150 km) distances; for the rest of the distances, the percentage of rejections of the Pareto distribution remain stable or even decreases.[16] Importantly, the percentage of rejections is greater in 2011 than it is in 1900 (almost double for distances longer than 100 km). These results indicate that the Pareto distribution is a plausible approximation of the real behaviour of the data in our geographical samples in all cases and for any distance in both years. This means that most of the possible combinations of neighbouring cities, for which economic interactions and migratory flows are significant, are Pareto-distributed, regardless of city size (we do not impose any size restriction).

After concluding that the Pareto distribution is a plausible description of city sizes, we estimate the Pareto exponent. While previously we estimated the parameter by ML to conduct the goodness-of-fit test, we now apply Gabaix and Ibragimov's Rank-$1/2$ estimator. The reason is that this estimator performs better in small samples (Gabaix and Ibragimov, 2011). Furthermore, Gabaix and Ibragimov (2011) suggested

[13] We use the 0.1 reference value for the p-value, as Clauset et al. (2009) recommended. 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 small sample sizes, the log-likelihood cannot be computed and, thus, the test cannot be carried out. Single-city samples are also excluded. Therefore, the number of tests by distance is not constant. The number of tests carried out by distance ranges from 7,579 to 7,959 in 1900 and from 7,584 to 7,959 in 2011.

[15] The Bonferroni correction of the p-values for multiple comparisons yields similar results, although with a lower percentage of rejections.

[16] These graphs indicate that 200 km is a conservative threshold, because the actual limit is probably a lower value. Actually, if we set the limit to the distance at which K-density crosses the confidence interval from above to below, i.e., where localization turns dispersion (see Figures 3 and 4), the flip happens at around 100 km out. Note that if we consider this limit, we can observe a huge increase in rejections of the Pareto distribution for distances above 100 km in both years.

that their estimator produces more robust results than the ML estimator under deviations from power laws.

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

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

where ξ is the error term and b and a are the parameters that characterize 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 greater 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, which appears when Pareto's exponent of the distribution is equal to unity ($a = 1$). This means that the rank-size relationship is constant: The population of the second city is one-half that of the first, the population of the third city is one-third that of the first, and so on.

Equation (2) is estimated by OLS for all our geographical samples by distance. This means that, for each city, we obtain 39 different estimates of the Pareto exponent. This iterative estimation of the Pareto exponent by distance is repeated starting from every city. After running all the regressions with the geographical samples defined as explained above, we obtain 310,096 Pareto exponent–distance pairs (single-city samples are excluded). To summarise all these point-estimates, we conduct a non-parametric estimation of the relationship between distance and the estimated Pareto exponents using a local polynomial smoothing.[17] Figures 5(b) and 5(d) display the results, including the 95% confidence intervals. The graphs show similar behaviour in both 1900 and 2011: a continuous decrease of the Pareto exponent with distance. The only difference is in how close the estimates are to Zipf's law; while in 1900 the estimated coefficients converge to the value 1 (the longer the distance, the closer to 1 are the estimated coefficients; at the longest distance, 200 km, they almost coincide), meaning that Zipf's law holds for large samples of cities, in 2011 the value 1 falls within the

[17] We used the lpolyci command in STATA with the following options: local mean smoothing, a Gaussian kernel function, and a bandwidth determined using Silverman's (1986) rule-of-thumb.

confidence bands only for short distances (20–25 km). Therefore, Zipf's law holds in 1900 for city sizes considering all possible geographical samples (or sub-regions) of cities in wide areas (greater than 200 km), whereas in 2011 we cannot reject Zipf's law only for geographical samples at very short distances. This implies a decreasing validity over time of Zipf's law for Spanish cities.

The decreasing Pareto exponent converges to the value estimated for the whole sample of cities in both years (shown in Table 2), represented by the continuous horizontal lines. González-Val (2019a) argued that a possible explanation for this result could be that, as distance increases, so does the number of cities within the samples, which pulls down the coefficient (Eeckhout, 2004). The total Spanish land area is about 500,000 km^2, and the average land area of a municipality is 62.18 km^2, or a circle with a radius of roughly 4.5 km. Considering that the surface area πr^2 of a circle is a quadratic function of its radius r, at the minimum radius of 10 km the average city would have about 5-6 neighbours; as distance increases, the number of cities included in the circles naturally also increases because the number of cities asymptotically will be a quadratic function of r.

To address this possible issue, we run placebo regressions to test whether sample size is the only factor driving our results, as González-Val (2019a) suggested. Previously we constructed 39 geographical samples starting from each city, representing all the possible combinations of cities within a 200-km radius. Each geographical sample includes a particular number of cities; thus, we have 310,401 sample sizes. Now, we construct the same number of samples (39) starting from each city, but instead of including the 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. Next, using the Gabaix and Ibragimov (2011) specification (Equation (2)), we estimate the Pareto exponent for all these random samples of cities. Sample size is the same in random and geographical samples, but they share only 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 39 values of the difference between the Pareto exponents estimated using geographical and random samples.

This gives us 310,096 values, which we summarise by conducting a non-parametric estimation using a local polynomial smoothing of the relationship between distance and the difference between the Pareto exponents estimated using geographical and random samples. Figure 6 shows the results, including the 95% confidence bands. Note that this time, the x-axis represents sample size instead of distance. For small sample sizes, the difference between Pareto exponents estimated by using geographical and random samples is positive but decreases with sample size. As sample size increases, in 2011 the difference stabilises around a positive value, significantly different from zero, while in 1900 the difference even increases with sample size. The explanation for this result is that close cities have similar sizes, but as distance and sample size increase, the behaviour of the mean size and standard deviation is not monotonous (see Figure 7). However, the random selection of cities leads to random samples including cities whose mean size and standard deviations tend to the distributional values, when sample sizes are large. Thus, if there is any spatial pattern in city sizes, as we argue, significant deviations arise in the estimated Pareto exponents. This significant positive difference between the Pareto exponents estimated using geographical and random samples means that geography has a significant effect on the value of the Pareto exponent: Pareto exponents estimated using geographical samples of nearby cities are (on average) higher than those obtained with random samples of cities, regardless of sample size (although the difference is higher for small sample sizes). This result is similar to that found by González-Val (2019a) for US cities, confirming that neighbouring cities are more homogeneous in city size than random samples of cities. Nevertheless, our results indicate that nearby cities are more homogeneous in Spain than in the US, because the positive difference between the Pareto exponents estimated using geographical and random samples is higher for the Spanish municipalities than for the US cities.

The second classical distribution considered is the lognormal distribution. The two parameters of this distribution are μ and σ, namely, the mean and standard deviation of $\ln S_i$, which denotes the natural logarithm of city size. For many years, the lognormal distribution has been considered for purposes of studying city size (Richardson, 1973). More recently, Eeckhout (2004) fit 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. Lee and Li

(2013) modified the Roback model to generate a city size distribution that asymptotically follows the lognormal distribution.

Again, first, we compute a statistical test by distance to assess the validity of the distribution. The standard test to check the lognormal behaviour of a sample is the KS test, previously used with city sizes by Giesen et al. (2010) and González-Val et al. (2015), among others. As González-Val (2019c) pointed out, one well-known inconvenience of this test is its relatively low power: With very large sample sizes, it tends to systematically reject the null hypothesis unless the fit is almost perfect. Therefore, we expect that the power of the test will decrease with distance as the sample size increases. The KS test null hypothesis is that the two samples (the actual data and the fitted lognormal distribution) come from the same distribution.

Figure 7(a) shows the result of the KS test by distance. For each distance, the graphs represent the percentage of p-values lower than 0.05 over the total number of tests carried out at that distance (again single-city samples are excluded).[18,19] Support for the lognormal distribution clearly decreases with distance; in 1900 the increasing line is almost linear, while in 2011 the relationship seems concave. Therefore, the increase in rejections with distance is faster in 2011 than it is in 1900. For distances longer than 90 km, the percentage of rejections soon rises to higher than 50%, and for the longest distances the test rejects the lognormal distribution in most of the cases (almost 80%) in both years.

Thus, the lognormal distribution is valid only for short distances. 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 310,401 mean- and standard deviation-distance pairs. Again, to summarise all these values, we estimate the non-parametric distance–mean and distance–standard deviation relationships using a local polynomial smoothing. The rest of the panels in Figure 7 (b and c) display the estimates, including the 95% confidence intervals. The results are similar in the two years considered. The mean decreases with distance; it is important to note that, contrary to what González-Val (2019c) found for US cities, for Spanish cities, for the distances considered, the

[18] 5% is the significance level usually considered in the literature. If we use the 10% level, as in the Pareto test, we obtain similar results to those shown in Figure 7(a).
[19] Again, the Bonferroni correction of the p-values for multiple comparisons provides similar results, although with a lower percentage of rejections.

average values diverge (with distance) from the mean of the whole sample (Table 2), represented by the horizontal line. Nevertheless, the standard deviation increases with distance and soon converges to the value for the whole sample.

5. Conclusions

In this paper, we adopted the approach proposed by González-Val (2019a), who argued that the proper statistical function of city size distribution is a matter of both size and distance. Therefore, instead of focusing only on whether the Pareto distribution (and Zipf's law) holds for the largest cities, we are interested in the city size distribution of neighbouring cities. Using un-truncated data from all Spanish cities in 1900 and 2011, we study the spatial distribution of cities, and how the city size distribution varies with distance throughout the course of over a century.

First, K-densities are estimated using the method of Duranton and Overman (2005) to identify different spatial patterns depending on city size, with similar results obtained in both 1900 and 2011. We have found that small Spanish cities are concentrated for almost all distances, while medium-sized and large cities show a dispersion pattern beyond a certain threshold. Overall, we identify 75–100 km as the general boundary of urban subsystems (on average), as large city pairs beyond this distance are driven by dispersion and, thus, belong to different urban subsystems. These results are consistent with the central place theory, pointing to an urban hierarchy with different tiers of city sizes. At the upper tiers, there is a low number of dispersed medium-sized and large cities, while at the lower tiers we find many small cities (a bit over 80% of the total number of cities) concentrated. These geographical patterns support a hierarchical system of cities similar to that found by González-Val (2019b) in the US, in which the central city of each subsystem would be far away from the other ones. In the words of Cronon (1991): "Cities were like stars or planets, with gravitational fields that attracted people and trade like miniature solar systems."

Second, by using the distance-based approach proposed by González-Val (2019a), we analysed the influence of distance on the city size distribution parameters, considering the two traditional distributions used to fit population data: Pareto and lognormal. By using all the possible combinations of cities within a 200-km radius (a conservative threshold, twice the observed limit of the urban subsystems), we produced results indicating that, in both periods, the Pareto distribution cannot be rejected in most

of the cases, regardless of city size. This means that the Pareto distribution fits well city size distribution for Spanish cities of all sizes as long as they are located nearby. On the other hand, the lognormal distribution is valid only for short distances.

Although our results support the validity of the Pareto distribution (the traditional benchmark distribution) at the local level, evidence supporting Zipf's law weakens over time. While Zipf's law holds in 1900 for city sizes considering all possible geographical samples (or sub-regions) of cities in wide areas (at the longest distance, 200 km, the estimated Pareto exponent is almost 1), in 2011 we cannot reject Zipf's law only for geographical samples at very short distances (20–25 km). This evidence may suggest that Zipf's law is becoming obsolete for Spanish cities, as its validity is reduced over time to a narrow set of distances. Some explanations for this change could be the low internal migration rates, an aging population, a very unequal spatial distribution of the population (nowadays over 75% of Spanish municipalities have fewer than 2,500 inhabitants), and the increasing weight of service activities (a constant or even decreasing returns to scale sector) while the microfoundations in most theoretical models are industrial productivity or technology shocks.

Finally, we run placebo regressions to ensure that sample size did not drive our results, confirming the significant effect of geography on the Pareto exponent because the exponents estimated using geographical samples of nearby cities are (on average) higher than those obtained from random samples of cities. This result indicates that neighbouring cities are more homogeneous in city sizes than are random samples of cities.

References

Arribas-Bel, D., M.-À. Garcia-López, and E. Viladecans-Marsal, (2020). Building(s and) cities: Delineating urban areas with a machine learning algorithm. Journal of Urban Economics, forthcoming.

Beltrán, F. J., A. Díez-Minguela, and J. Martínez-Galarraga, (2017). The Shadow of Cities: Size, Location, and the Spatial Distribution of Population in Spain. Cambridge Working Paper Economics 1749, Faculty of Economics, University of Cambridge.

Bentolila, S., (1997). Sticky labor in Spanish regions. European Economic Review, 41: 591–598.

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.

Cheshire, P. C., and S. Magrini, (2006). Population Growth in European Cities: Weather Matters – but only Nationally. Regional Studies, 40(1): 23–37.

Christaller, W., (1933). Die Zentralen Orte in Suddeutschland. Translated by C. W. Baskin (1966) to Central Places in Southern Germany, Prentice-Hall, Englewood Cliffs, NJ.

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

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

Cronon, W., (1991). Nature's Metropolis: Chicago and the Great West. New York: W.W. Norton.

Cuberes, D., K. Desmet, and J. Rappaport, (2019). Urban Growth Shadows. Federal Reserve Bank of Kansas City, Research Working Paper no. 19-08, November. Available at https://doi.org/10.18651/RWP2019-08

Dao, M., D. Furceri, and P. Loungani, (2017). Regional Labor Market Adjustments in the United States: The Role of Recessions. Review of Economics and Statistics, 99: 243–257.

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

Duranton, G., and H. G. Overman, (2005). Testing for Localization Using Microgeographic Data. Review of Economic Studies, 72: 1077–1106.

Duranton, G., and H. G. Overman, (2008). Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data. Journal of Regional Science, 48(1): 213–243.

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

Fazio, G., and M. Modica, (2015). Pareto or log-normal? Best fit and truncation in the distribution of all cities. Journal of Regional Science, 55(5): 736–756.

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

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.

Garcia-López, M.A., A. Holl, and E. Viladecans-Marsal, (2015). Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. Journal of Urban Economics, 85: 52-67.

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.

Goerlich, F.J., J. Azagra, M. Mas, and P. Chorén, (2006). La localización de la población española sobre el territorio. Un siglo de cambios: Un estudio basado en series homogéneas (1900-2001). Bilbao: Fundación BBVA.

Goerlich, F.J., F. Ruiz, P. Chorén, and C. Albert, (2015). Cambio en la estructura y localización de la población. Una visión de largo plazo (1842-2011). Bilbao: Fundación BBVA.

Goldstein, M. L., S. A. Morris and G. G. Yen, (2004). Problems with fitting to the Power-law distribution. The European Physical Journal B - Condensed Matter, 41(2): 255–258.

González-Val, R., (2019a). US city size distribution and space. Spatial Economic Analysis, forthcoming.

González-Val, R., (2019b). The spatial distribution of US cities. Cities, 91: 157–164.

González-Val, R., (2019c). Lognormal city size distribution and distance. Economics Letters, 181: 7–10.

González-Val, R., (2019d). Historical urban growth in Europe (1300–1800). Papers in Regional Science, 98(2): 1115–1136.

González-Val, R., L. Lanaspa, and F. Sanz-Gracia, (2014). New evidence on Gibrat's Law for cities. Urban Studies, 51(1): 93–115.

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.

González-Val, R., D. A. Tirado, and E. Viladecans-Marsal, (2017). Market potential and city growth: Spain 1860–1960. Cliometrica, 11(1): 31–61.

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 S. Skouras, (2013). US city size distribution: Robustly Pareto, but only in the tail. Journal of Urban Economics, 73: 18–29.

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

Lanaspa, L., A. M. Perdiguero, and F. Sanz, (2004). La distribución del tamaño de las ciudades. El caso de España (1900-1999). Revista de Economía Aplicada, 34(vol. XXII): 5–16.

Lanaspa L, F. Pueyo, and F. Sanz, (2003). The Evolution of Spanish Urban Structure during the Twentieth Century. Urban Studies, 40: 567–580.

Lee, S., and Q. Li., (2013). Uneven landscapes and city size distributions. Journal of Urban Economics, 78: 19–29.

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

Lösch, A., (1940). The Economics of Location. Fischer Jena: English translation. Yale University Press, New Haven, CT, 1954.

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

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.

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

Richardson, H. W., (1973). Theory of the distribution of city sizes: Review and prospects. Regional Studies, 7: 239–251.

Romero Valiente, J. M., (2003). Migraciones. In: Tendencias demográficas durante el siglo XX en España, A. Arroyo Pérez ed., Instituto Nacional de Estadística, 209–253.

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

Silverman, B. W., (1986). Density Estimation for Statistics and Data Analysis. 1st Edition. Chapman and Hall/CRC. Cleveland.

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

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

Table 1. Bilateral physical distances between the 10 largest cities in Spain

A. 1900

Rank	City	Population	S_{MAD}/S	Bilateral distances (km)								
1	Madrid	575,675	1.0	.								
2	Barcelona	539,103	1.1	504.6	.							
3	Valencia	215,687	2.7	300.8	302.8	.						
4	Sevilla	147,271	3.9	390.8	830.4	541.5	.					
5	Málaga	134,849	4.3	415.0	769.9	468.0	158.0	.				
6	Murcia	109,930	5.2	348.2	472.0	178.2	433.1	323.1	.			
7	Cartagena	103,373	5.6	390.0	500.4	215.1	442.6	320.0	44.6	.		
8	Zaragoza	98,204	5.9	273.4	256.2	246.2	646.1	627.8	408.9	451.1	.	
9	Bilbao	91,337	6.3	323.0	468.1	471.2	702.6	737.9	605.5	650.1	244.6	.
10	Granada	75,570	7.6	359.5	682.4	380.0	213.0	88.8	235.6	236.0	550.2	678.6

B. 2011

Rank	City	Population	S_{MAD}/S	Bilateral distances (km)								
1	Madrid	3,198,645	1.0	.								
2	Barcelona	1,611,012	2.0	504.6	.							
3	Valencia	792,054	4.0	300.8	302.8	.						
4	Sevilla	698,041	4.6	390.8	830.4	541.5	.					
5	Zaragoza	678,115	4.7	273.4	256.2	246.2	646.1	.				
6	Málaga	561,435	5.7	415.0	769.9	468.0	158.0	627.8	.			
7	Murcia	437,666	7.3	348.2	472.0	178.2	433.1	408.9	323.1	.		
8	Bilbao	351,355	9.1	323.0	468.1	471.2	702.6	244.6	737.9	605.5	.	
9	Alicante	329,325	9.7	358.4	407.0	126.0	495.2	369.7	391.3	69.3	583.4	.
10	Córdoba	328,326	9.7	296.6	711.5	421.5	120.0	536.1	132.8	320.5	618.1	379.5

Notes: Source: INE Censuses and Goerlich et al. (2006, 2015). S_{MAD}/S is the quotient of Madrid's population divided by i's population. Bilateral distances are calculated using the haversine distance measure.

Table 2. Descriptive statistics

A. Descriptive statistics

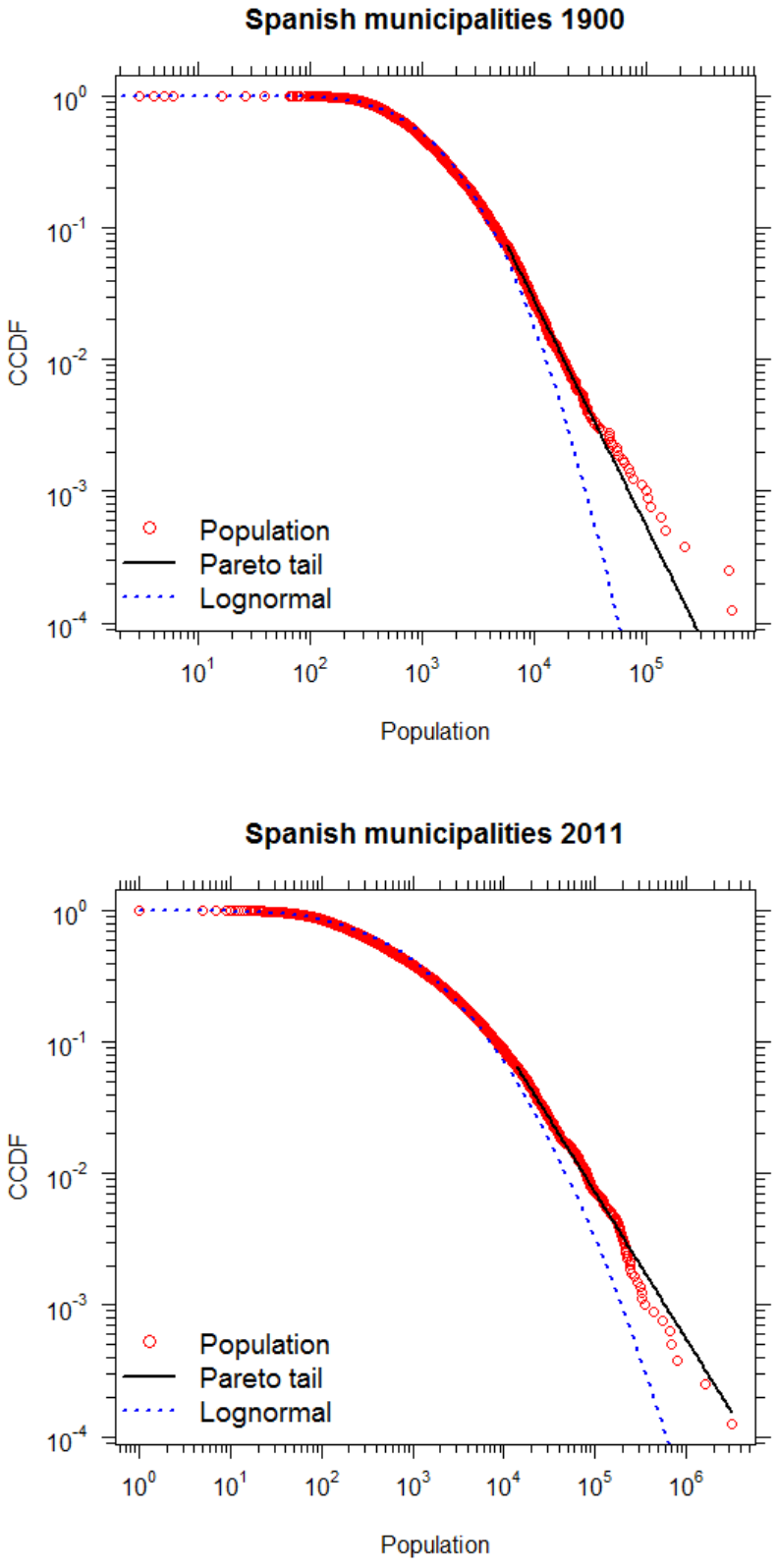
Year	Mean size	Standard deviation	Minimum	Maximum	Cities
1900	2,278.10	10,357.30	3	575,675	7,959
2011	5,462.02	46,668.18	1	3,198,645	7,959

B. Statistical Distributions

	Pareto distribution		Lognormal distribution	
Year	Pareto exponent	Standard error	Mean	Standard deviation
1900	0.89	0.014	6.97	1.06
2011	0.52	0.008	6.52	1.83

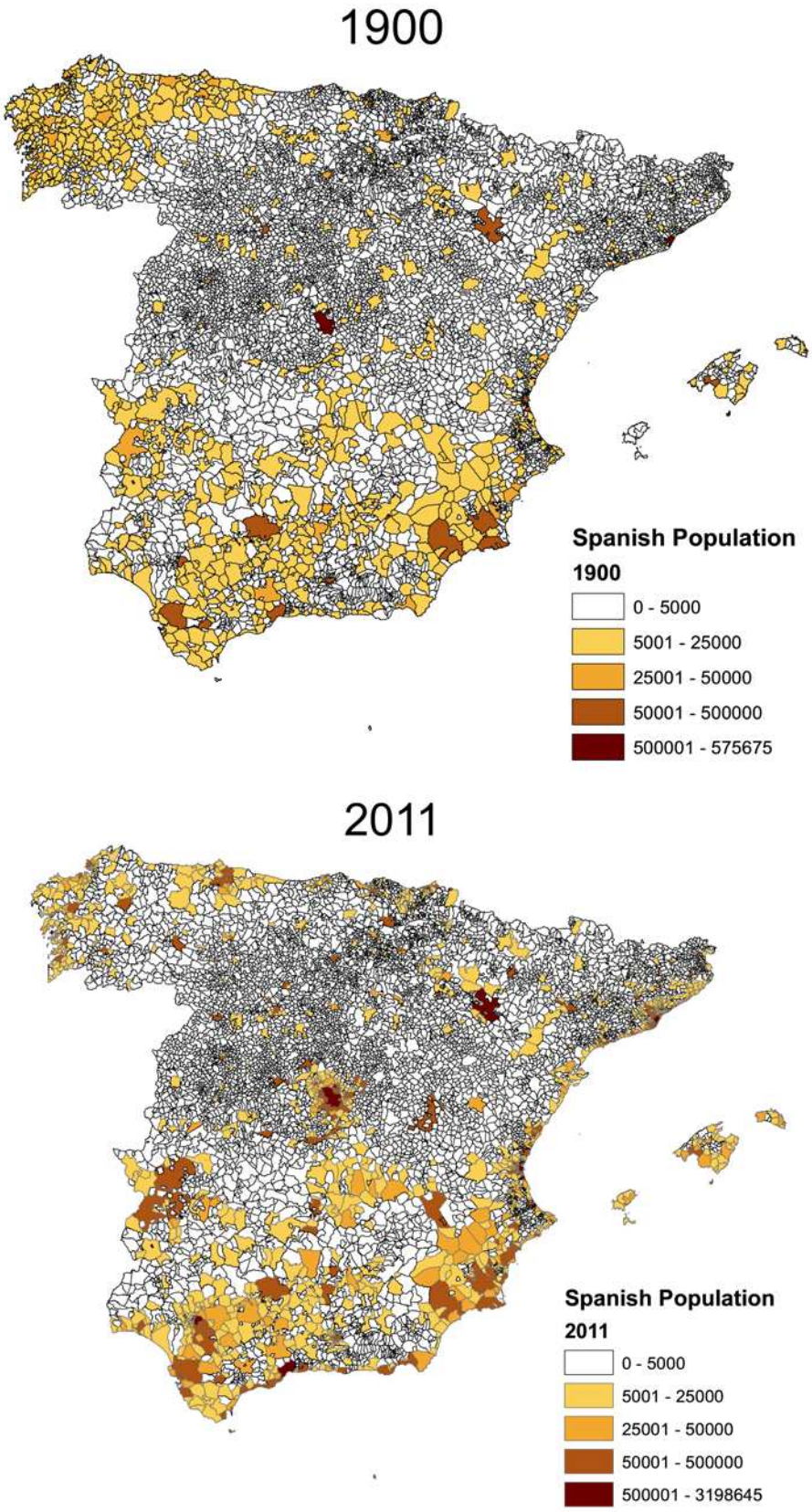
Notes: The Pareto exponent is estimated using Gabaix and Ibragimov's Rank-$1/2$ estimator. Standard errors are calculated by applying Gabaix and Ioannides's (2004) corrected standard errors: $\text{GI s.e.} = \hat{a} \cdot (2/N)^{1/2}$, where N is the sample size. The lognormal parameters are estimated by maximum likelihood.

Figure 1. Spanish city size distribution in 1900 and 2011



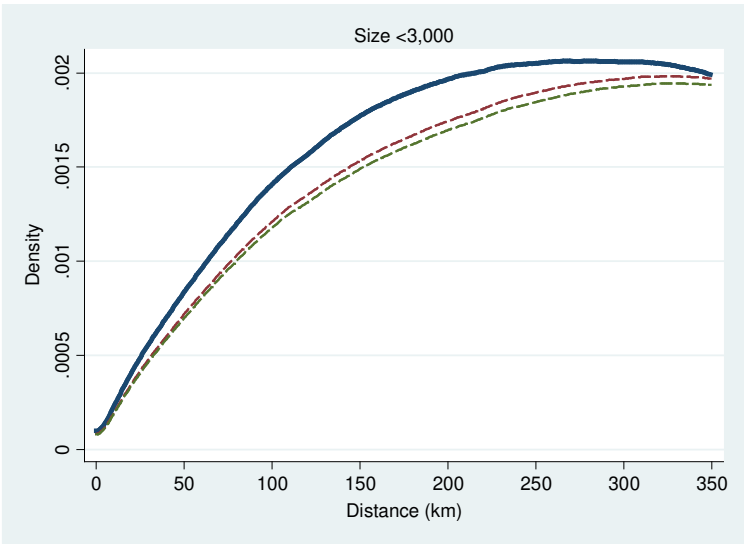
Notes: The Pareto upper-tail threshold is set at 5,717 and 14,024 in 1900 and 2011, respectively, using Clauset et al.'s (2009) methodology.

Figure 2. City sizes in 1900 and 2011

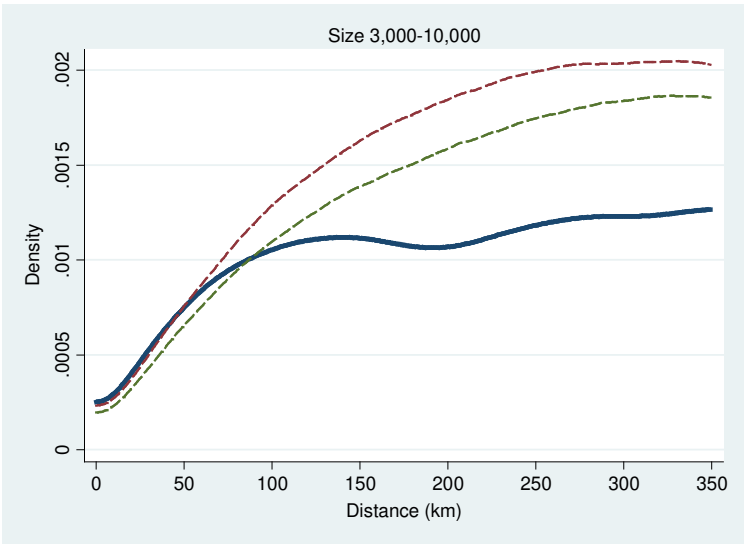


Notes: Geographical boundaries are defined according to the census in 2001 (Goerlich et al., 2006, 2015).

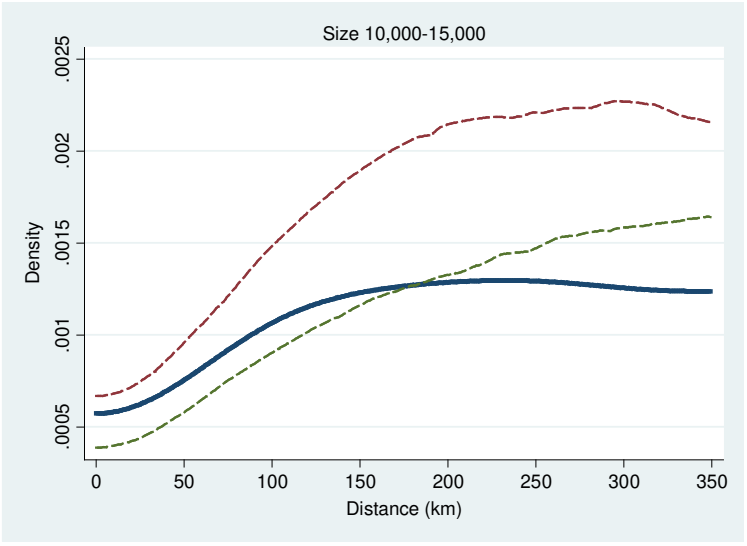
Figure 3. Spatial distribution of cities by size, Spanish municipalities in 1900



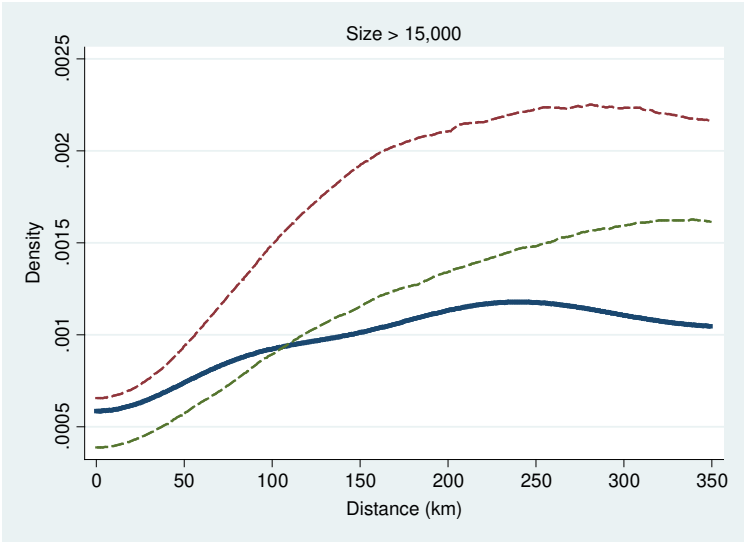
(a) <3,000 (6,629 cities)



(b) 3,000–10,000 (1,119 cities)



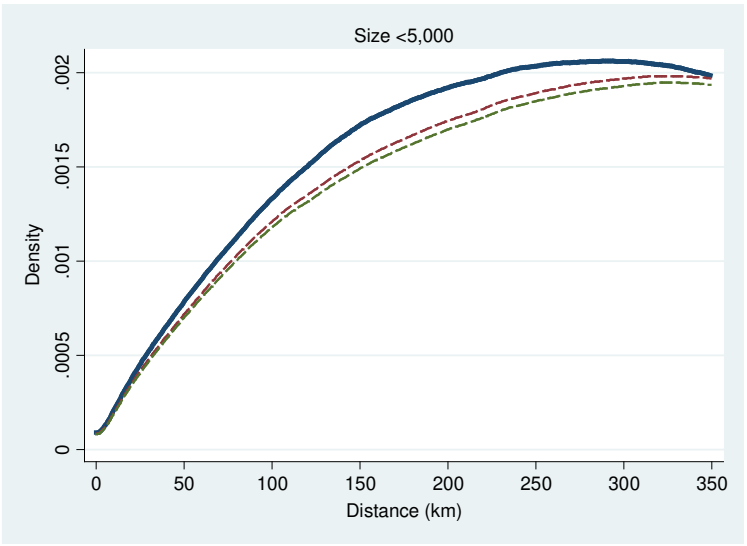
(c) 10,000–15,000 (105 cities)



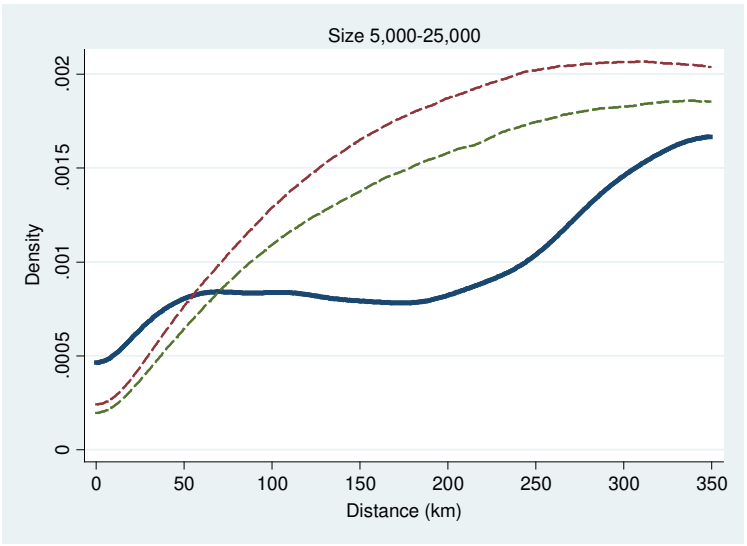
(d) >15,000 (106 cities)

Notes: K-densities are estimated using the method of Duranton and Overman (2005). Dashed lines represent the 95% global confidence bands, based on 2,000 simulations.

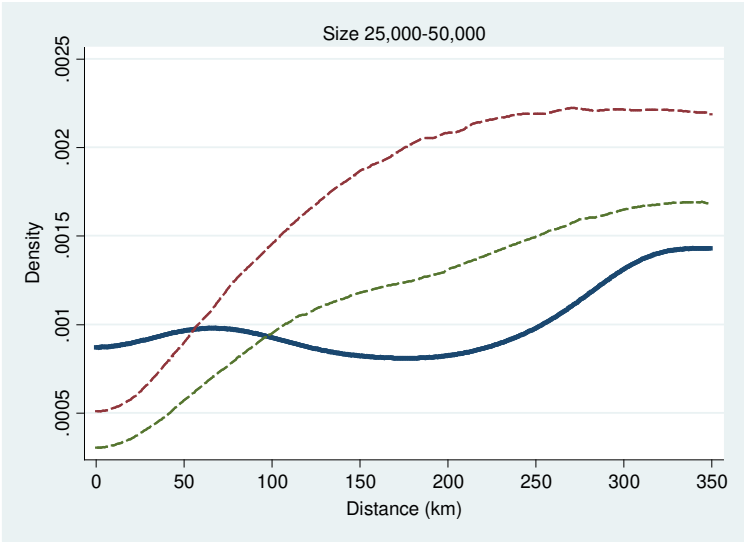
Figure 4. Spatial distribution of cities by size, Spanish municipalities in 2011



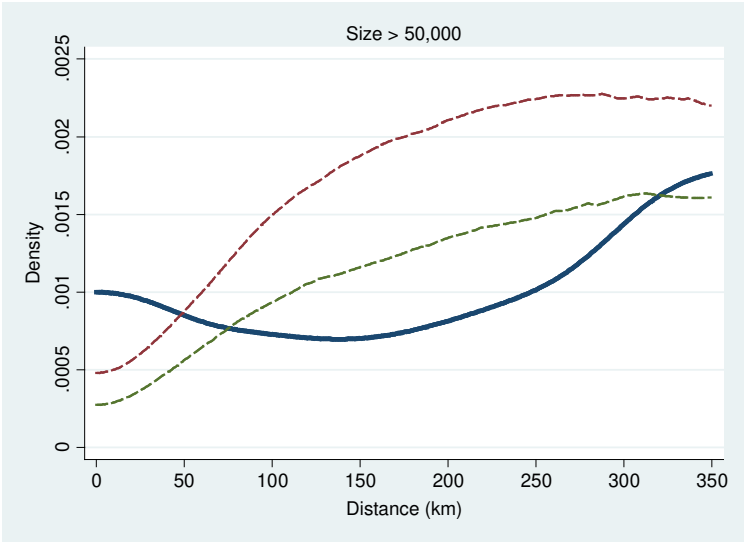
(a) <5,000 (6,754 cities)



(b) 5,000–25,000 (932 cities)



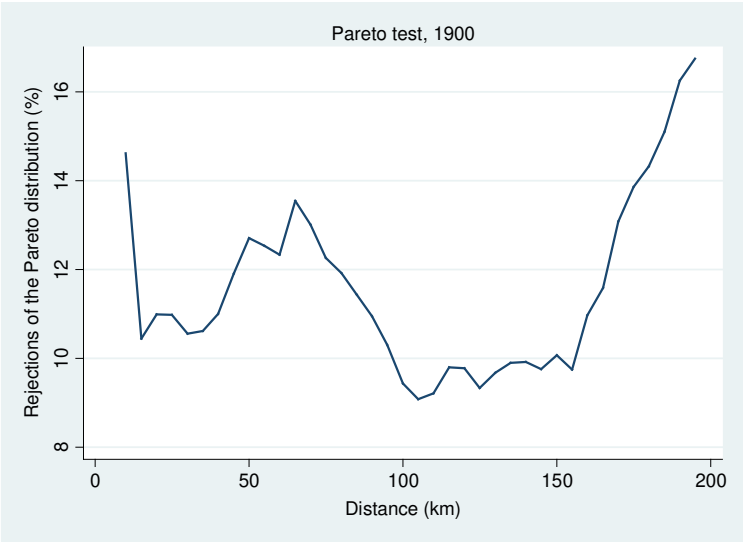
(c) 25,000–50,000 (140 cities)



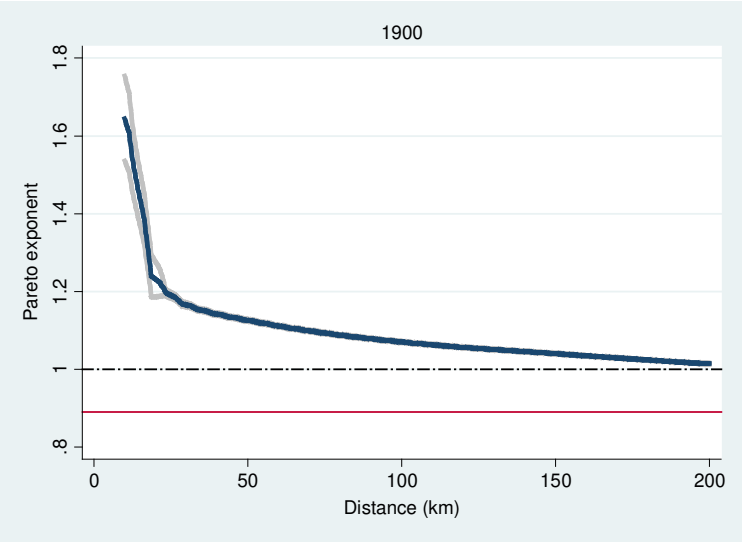
(d) >50,000 (133 cities)

Notes: K-densities are estimated using the method of Duranton and Overman (2005). Dashed lines represent the 95% global confidence bands, based on 2,000 simulations.

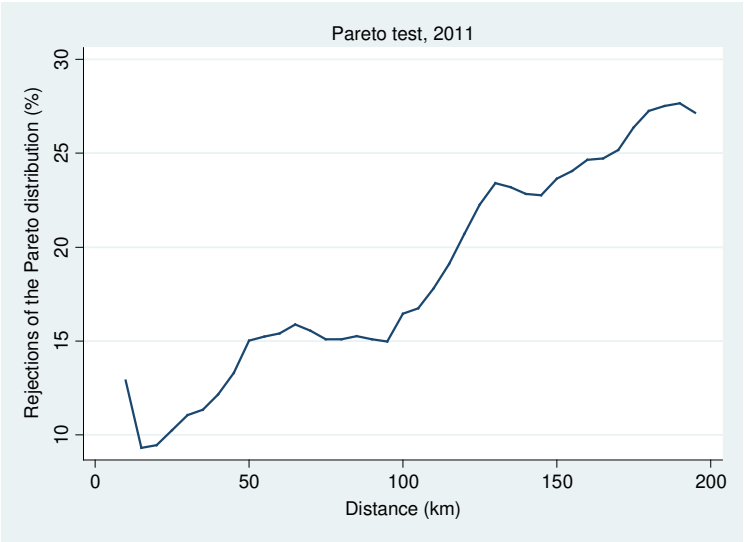
Figure 5. Pareto distribution over space: Distribution test and Pareto exponent by distance in 1900 and 2011



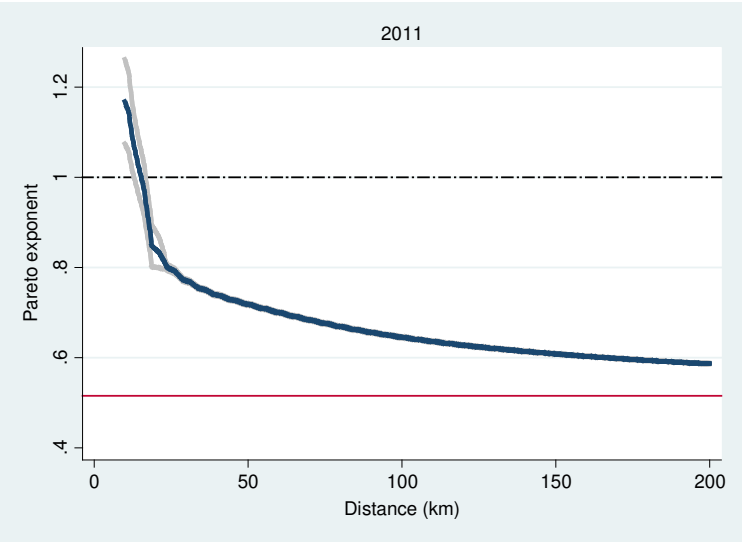
(a) Pareto test (1900)



(b) Pareto exponent (1900)



(c) Pareto test (2011)



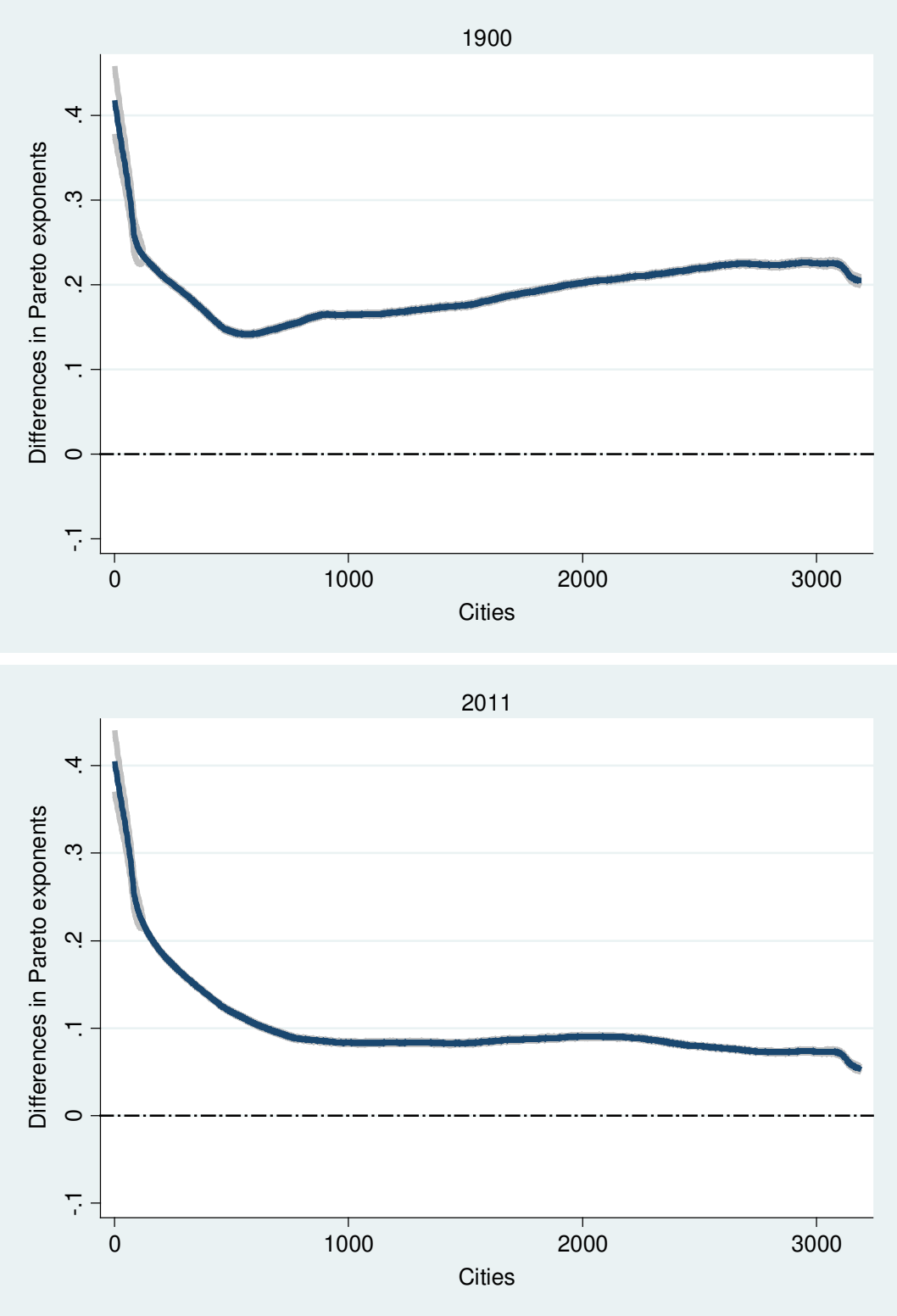
(d) Pareto exponent (2011)

Notes:

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

Figures (b) and (d) show the non-parametric relationship between distance and the estimated Pareto exponents including the 95% confidence intervals, based on 310,096 Pareto exponent–distance pairs. The horizontal line represents the estimated Pareto exponent for the entire sample of cities (see Table 2).

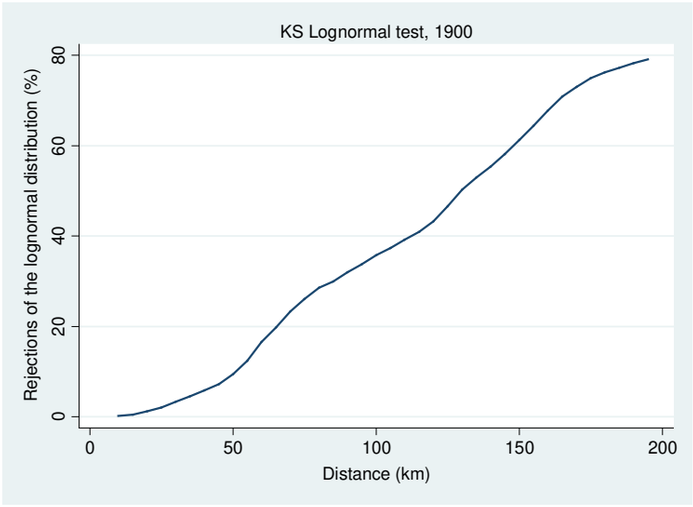
Figure 6. Placebo regressions: Differences between geographical samples and random samples by sample size

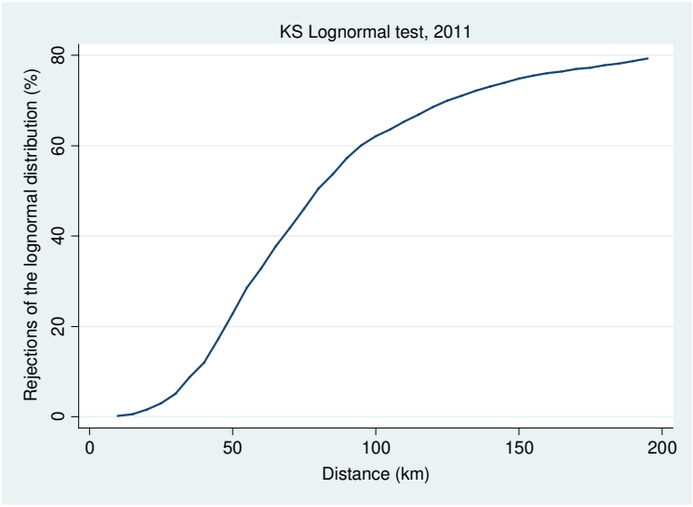


Notes:

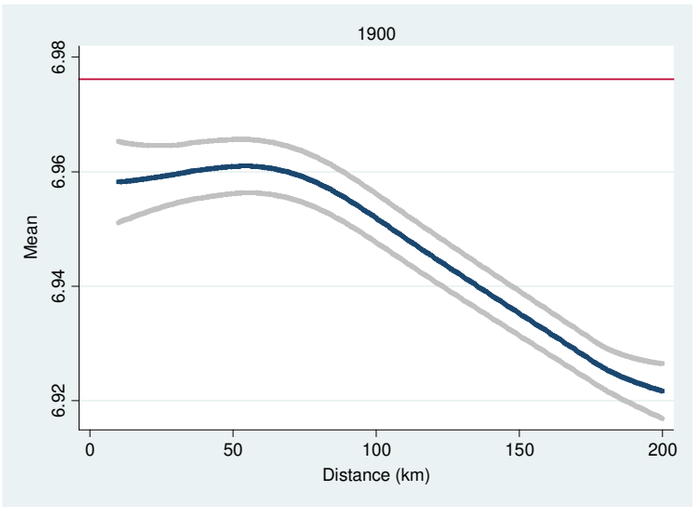
Figures shows the non-parametric relationship between distance and the difference between Pareto exponents estimated using geographical and random samples, including the 95% confidence intervals, based on 310,096 observations.

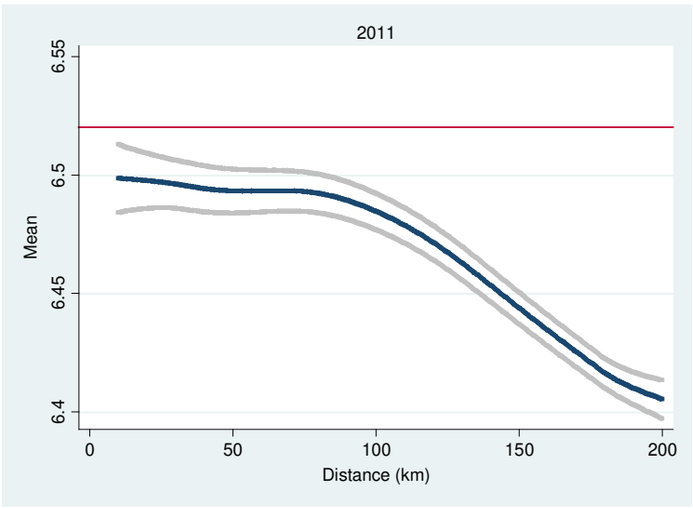
Figure 7. Lognormal distribution over space: Distribution test, mean and standard deviation by distance



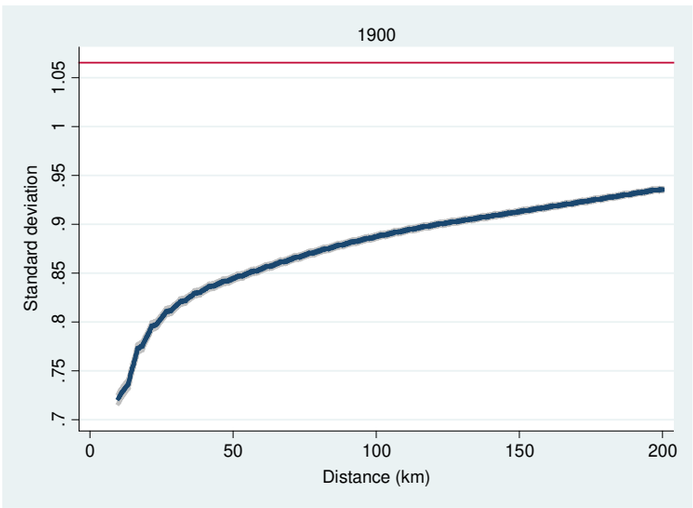


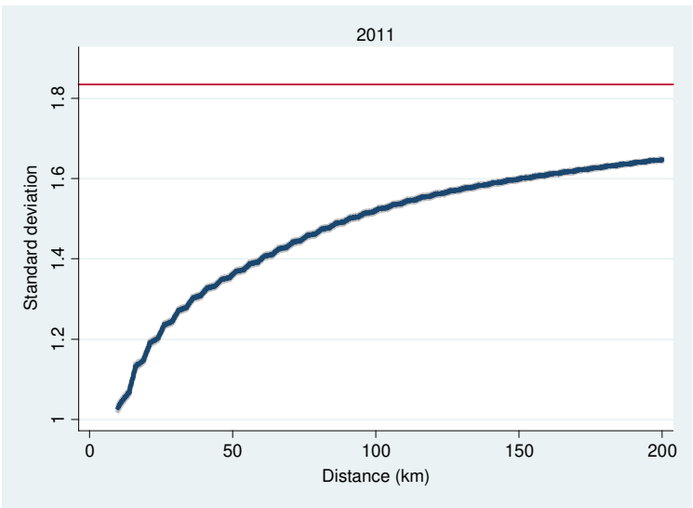
(a) KS test





(b) Mean





(c) Standard deviation

Notes:

Figure (a): Percentage of rejections of the Kolmogorov–Smirnov test of the lognormal distribution at the 5% level.

Figures (b) and (c) show the non-parametric distance–mean and distance–standard deviation relationships, respectively, including the 95% confidence intervals, based on 310,401 observations. The horizontal lines represent the values for the entire sample of cities (see Table 2).