# The spatial distribution of US cities

González-Val, Rafael

Universidad de Zaragoza  IEB

September 2018

# The spatial distribution of US cities

Rafael González-Val

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

***Abstract*:** In this paper, we consider the distribution of bilateral distances between all pairs of cities to estimate K-densities using the methodology by Duranton and Overman (2005), identifying different spatial patterns. By using data from different definitions of US cities in 2010 (places, urban areas, and core-based statistical areas), we analyse the spatial distribution of cities, finding significant patterns of dispersion depending on the city size and city definition. Our results lend support to a hierarchical system of US cities in which the central cities of each subsystem are far away from each other.

***Keywords*:** space, city size, urban hierarchy, distance-based approach.

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

***Acknowledgements:***

Financial support was provided by the Spanish Ministerio de Economía y Competitividad (ECO2017-82246-P and ECO2016-75941-R projects), the DGA (ADETRE research group) and FEDER.

## 1. Introduction

City size distribution is the subject of numerous empirical investigations by urban economists, statistical physicists, and urban geographers. Over the years, the Pareto distribution (Pareto, 1896), also known as a power law, has generated a huge amount of research and considerable acceptance. The well-known Zipf's law (Zipf, 1949) is a particular case of the Pareto distribution in which the Pareto exponent is equal to one, which means that the second-largest city in a country is exactly half the size of the largest one, the third-largest city is one-third of the size of the largest, and so on. A lot of studies test the validity of this law (see the surveys by Cheshire, 1999, Nitsch, 2005; Soo, 2005) for many different countries.

Recent developments in this literature come from the new sample selection criterion by Eeckhout (2004), who demonstrates the statistical importance of considering the whole un-truncated sample and concludes that city size distribution is actually lognormal rather than Pareto. Many studies published since then consider un-truncated data (Giesen et al., 2010; González-Val et al., 2015; Ioannides and Skouras, 2013), and, after proposing new distributions, the current state of the art (Giesen and Suedekum, 2014; Ioannides and Skouras, 2013) is that, although most of the distribution is non-linear, the Pareto distribution (and Zipf's law) holds *for the largest cities*. This claim supposedly reconciles the old body of empirical literature focused on the largest cities with the new wave of empirical studies using un-truncated city sizes.

Nevertheless, the standard statistical analysis omits one important issue: the spatial distribution of the elements in the upper-tail. An important feature of city size distribution is the spatial dependence among the elements of the distribution: there is a relationship between the population of a city and the populations of nearby cities, because cities are connected through migratory flows. Actually, an essential assumption in urban models to obtain the spatial equilibrium is free migration across cities. However, the upper-tail of the distribution contains large cities that are very far away from each other. Table 1 shows the bilateral physical distances between the 10 largest cities in the United States (US) in 2010, using three different city definitions: places, urban areas, and core-based statistical areas. In the three samples, New York is the largest city, and $S_{NY}/S$ (the quotient between New York's population and city *i*'s population) reports how closely these top-10 cities align with Zipf's law (the quotient represents the so-called 'rank-size rule').[1] However, the important point is that the average physical distance between these cities is 1,241.3, 1,070.5, and 1,073.7 miles for places, urban areas, and core-based statistical areas, respectively. Therefore, there is a great distance between these largest cities (on average).[2]

Rauch (2014) investigates whether there are significant migrations for long distances.[3] Using the 2000 US census, he obtains evidence for moved distances. He creates bins of 100 kilometres in size (approximately 62 miles), concluding that over 68% of observations, the large majority of people, fall into the bin with a distance between 0 and 100 km, and that the majority of US citizens live near their place of birth. Only a share of 0.00017 of all people are included in the largest distance in his data set, the distance between California and Maine, with roughly 2,610 miles: those who either were born in California and live in Maine or were born in Maine and live in California. Rauch (2014) also estimates the relationship between the number of people and the distance between home and place of birth using a standard gravity equation, finding that this relationship decreases with distance.

[1] The rank size rule is a deterministic rule, which is not exactly equivalent to Zipf's law but can be a good approximation (Gabaix and Ioannides, 2004).

[2] In Section 4, we carry out an analysis of the spatial distribution of all cities, not just the 10 largest.

[3] Obviously significant commuting cannot occur across such wide distances, because commuting usually takes place within metropolitan areas from surrounding cities to the central place, so in the discussion we focus only on migration.

This empirical evidence indicates that there are no significant migratory flows between the largest cities in the upper-tail distribution because they are so far from each other. This means that, although New York, Los Angeles, and Miami are cities within the same country, actually they are the centres of different urban systems. The empirical literature on city size distribution usually omits this spatial issue; thus, the interpretation of the results is reduced to identifying the Pareto upper-tail regardless of whether there is any meaningful relationship between the largest cities.

In this paper, we deal with this spatial issue by analysing the distances between cities and thus finding some empirical facts about the spatial distribution of US cities. Our paper is inherently related to the system-of-cities literature.[4] Basically, theories stemming from this literature generate different subsystems of cities, composed of a few large cities surrounded by many medium-sized and small cities. The seminal paper that analyses how a group of cities develops and grows in a theoretical framework is by Henderson (1974). Henderson's modelling uses a general equilibrium analysis that provides an overview of the basic theoretical propositions about a system of cities.[5] An important question that the Henderson model addresses is how cities' populations relate to each other. The first step in assessing the validity of these theories entails an analysis of the spatial distribution of cities.

Some theoretical papers study the spatial interactions between surrounding cities. The literature often considers cities' market potential as a good proxy for agglomeration economies, although the direction of the effect of changes in the market potential on city growth is unclear. The New Economic Geography (NEG) theory literature (Fujita et al., 1999; Krugman, 1991, 1996) in many cases predicts a hierarchy

[4] A comprehensive review of the vast literature on this topic is beyond the scope of this paper.

[5] Other examples of early theoretical papers on the systems of cities are those by Henderson (1982a, 1982b), Henderson and Ioannides (1981), Hochman (1981), and Upton (1981).

of cities, in which the availability of services increases when moving towards the top of the hierarchy. Although the greater market potential should foster growth (the rationale being that nearby cities offer a larger market and hence more possibilities for selling products), this hierarchy can also generate ‘agglomeration growth shadows’, in which the spatial competition near higher-tiered centres constrains the growth of local businesses (Partridge et al., 2009). Location theory and hierarchy models (Dobkins and Ioannides, 2001) suggest that increasing market potential could affect city growth negatively, because the forces of spatial competition separate the larger cities from each other, so the bigger a city grows, the smaller its neighbouring cities will be.

Nevertheless, although there is a sizeable body of theoretical research, the empirical evidence remains limited to a few papers, probably because systems of cities are difficult to isolate as scientific objects of study (Pumain, 2006). Partridge et al. (2009) study whether proximity to same-sized and higher-tiered urban centres affected the patterns of 1990–2006 US county population growth. Their results show that, rather than casting NEG agglomeration shadows on nearby growth, larger urban centres generally appear to have had positive growth effects for more proximate places with fewer than 250,000 people, although there is some evidence that the largest urban areas cast growth shadows on proximate medium-sized metropolitan areas and that there is spatial competition among small metropolitan areas. Dobkins and Ioannides (2001) explore the spatial interactions among US cities by using a data set of metro areas from 1900 to 1990 and spatial measures including distance from the nearest larger city in a higher-tier, adjacency, and location within US regions. They find that, among cities that enter the system, larger cities are more likely to locate near other cities, and older cities are more likely to have neighbours. Hsu et al. (2014) find strong empirical support for what they call ‘the spacing-out property’ in the US: larger cities tend to be widely spaced, with smaller cities grouped around these centres.

In this paper, Table 1 provides some anecdotal evidence on the great bilateral distances between the top 10 largest cities in the US. However, to corroborate whether this spatial pattern is consistent with the geographical distribution of all cities, we must carry out a systematic analysis of the spatial distribution of cities by considering space as continuous. In Section 4, we study the spatial distribution of cities by using Duranton and Overman’s (2005) methodology, which we present in Section 3. We obtain evidence supporting a dispersion pattern of the largest cities, which would indicate the existence of multiple urban subsystems. Section 2 presents the database that we use, and Section 5 concludes.

## 2. Data

There are various ways of defining a city. The US Census Bureau provides statistical information for several geographical levels, and the US city size distribution has been analysed using different spatial units: states (Soo, 2012), counties (Beeson et al., 2001; Desmet and Rappaport, 2017), minor civil divisions (Michaels et al., 2012), metropolitan areas (Black and Henderson, 2003; Dobkins and Ioannides, 2000, 2001; Ioannides and Overman, 2003), and economic areas, defined by the Bureau of Economic Analysis (Berry and Okulicz-Kozaryn, 2012), or by using the city clustering algorithm (Rozenfeld et al., 2011).

In this paper, we consider three different definitions of cities: places, urban areas, and core-based statistical areas. Table 2 shows the descriptive statistics. Our data come from the 2010 US decennial census. The geographical coordinates (latitude and longitude) needed to compute the bilateral distances between cities are obtained from the 2010 Census US Gazetteer files.[6]

The generic denomination 'places' includes, since the 2000 census, all incorporated and unincorporated places. The US Census Bureau uses the generic term 'incorporated place' to refer to a type of governmental unit incorporated under state law as a city, town (except the New England states, New York, and Wisconsin), borough (except in Alaska and New York City), or village and having legally prescribed limits, powers, and functions. 'Unincorporated places' (which were renamed Census Designated Places, CDPs, in 1980) designate a statistical entity, defined for each decennial census according to the Census Bureau guidelines, and comprise a densely settled concentration of population that is not within an incorporated place but is locally identified by a name. Unincorporated places are the statistical counterpart of incorporated places, and the difference between them, in most cases, is merely political and/or administrative. These places have been used in recent empirical analyses of US city size distribution (Eeckhout, 2004, 2009; Giesen et al., 2010; González-Val, 2010; Levy, 2009), because they do not impose any truncation point (populations range from 1 to 8,175,133 inhabitants).

'Urban area' is the generic term for urbanized areas and urban clusters. An urbanized area consists of a densely developed area that contains 50,000 or more people, while urban clusters consist of a densely developed area that has at least 2,500 people (which is the minimum population threshold; see Table 2) but fewer than 50,000 people. The US Census Bureau defines urban areas once a decade after the population totals for the decennial census have become available and classifies all territories and populations located within an urbanized area or urban cluster as urban and all areas

[6] Although, as the main text indicates, there are several definitions of cities in the US, the Census US Gazetteer files only provide coordinates for places, urban areas, and core-based statistical areas. Thus, the use of any other city definition would imply the use of non-official geographical coordinates.

outside as rural. Garmestani et al. (2005, 2008) use this definition of urban areas in previous empirical studies. Furthermore, urban areas are used as the cores on which core-based statistical areas are defined.

'Core-based statistical areas' (CBSAs) consist of the county, counties, or equivalent entities associated with at least one core (urbanized area or urban cluster) with a population of at least 10,000 (actually, the minimum population in the sample is 13,477), plus adjacent counties with a high degree of social and economic integration with the core, as measured through commuting ties with the counties associated with the core. Thus, CBSAs have economic meaning because they include the core area with a substantial population nucleus together with adjacent communities with a high degree of economic and social integration with that core, according to the US Census Bureau commuting data. CBSAs include both metropolitan and micropolitan statistical areas. Metropolitan statistical areas are CBSAs associated with at least one urbanized area that has a population of at least 50,000, while micropolitan statistical areas are CBSAs associated with at least one urban cluster that has a population of at least 10,000 but fewer than 50,000 people. Thus, our city definitions are nested; most places are included in urban areas, and most urban areas and places are located inside CBSAs.

Any of these spatial units have pros and cons. Most of the population of the country is included in the three samples (73.3% of the total US population is located in places, 81.9% lives in urban areas, and 93.9% is included in CBSAs). Places are the administratively defined cities (legal cities), and their boundaries make no economic sense, although some factors, such as human capital spillovers, are thought to operate at a very local level (Eeckhout, 2004). Urban areas represent urban agglomerations, making sure that rural locations are excluded. In addition, CBSAs are more natural economic units, covering huge areas that are meant to capture labour markets.

Nevertheless, Eeckhout (2004) demonstrates the statistical importance of considering the whole sample, suggesting the use of places (un-truncated data) rather than urban areas or CBSAs.

## 3. Methodology

We define four groups of city sizes depending on their population (5,000–25,000, 25,000–50,000, 50,000–100,000, and larger than 100,000 inhabitants) and then study how cities of a similar size are distributed in space. The criterion used to define the thresholds of the different groups is that the number of cities in each of the categories should be similar across the different definitions of cities, especially for the groups containing the large cities (50,000–100,000 and more than 100,000 inhabitants).[7] The sample size of each group is an important issue for two reasons. First, a similar number of cities within the groups simplifies comparisons across the different city definitions. Second, if the number of cities within one group was low, the confidence bands built based on counterfactuals (more on this below) would be wide and thus could bias our conclusions. Obviously, any choice of groups inevitably involves a certain amount of arbitrariness; we explored alternative cut-off points, although these were not very different from the groups finally chosen, and the qualitative results remained the same.[8]

We follow the methodology by Duranton and Overman (2005, 2008). This empirical procedure is used extensively to study the spatial distribution of firms, but, to

[7] We acknowledge that the definition of the city size groups is ad hoc. In particular, the largest group may seem broad, putting together cities like New York, with multiple millions of inhabitants, and cities like Richmond in California, with 103,701 inhabitants. From a system-of-cities perspective, New York hardly sits at the same hierarchical level as Richmond, but, if we consider a higher cut-off point for the top city size group, the number of cities within it would be low, and, as the main text explains, this is what we try to avoid. For instance, only 33 places, 80 urban areas, and 105 CBSAs have more than 500,000 inhabitants.

[8] Alternatively, we tried relative thresholds. The upper limits were 1, 2, 5, and ∞ times the average city size. The problem in this case was that some of these groups include a very low number of cities, especially in the case of urban areas and CBSAs. These results are available on request.

our knowledge, this is the first time that it is applied to analyse the spatial distribution of cities. This approach considers the distribution of bilateral distances between all the pairs of cities in each group. Then, we test whether the observed distribution of bilateral distances for each category of cities of similar sizes is significantly different from a randomly drawn set of bilateral distances. To be able to test this hypothesis, we build global confidence intervals around the expected distribution based on the simulated random draws. Cities of a particular size will be significantly localized or dispersed if their distribution of bilateral distances falls outside the global confidence intervals.

First, we calculate the bilateral distance between all the 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

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

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 900 miles. This threshold is approximately the median distance between all the pairs of cities (848 miles for places, 857 for urban areas, and 820 for CBSAs, to be precise). One fundamental difference between the analysis of the spatial distribution of firms and that of cities is the geographical scope; while most of the firms within an industry usually concentrate in a cluster of short distances, cities' distribution usually covers all the territory of the country and thus we must consider longer distances (up to 900 miles).

Second, to identify whether the location pattern of cities of a considered size is significantly different from randomness, we need to construct counterfactuals by first drawing locations from the overall cohort of cities and then calculating the set of bilateral distances. We consider that the set of all existing 'sites' ($S$), that is, 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, Los Angeles could be located in any other place in the US where a city exists. For each group of cities, we run 2,000 simulations. For each simulation, the density of distances between pairs of cities is calculated as if the same number of cities within the group was allocated across the set $S$ of all possible locations: all the cities for urban areas (3,592) and CBSAs (945) and a random sample of 15,000 in the case of places.[9] Sampling is performed without replacement. Thus, for any of the four groups of cities $A$ with $n$ cities, we generate our counterfactuals $\widetilde{A}_m$ for $m = 1,2,...,2000$ by sampling $n$ elements without replacement from $S$, so each simulation is equivalent to a random redistribution of cities across all the possible sites.

Finally, we compare the actual kernel density estimates with the simulated counterfactuals. To analyse the statistical significance of the localization pattern of cities compared with randomness, we construct global confidence bands using the simulated counterfactual distributions, following the methodology of Duranton and Overman (2005, 2008). $\overline{K}(d)$ denotes 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 900 miles. Similarly, the lower global confidence band $\underline{K}(d)$ is such that it is hit by 5% of the randomly generated K-densities that are not localized. Thus, when the estimated K-densities lie within the global confidence bands for distance $d$, $\overline{K}(d) > \hat{K}(d) > \underline{K}(d)$, the spatial location of cities is not significantly different from randomness. Deviations from randomness involve a localization pattern if graphically the estimated K-densities lie above the upper global confidence band for at least one distance $d$, that is, when

[9] This figure of 15,000 is more than half of the total number of places in the distribution (28,738). We cannot use the full sample of cities because of computational limitations.

$\hat{K}(d) > \bar{K}(d)$. Analogously, when graphically $\hat{K}(d) < \underline{K}(d)$ and the K-densities fall below the lower global confidence band for any distance $d$ , a dispersion pattern can be observed.

These global confidence bands may represent the idea that, contrary to clusters of firms, which are usually spatially distributed in a low number of locations, cities are distributed across the entire country. If we take into account the fact that, for any distance, the surface area considered is a circle ( $\pi r^2$ ), that is, a quadratic function of its radius $r$ (i.e. distance), the number of cities asymptotically will be a quadratic function of $r$ and, as the distance increases, the number of cities included in the circles naturally also increases. This means that any random redistribution of cities across space will lead, in most cases, to an increase of cities with distance; thus, the global confidence bands may be upward sloping, as our figures show. However, for long distances, the bands may be downward sloping because of physical space limitations (the distance increases but the area covered does not due to geographical boundaries such as the coastline) and the wide extensions of non-populated and rural areas in the US.

Furthermore, there are important dissimilarities between firms and cities that make the interpretation of the results different in the case of cities. In the case of cities, the localization or dispersion patterns supply information on the kind of hierarchy exhibited by US cities. The classic notion of an urban hierarchy consists of cities of different sizes, power, and influence within a given regional, national, or even broader territory (Pumain, 2006). From this notion, urban systems are composed of different tiers of cities ordered by size, and, if we are able to detect clusters of cities of similar sizes, this could provide evidence on the spatial patterns of urban hierarchies.

Let us consider an example. City pairs like New York City and Philadelphia, San Antonio and Austin, and San Francisco and Sacramento all fall into the fourth category (large cities), they all sit within 100 miles of each other, and each pair is part of a different urban system. This probably implies a high density for short distances between 0 and 100 miles, and this category would thus show localization between these distances. Now consider Los Angeles and its surrounding cities (e.g. Pasadena), almost 400 miles away from San Francisco and Sacramento. These cities are also close to each other, which still implies localization between 0 and 100 miles, but, at the same time, there would also be a peak of density for distances around 400 miles. A multiplicity of peaks in our K-densities by distance may thus indicate levels of cities of similar sizes in different urban hierarchies. However, the distances between cities in California and Texas or California and New York are so long that the possible peaks of density in these cases would not show up in our analysis, as they are more than 900 miles from each other.

## 4. Results

Figures 1, 2, and 3 show the results for the three definitions of cities considered, places, urban areas, and CBSAs, respectively. Regarding places (Figure 1), the estimated K-densities fall outside the bands in almost all cases, pointing to a clear non-random location pattern. We observe a high density for distances between zero and 100 miles for all city sizes, pointing to a localization pattern for cities of similar sizes in any category. Around 100 miles, the K-density crosses the confidence interval from above to below and thus localization turns into dispersion. Therefore, we can set 100 miles as the boundary of urban subsystems (on average), because city pairs within this distance are likely to be driven by localization and pairs above this distance are driven by dispersion and hence belong to different urban subsystems. For distances beyond 100 miles, different patterns of concentration emerge depending on the city size. For small and medium-sized cities (5,000–25,000 and 25,000–50,000 inhabitants), the densities fall within the bands (or they are very close to the lower band) until roughly 300 miles, indicating that the location of these cities is not significantly different from randomness. For longer distances (300 to 900 miles), the density of small and medium-sized cities continues to increase but at a slower pace than random location would involve, pointing to a soft dispersion pattern. Furthermore, we observe multiple peaks of density at different distances for all city sizes, supporting a spatial pattern of urban hierarchies in US places.

The geographical pattern of large cities (50,000–100,000 and more than 100,000 inhabitants) is different. Starting from a localization pattern (some examples were mentioned in the previous section: pairs like New York City and Philadelphia, San Antonio and Austin and San Francisco and Sacramento), from zero to 200 miles, the density decreases. This means that, when all cities are considered, some large cities are located close to each other but the rest are farther apart. This drop in density, illustrated in both Figure 1(c) and Figure 1(d), can be interpreted as the ‘agglomeration shadow’ of big cities: larger cities tend to be widely spaced. Although the densities recover the initial levels at a distance of around 300 miles, there is a clear dispersion pattern in the location of these cities for all distances longer than 200 miles.

If we consider urban areas rather than places, the results are quite different. Figure 2 shows that, for most of the size categories (5,000–25,000, 25,000–50,000, and 50,000–100,000 inhabitants), the densities increase with distance, but we cannot reject a spatial pattern different that is from randomness, as the K-densities fall within the global confidence bands for most of the distances (or they coincide with the lower band in the case of cities with populations between 25,000 and 50,000). Only for long distances does a dispersion pattern appear: 500 and 600 miles in Figures 2(b) and 2(c), respectively. For the largest cities (more than 100,000 inhabitants), Figure 2(d) shows that the increasing density with distance is below the bands from 200 miles, indicating a dispersion pattern too.

Finally, Figure 3 shows the results for the CBSAs. In this case, a localization pattern is only observed for small and medium-sized cities (Figures 3(a) and 3(b)) at long distances. For the size categories 50,000–100,000 and more than 100,000 inhabitants, we cannot reject a random spatial pattern, as the K-densities lie within the global confidence bands for almost all the distances (or they coincide with the upper band for short distances). This time, most of the pairs of highly populated places located close to each other might actually be within the same CBSA, so all that information is missing and it is hard to find any significant spatial patterns. Only for large cities (Figure 3(d)) does a weak dispersion pattern appear from 600 miles.

The lack of a clear spatial pattern for urban areas and CBSAs is not surprising; remember that these city definitions are nested and that most places are included in urban areas and most urban areas and places are located inside CBSAs. Urban areas include urbanized areas and urban clusters, and a minimum population threshold of 50,000 and 2,500 inhabitants, respectively, is imposed. At the same time, CBSAs' core is, at least, one urbanized area or urban cluster with a minimum population of 10,000 inhabitants. This means that, as we move from places to urban areas and from urban areas to CBSAs, the level of spatial aggregation increases. As Sánchez-Vidal et al. (2014) indicate, if the city definition requires a minimum population, most of the interactions between central and surrounding cities actually take place within these aggregate geographical units, so that information is missing. For instance, the number of cities in the size groups 25,000–50,000 and 50,000–100,000 clearly decreases when we move from places to urban areas (see Figures 1, 2, and 3). Thus, from this point of view, places are more appropriate than urban areas and CBSAs to analyse the spatial distribution of cities.

## 5. Conclusions

By using data from different definitions of US cities in 2010, we study the distribution of cities in space. K-densities, estimated using the methodology by Duranton and Overman (2005), allow us to identify different spatial patterns depending on the city size and city definition. The city definition is a key issue for any study dealing with population data; here, we consider three options: places, urban areas, and core-based statistical areas. For places and urban areas, we obtain some significant patterns, but for CBSAs, which are the spatial units covering wider geographical areas, the evidence is less conclusive. As most of the spatial interactions take place between nearby cities–at least, in terms of migration (Rauch, 2014) and commuting–it is not surprising that places, the lowest spatial unit considered, is the city definition exhibiting multiple and more pronounced spatial patterns.

Overall, focusing on places, we obtain a dispersion pattern regardless of city size for long distances. Starting from localization for distances between zero and 100 miles for all city sizes, 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. Furthermore, for large cities, the densities decrease from zero to 200 miles, indicating that, when all big cities are considered, some large cities are located close to each other but the rest are farther apart, pointing to different centres of urban systems.

These geographical patterns support a hierarchical system of US cities in which the central city of each subsystem is far away from others. We could set 100 miles as the boundary of urban subsystems (on average), because, according to our results, city pairs within this distance are likely to be driven by localization and pairs beyond this distance are driven by dispersion and hence belong to different urban subsystems. A more conservative threshold would be a distance between 200 and 300 miles, for which the density of large cities recovers the initial values, suggesting a kind of agglomeration shadow that large cities cast on nearby big cities (50,000–100,000 and more than 100,000 inhabitants) until a distance of 300 miles. The evidence regarding urban areas is less conclusive; only for cities with populations greater than 100,000 and distances longer than 200 miles do we obtain a significant dispersion pattern. In the case of CBSAs, almost no spatial pattern can be discerned. Only for the top large cities does a weak dispersion pattern appear from 600 miles.

## References

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

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

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

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

Desmet, K., and J. Rappaport. 2017. "The settlement of the United States, 1800 to 2000: The Long Transition to Gibrat's Law." *Journal of Urban Economics*, 98: 50–68.

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

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

Duranton, G., 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.

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

Fujita, M., P. Krugman, and A. J. Venables. 1999. *The spatial economy: cities, regions and international trade*. The MIT Press. Cambridge

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

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

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

Giesen, K., and J. Südekum. 2014. "City Age and City Size." *European Economic Review* 71: 193–208.

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

González-Val, R. 2010. "The evolution of the US city size distribution from a long-run perspective (1900–2000)." *Journal of Regional Science* 50(5): 952–972.

González-Val, R., 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.

Henderson, J.V. 1974. "The Sizes and Types of Cities." *American Economic Review* 64: 640–656.

Henderson, J.V. 1982a. "Systems of Cities in Closed and Open Economies." *Regional Science and Urban Economics* 12: 325–350.

Henderson, J.V. 1982b. "The Impact of Government Policies on Urban Concentration." *Journal of Urban Economics* 12: 280–303.

Henderson, J.V., and Y. Ioannides. 1981. "Aspects of Growth in a System of Cities." *Journal of Urban Economics* 10: 117–139.

Hochman, O. 1981. "Land Rents, Optimal Taxation, and Local Fiscal Independence in an Economy with Local Public Goods." *Journal of Public Economics* 59–85.

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

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

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

Krugman, P. 1991. “Increasing returns and economic geography.” *Journal of Political Economy* 99: 483–499.

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

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

Michaels, G., F. Rauch, and S. J. Redding. 2012. “Urbanization and Structural Transformation.” *The Quarterly Journal of Economics* 127(2): 535–586.

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

Pareto, V. 1896. *Cours d’économie politique*. Geneva: Droz

Partridge, M. D., D. S. Rickman, K. Ali, and M. R. Olfert. 2009. “Do New Economic Geography agglomeration shadows underlie current population dynamics across the urban hierarchy?” *Papers in Regional Science* 88(2): 445–466.

Pumain, D. 2005. “Alternative Explanations of Hierarchical Differentiation in Urban Systems.” In: *Hierarchy in Natural and Social Sciences*, Methodos Series, Vol. 3, D. Pumain ed., Springer: Dordrecht, 169–222.

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

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.

Sánchez-Vidal M., R. González-Val, and E. Viladecans-Marsal. 2014. “Sequential city growth in the US: Does age matter?” *Regional Science and Urban Economics* 44(1): 29–37.

Soo, K. T. 2005. “Zipf’s Law for cities: a cross-country investigation.” *Regional Science and Urban Economics* 35: 239–263.

Soo, K. T. 2012. “The size and growth of state populations in the United States.” *Economics Bulletin* 32(2): 1238–1249.

Upton, C. 1981. “An Equilibrium Model of City Sizes.” *Journal of Urban Economics* 10: 15–36.

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 the US in 2010**

A. Places

| Rank | City | Population | $S_{NY}/S$ | Bilateral distances (miles) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | New York, NY | 8,175,133 | 1.0 | . | | | | | | | | | |
| 2 | Los Angeles, CA | 3,792,621 | 2.2 | 2,459.0 | . | | | | | | | | |
| 3 | Chicago, IL | 2,695,598 | 3.0 | 717.7 | 1,748.8 | . | | | | | | | |
| 4 | Houston, TX | 2,099,451 | 3.9 | 1,419.3 | 1,378.8 | 937.3 | . | | | | | | |
| 5 | Philadelphia, PA | 1,526,006 | 5.4 | 77.5 | 2,399.3 | 666.5 | 1,343.0 | . | | | | | |
| 6 | Phoenix, AZ | 1,445,632 | 5.7 | 2,140.5 | 364.3 | 1,444.8 | 1,015.1 | 2,077.0 | . | | | | |
| 7 | San Antonio, TX | 1,327,407 | 6.2 | 1,583.0 | 1,207.7 | 1,047.1 | 189.7 | 1,507.9 | 846.7 | . | | | |
| 8 | San Diego, CA | 1,307,402 | 6.3 | 2,427.1 | 111.1 | 1,723.6 | 1,298.6 | 2,365.1 | 296.4 | 1,122.9 | . | | |
| 9 | Dallas, TX | 1,197,816 | 6.8 | 1,370.9 | 1,249.0 | 798.7 | 223.6 | 1,297.9 | 886.9 | 252.0 | 1,181.1 | . | |
| 10 | San Jose, CA | 945,942 | 8.6 | 2,550.4 | 296.4 | 1,832.7 | 1,602.1 | 2,497.2 | 604.8 | 1,443.7 | 407.4 | 1,446.4 | . |

B. Urban areas

| Rank | City | Population | $S_{NY}/S$ | Bilateral distances (miles) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | New York-Newark, NY-NJ-CT | 18,351,295 | 1.0 | . | | | | | | | | | |
| 2 | Los Angeles-Long Beach-Anaheim, CA | 12,150,996 | 1.5 | 2,442.1 | . | | | | | | | | |
| 3 | Chicago, IL-IN | 8,608,208 | 2.1 | 726.4 | 1,723.4 | . | | | | | | | |
| 4 | Miami, FL | 5,502,379 | 3.3 | 1,066.9 | 2,313.9 | 1,165.9 | . | | | | | | |
| 5 | Philadelphia, PA-NJ-DE-MD | 5,441,567 | 3.4 | 86.8 | 2,375.6 | 669.6 | 994.7 | . | | | | | |
| 6 | Dallas-Fort Worth-Arlington, TX | 5,121,892 | 3.6 | 1,380.0 | 1,219.5 | 796.9 | 1,104.5 | 1,298.8 | . | | | | |
| 7 | Houston, TX | 4,944,332 | 3.7 | 1,419.5 | 1,360.6 | 931.2 | 957.3 | 1,334.2 | 229.0 | . | | | |
| 8 | Washington, DC-VA-MD | 4,586,770 | 4.0 | 212.2 | 2,280.2 | 598.4 | 896.5 | 125.4 | 1,181.9 | 1,211.0 | . | | |
| 9 | Atlanta, GA | 4,515,419 | 4.1 | 741.7 | 1,927.8 | 586.0 | 582.5 | 654.9 | 732.7 | 706.4 | 529.6 | . | |
| 10 | Boston, MA-NH-RI | 4,181,019 | 4.4 | 185.7 | 2,581.7 | 858.4 | 1,231.8 | 272.4 | 1,552.6 | 1,600.6 | 397.7 | 927.2 | . |

Notes: Source: US Census 2010. $S_{NY}/S$ is the quotient between New York's population and city i's population. Bilateral distances calculated using the haversine distance measure based on Gazetteer coordinates.

**Table 1 (cont.). Bilateral physical distances between the 10 largest cities in the US in 2010**

C. Core based statistical areas

| Rank | City | Population | $S_{NY}/S$ | Bilateral distances (miles) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | New York-Newark-Jersey City, NY-NJ-PA | 19,567,410 | 1.0 | . | | | | | | | | | |
| 2 | Los Angeles-Long Beach-Anaheim, CA | 12,828,837 | 1.5 | 2,445.3 | . | | | | | | | | |
| 3 | Chicago-Naperville-Elgin, IL-IN-WI | 9,461,105 | 2.1 | 724.2 | 1,726.8 | . | | | | | | | |
| 4 | Dallas-Fort Worth-Arlington, TX | 6,426,214 | 3.0 | 1,389.2 | 1,220.5 | 800.3 | . | | | | | | |
| 5 | Philadelphia-Camden-Wilmington, PA-NJ-DE-MD | 5,965,343 | 3.3 | 101.5 | 2,376.0 | 666.8 | 1,299.0 | . | | | | | |
| 6 | Houston-The Woodlands-Sugar Land, TX | 5,920,416 | 3.3 | 1,427.7 | 1,369.0 | 933.8 | 234.1 | 1,330.8 | . | | | | |
| 7 | Washington-Arlington-Alexandria, DC-VA-MD-WV | 5,636,232 | 3.5 | 236.7 | 2,267.9 | 584.5 | 1,169.5 | 136.3 | 1,195.6 | . | | | |
| 8 | Miami-Fort Lauderdale-West Palm Beach, FL | 5,564,635 | 3.5 | 1,089.4 | 2,306.7 | 1,163.9 | 1,096.6 | 998.4 | 941.6 | 896.0 | . | | |
| 9 | Atlanta-Sandy Springs-Roswell, GA | 5,286,728 | 3.7 | 761.0 | 1,928.6 | 592.0 | 731.5 | 660.0 | 698.6 | 524.4 | 574.6 | . | |
| 10 | Boston-Cambridge-Newton, MA-NH | 4,552,402 | 4.3 | 186.0 | 2,587.2 | 860.5 | 1,563.5 | 287.3 | 1,610.0 | 422.6 | 1,254.1 | 947.0 | . |

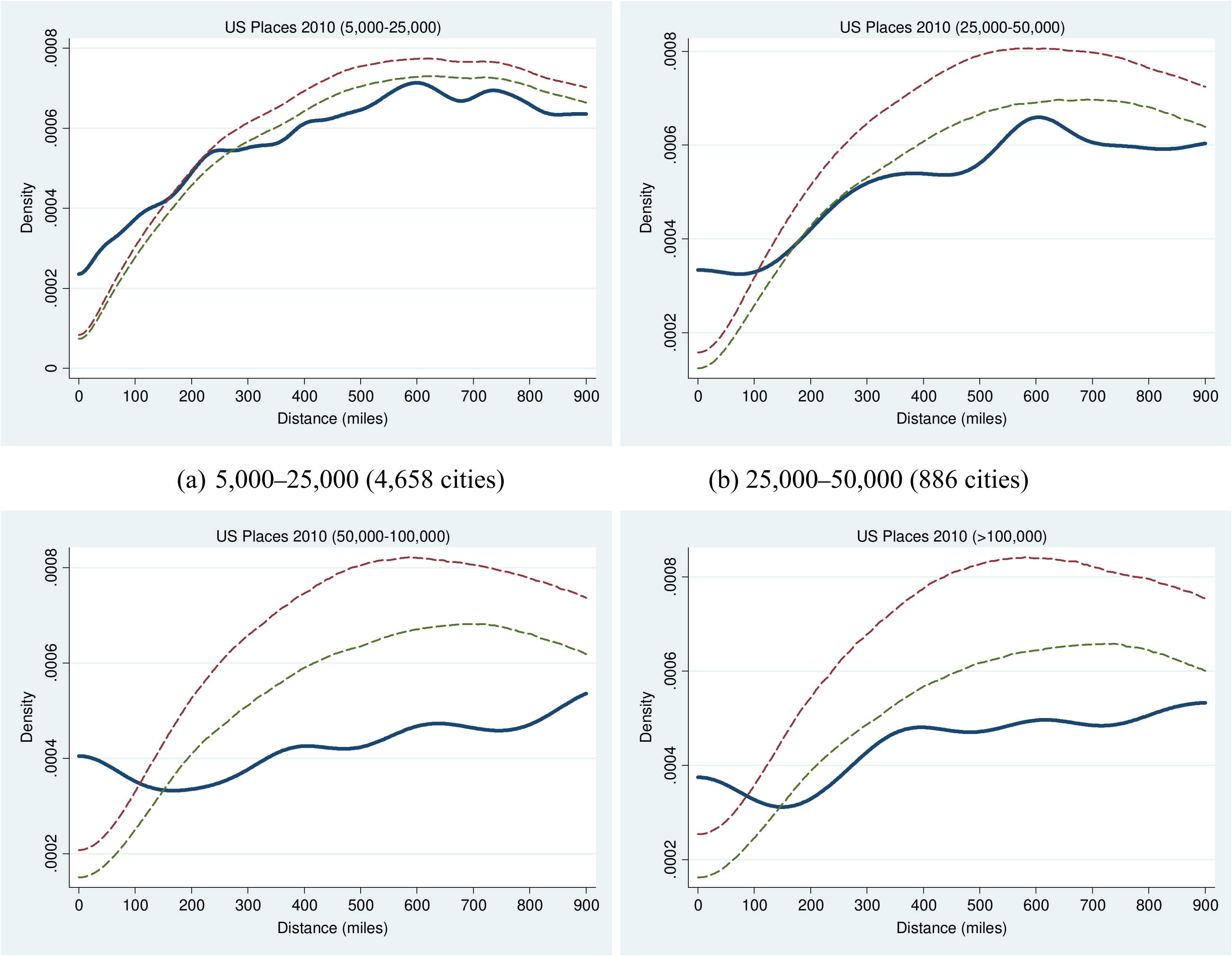
Notes: Source: US Census 2010. $S_{NY}/S$ is the quotient between New York's population and city i's population. Bilateral distances calculated using the haversine distance measure based on Gazetteer coordinates.

**Table 2. Descriptive statistics**

| City definition | Cities | Mean size | Standard deviation | Minimum | Maximum | Percentage of US population |
|---|---|---|---|---|---|---|
| Places | 28,738 | 7,880.2 | 66,192.9 | 1 | 8,175,133 | 73.3% |
| Urban areas | 3,592 | 70,363.7 | 495,447.5 | 2,500 | 18,351,295 | 81.9% |
| Core based statistical areas | 945 | 310,927.4 | 1,049,872.2 | 13,477 | 19,567,410 | 93.9% |

Source: US Census 2010.

**Figure 1. Spatial distribution of cities by size, US places in 2010**



(a) 5,000–25,000 (4,658 cities)

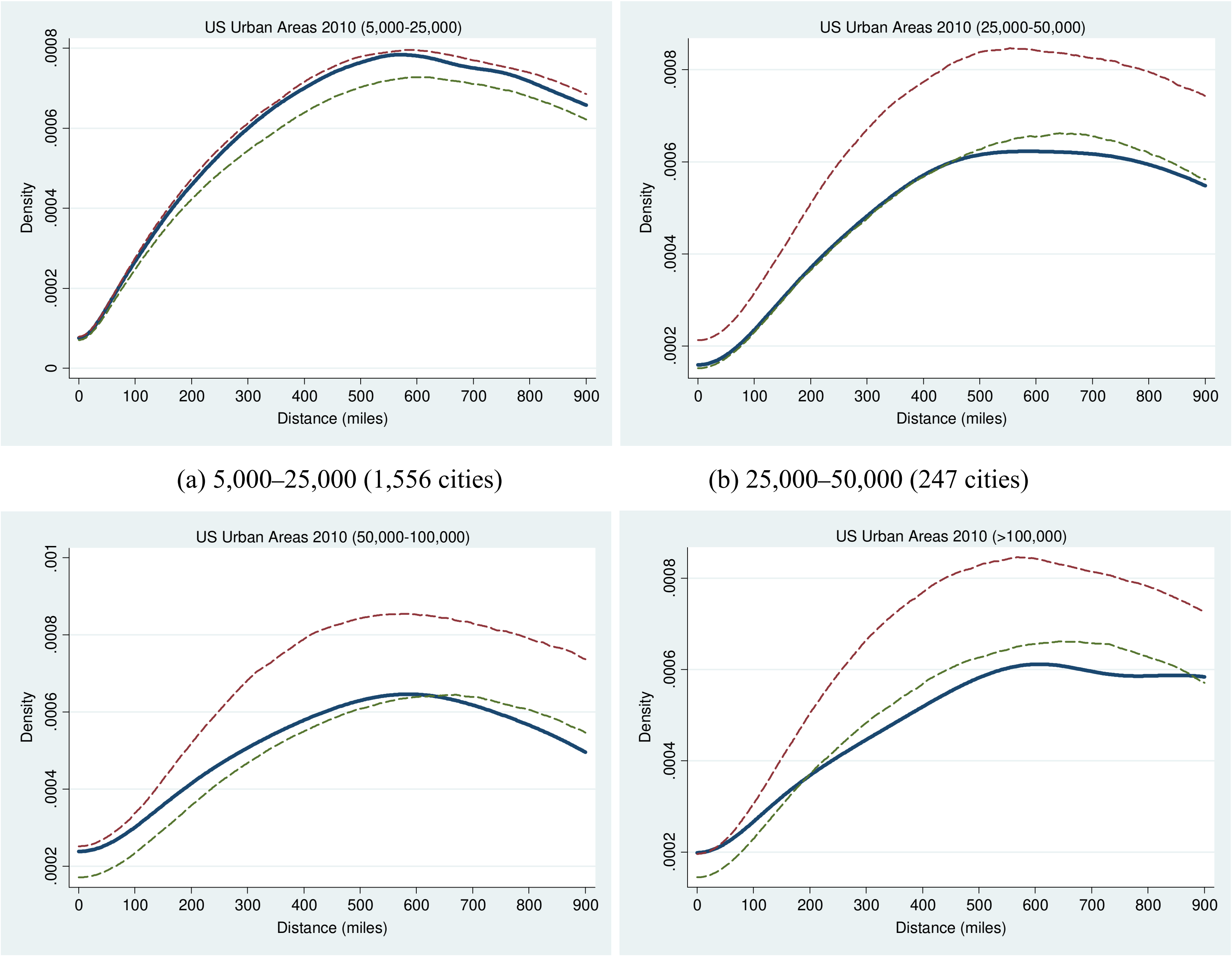
(b) 25,000–50,000 (886 cities)

(c) 50,000–100,000 (488 cities)

(d) >100,000 (280 cities)

Notes: Places include incorporated places and census designated places. K-densities are estimated using the methodology of Duranton and Overman (2005). Dashed lines represent the 95% global confidence bands, based on 2,000 simulations.

**Figure 2. Spatial distribution of cities by size, US urban areas in 2010**



(a) 5,000–25,000 (1,556 cities)

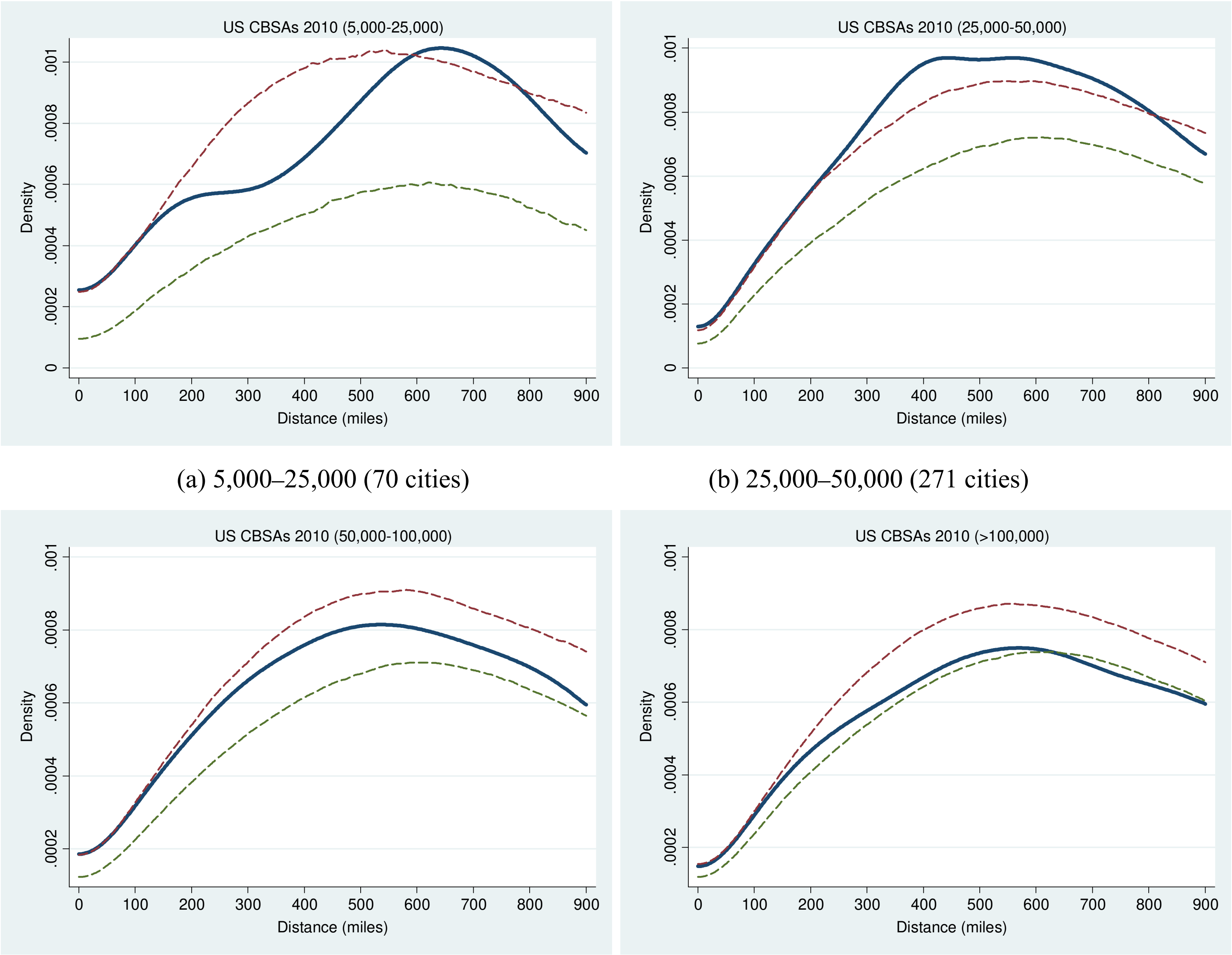
(b) 25,000–50,000 (247 cities)

(c) 50,000–100,000 (199 cities)

(d) >100,000 (298 cities)

Notes: Urban areas include urbanized areas and urban clusters. K-densities are estimated using the methodology of Duranton and Overman (2005). Dashed lines represent the 95% global confidence bands, based on 2,000 simulations.

**Figure 3. Spatial distribution of cities by size, US core-based statistical areas in 2010**



(a) 5,000–25,000 (70 cities)

(b) 25,000–50,000 (271 cities)

(c) 50,000–100,000 (214 cities)

(d) >100,000 (374 cities)

Notes: Core-based statistical areas include metropolitan and micropolitan statistical areas. K-densities are estimated using the methodology of Duranton and Overman (2005). Dashed lines represent the 95% global confidence bands, based on 2,000 simulations.