The size distribution of employment centers within the U.S. Metropolitan Areas

Daniel Arribas-Bel* Arturo Ramos†

Fernando Sanz Gracia‡

Abstract

This study tackles the description of the size distribution of urban employment centers or, in other words, the size of areas within cities with significantly high density of workers. Certainly, there exists a branch of urban economics that has paid substantial attention to urban employment centers, but the efforts have been focused on identification methodologies. In this paper, we build on such body of research and combine it with insights from the latest contributions in the sister subfield of city size distributions to push the agenda forward in terms of the understanding of these phenomena. We consider the 359 Metropolitan Statistical Areas (MSAs) in the United States in the year 2000 and reach three main conclusions: first, employment center sizes are more unevenly distributed than city sizes; second, the two functions that better describe city size distributions, namely the lognormal and the double Pareto-lognormal, also offer a good fit for the case of centers, particularly the last one; and third, several interesting statistically significant relationships (correlations) between variables related to centers and MSAs are deduced. Further experiments with a different technique of center identification suggest that the results are fairly robust to the method of choice.

*Department of Spatial Economics, VU University Amsterdam (Netherlands) darribas@feweb.vu.nl. This research was supported and funded by the Spanish Ministry of Education and Science (AP20063563 and ECO2009-09332). The authors would like to thank the Editor and three anonymous referees for their useful comments and suggestions; all remaining errors are the sole responsibility of the authors.

†Department of Economic Analysis, Universidad de Zaragoza (SPAIN) aramos@unizar.es

‡Department of Economic Analysis, Universidad de Zaragoza (SPAIN) fsanz@unizar.es

1 Introduction

Rigorous study of the size distribution of a measurable phenomenon has a long tradition in a wide range of disciplines such as physics, biology or geology. The social sciences have also been keen on this sort of research and, among all of them, economics has probably taken the lead. In particular, three have been the phenomena where it has focused. One, the analysis of the distribution of income among individuals; two, the distribution of firm size and its evolution over time; and three, the study of city size distributions, where size is taken as population. This work tackles the description of an entity that, to the best of our knowledge, has remained largely ignored by research: urban employment centers. Certainly, there exists a branch of urban economics that has paid substantial attention to urban employment centers, but the efforts have been rather focused on establishing an appropriate definition and the right technique to discern which areas are and which are not part of a center within an urban region. Although there is not a unique definition of employment center, there seems to be a common agreement on its general characteristics; as stated in McMillen and Smith (2003), an employment center is defined as *an area with significantly higher employment densities than surrounding areas.* Also, it *should be large enough to have a significant effect on the overall spatial structure of the urban area.* In this paper, the departing point is already the result of one of such definitions and the main purpose is to advance the agenda in terms of the understanding of these phenomena by analyzing their size distribution.

It is not necessary then to justify the study of city size distributions and its relevance for economists, geographers or regional scientists. At the heart of this interest is the fact that cities are probably the clearest manifestation of positive economies of agglomeration (see Fujita and Thisse, 2002). These are channeled into more than proportional increments of productivity (Ciccone and Hall, 1996) and are also associated with innovation processes, human capital accumulation (Moretti, 2004) and, ultimately, wealth creation. Although there exists certain consensus around the existence of the previous mechanisms[1], it is less clear to what extent the city is the true spatial scale at which they operate or whether it has been used as proxy, taking advantage of better data availability. The question that underlies the discussion is at which scale these agglomeration forces truly operate; or in other words, if we could choose the unit regardless of the availability of data, would we stick to urban regions or would we zoom in a bit more to try to spatially delineate the economic micro-foundations of agglomeration economies (Duranton and Puga, 2004)?

We consider that the spatial extent of such forces is better captured by intra-urban areas of high employment density, that is, by employment centers. This statement contains the initial hypothesis of our work, and we would like to justify it in more detail, motivating in that way our choice of urban employment

[1] Many authors agree about the positive effects of agglomeration through the existence of the so-called economies of agglomeration. However the consensus is not complete: see Beaudry and Schiffauerova (2009) and de Groot et al. (2009).

centers as the basic geographic unit of all our analysis. The key concept is that of agglomeration economies; it is clear that they occur in cities, but we wonder about why it is necessary to descend to a smaller unit. In their excellent survey article about urban spatial structure, Anas et al. (1998) express the following ideas:

> "[...] One class of agglomeration economies is *intrafirm* economies of scale and scope that *takes place at a single location.* Another class is positive technological and pecuniary externalities that arise between economic agents *in close spatial proximity* [...]"
>
> P. 1427, emphasis is ours

Evidently, according to the highlighted words in this quote, the urban employment centers capture the generation and diffusion of these agglomeration economies better than larger urban scales. Furthermore, great part of the theoretical literature about the internal structure of cities, their land use and the existence of polycentrism defines productive technology in such a way that the effect of one producer on the productivity of another is assumed to be a decreasing function of the distance between the two (Fujita and Ogawa, 1982; Lucas, 2001). This production externality causes that " employment at any site is more productive the higher is employment at neighboring sites" (Lucas and Rossi-Hansberg, 2002, p. 1469). Thus, employment centers are by definition areas of high density and concentration of workers; as a consequence, it is in those places where an important part of the economic activity takes place and where the agglomeration economies are developed with greater intensity. All this together makes the argument for urban employment centers as the object of study in this work.

At the same time, the location of urban employment centers as well as the potential existence of a geographical and/or functional hierarchy across their distribution defines the spatial structure of cities and has clear influence on important topics such as urban land prices or commuting times. In turn, the novelty of moving from population to employment centers has consequences over commuting. In fact, one of the key factors influencing the appearance of employment subcenters is the existence of high commuting costs. Subcenters have all, or almost all, of the advantages associated to the agglomeration, but are able to offer to the firms placed there lower land prices and the possibility of paying lower wages, since the commuting costs of workers reduce as well (McMillen and Smith, 2003). The policy relevance in this context is hard to exaggerate. Given these considerations, it seems natural to focus the analysis on this unit and question whether these entities are statistically distributed in similar ways as cities.

Against this background, we consider the 359 Metropolitan Statistical Areas (MSAs) in the United States in the year 2000. For each of them, we apply a center identification methodology that relies on Local Indicators of Spatial

Association (LISAs, Anselin, 1995) to detect areas of significantly high employment density within urban areas, finding 844 centers. Basing their size on the number of workers employed in each center, the main goal of this work is to take a first look at the distribution of the size of these employment centers.

The main findings of the analysis are three. In the first place, inequality in the size distribution increases as we move from population to employment of a city and is even more pronounced for the size of employment centers. Secondly, according to recent advances in urban economics, the functions that best describe the distribution of city sizes are the lognormal and the double Pareto-lognormal, particularly the last one. We find that these results also hold for the case of employment center size distribution. In third place, we deduce a set of statistically significant correlations between a few MSA and center variables: on average, the largest centers in absolute terms are the densest; the largest centers in relative terms tend to locate in smaller and poorer (less income per capita) MSAs; on the contrary, the relation between center density, the size and income per capita of their MSA is positive; finally, larger MSAs also have higher per capita income.

The remainder of the paper is organized as follows: Section 2 briefly surveys the main findings that the literature on city size distributions, as a reference point to this paper, has produced over several years; Section 3 describes the data employed as well as the intuition behind the center identification procedure; Section 4 constitutes the main contribution of the paper showing the results obtained regarding employment center size distributions. Section 5 describes the theoretical underpinnings of our empirical results. In Section 6, we perform sensitivity analysis in order to show that the results are not tied to the center methodology of choice; and Section 7 concludes with a few general remarks.

2 City Size Distribution: a brief survey of the main findings

The study of city size distributions has a long tradition in urban economics. To cite just a few examples, see Black and Henderson (2003), Ioannides and Overman (2003), Bosker et al. (2008), Giesen et al. (2010) and references therein. Historically, the Pareto distribution has generated more work than other distributions and greater acceptance. Its probability density function is

$$P(Size \geq x) = \frac{a}{x^b} \tag{1}$$

where a is a constant, $b > 0$ is the Pareto exponent and x is the number of inhabitants of each urban center. Considering the rank r (1 for the most populous center, 2 for the second, and so on) of the N cities we can obtain the well-known expression:

$$\ln r = c - b \ln x \tag{2}$$

where c is a constant. The expression relates the logarithm of rank with the logarithm of city size if the system follows a Pareto distribution. In the case of $b = 1$, we obtain the well-known Zipf's law or rank-size rule (see surveys on this subject by Cheshire, 1999, and Gabaix and Ioannides, 2004). If we consider the 135 largest MSAs in the USA in 1991, the Pareto exponent is 1.005 (Krugman, 1996, also shown in Gabaix, 1999), meaning that this law is fulfilled almost exactly. In summary, the Pareto distribution, and a particular case of it such as Zipf's law, are certainly useful for describing the size distribution of urban areas, especially the largest ones (upper tail).

However, Eeckhout (2004) proposes two ideas of interest here. One, that when all urban centers are taken, without any size restriction, Pareto's distribution breaks down and the best representation of the data is a lognormal function. Two, as a theoretical result: if the underlying distribution is lognormal, which generates a concave Zipf plot, the Pareto exponent decreases with sample size, meaning that a sample size can be found which verifies Zipf's law exactly. These first two contributions clearly show the importance of taking all cities, as doing otherwise can lead to skewed or spurious results.

Finally, in an influential paper, Giesen et al. (2010) use recent data for all urban centers (i.e. with no truncation point) in eight countries to show that, almost systematically, the distribution known as double Pareto-lognormal is the most suitable one from a statistical point of view, outperforming the lognormal. Similar results have been recently obtained in González-Val et al. (2013) for the 34 countries of the OECD. The double Pareto lognormal depends on four parameters, has a lognormal body and follows a power law (Pareto) in the upper and lower tails, although it is not possible to clearly delineate the lognormal body from the Pareto tails. It is also supported by a solid theoretical background (see Reed, 2002).

3 Center identification methodology and data

There is an ample literature that focuses on techniques to find and delineate urban employment centers. There has been a wide array of suggestions that vary in the way of the approach to the problem as well as in the complexity associated with the methodology. Giuliano and Small (1991) suggest to take ad-hoc cutoffs based on local knowledge and identify every area above them as centers and subcenters; Craig and Ng (2001) use spline quantiles to regress employment density on distance to the CBD to pick up areas significantly rising in density; in a series of papers (McMillen, 2001, McMillen and Smith, 2003, McMillen, 2004), McMillen applies locally weighted regression (LWR) in a two-step procedure. A recent strand of research (Paez et al., 2001, Baumont et al., 2004, Riguelle et al., 2007, Griffith and Wong, 2007) employs local indicators of spatial association (LISAs, Anselin, 1995) as the main methodology to find local clusters of high employment density. In this study, we join the latter strand and adopt LISAs as the main workhorse to identify and delineate urban employment centers.

Three main reasons guide and inform this decision: rather than relying on local knowledge to establish cutoffs for areas to be part of a center, it "offloads" the decision to statistical significance; it does not require the existence of any particular pattern (e.g. monocentric) for it to work, being flexible enough to accommodate both monocentric and more complicated polycentric landscapes; finally, because it does not require very sophisticated data inputs and may be easily automated, it is possible to apply it to a large dataset with many urban regions.

LISAs are a family of statistics designed to analyze spatial heterogeneity and to detect areas with significant deviations from the overall trend. Their use spans across many fields, from disparities in income growth (e.g. Rey and Montouri, 1999) to health (e.g., Mobley et al., 2006) or the examples from urban economics mentioned above for employment centers, to give a few examples. An advantage of these methods over other spatial analysis techniques recently popularized such as Ripley's K functions, for instance, is that unlike the latter, which were designed with point patterns in mind, they were especifically conceived for areal data, entities associated with polygons, like the ones we use in this study. The particular method we adopt is the local variant of the traditional Moran's I, which can be expressed as:

$$I_i = (\frac{z_i}{m_2}) \sum_j w_{ij} z_j, \tag{3}$$

where z_i is the standardized variable of interest for observation i, m_2 is its second moment (variance) and w_{ij} is the weight given to the spatial link between observations i and j, which relates to their spatial configuration. As an example, using a contiguity criterion as we will apply later on, $w_{ij} = 1$ if i and j share by any extent parts of their border, and $w_{ij} = 0$ otherwise. Significance may be calculated through simulation of the empirical distribution of I_i based on a permutation approach in which the neighbors of i are randomly shuffled a number of times to recalculate I_i and compare them to the actual value. In this context, a statistically significant positive (negative) I_i implies positive (negative) spatial autocorrelation in the vicinity of i: a situation in which i has a significantly high or low value and its neighbors equally have significantly high (low) or low (high) levels, respectively. Using auxiliary tools, it is possible to disentangle cases of positive spatial autocorrelation where a significant I_i is associated with a cluster of high values from cases of low values. Equally, it is possible to find out if negative correlation arises from a situation where observation i is a low value and its neighbors high or if it comes from a high value surrounded by low values. In particular, we use the so called "Moran scatterplot", a graphic in which z_i is displayed against $\sum_j w_{ij} z_j$, its so-called spatial lag. Cases of positive spatial autocorrelation appear on the upper-right and lower-left quadrants, while negative correlation locates in the remaining two other quadrants. More interestingly, spatial clusters of high values are found in the upper-right quadrant while clusters of low values locate in the lower-left.

"Spatial outliers" appear on the upper-left (lower-right) quadrant if it is a low (high) value surrounded by high (low) value neighbors.

Based on these techniques, we develop a simple and intuitive methodology to automatically detect urban employment centers. To do that, we first condense the notions of employment center from the literature cited above into the following, which becomes our operative definition:

> *An employment center is identified as a contiguous set of areas within an urban region in which each of them shows a spatial concentration of statistically significant high employment density at the 10% level.*

Next we *translate* each statement into a rule that needs to be met for an area to become a center of urban employment. Given an urban region, we begin with a set of LISA results for the employment density of its different areas (think of neighborhoods or transportation areas). This implies N statistics and its associated p-values. In this regard, a liberal 10% significance rule is used to ensure we capture any potential part of a center. The second step is to discard any observations that are not part of significant clusters of either high values surrounded by high values (HH) or high values surrounded by low values (HL). The intuition behind this is that we are looking for areas of significant high employment and these may be surrounded by other high employment areas, as in large centers, or by very low ones, as in smaller clusters located in residential areas, for example. At this point, we have in hand all the areas that form the employment centers of an urban region. The final step involves assigning them into unified clusters that can be treated as one in order to analyze their size. For this, a contiguity check is run to ensure all the candidates that are contiguous and hence part of only one center are counted as such, avoiding double counting issues. This yields the final number of centers as well as which areas are part of which. From here, it is straighforward to aggregate the size of every area that is part of a center and obtain its size in terms of workers.

Results of measuring city size are always somehow conditioned by the definition of city employed[2]. In the context of this paper however, because the analysis is focused on *employment center* size distribution, the role of the city definintion comes in to the extent it is the appropriate one to define and identify the centers. For this reason, we are inclined to adopt an economic sense of the city and choose the metropolitan region, which encapsulates the regional labour market and thus offers the appropriate boundary where the employment centers are to be found. This concept is operationalized by the Metropolitan Statistical Area (MSA), defined by the Office of Management and Budget, and constructed based on commuting flows. Employment centers in a MSA are identified at the Census Tract level, which offers a good balance between data availability and spatial resolution. The need for intra-city variation in employment in order to

[2]See Batty (2011) for an explanation of some of the challenges or González-Val (2012) for a recent overview of empirical issues, including the choice of the city definition.

be able to identify clusters of high values makes more exhaustive approaches, such as considering all human settlements as in Eeckhout (2004), although very appealing in principle, inconceivable in practice. The first panel in Table 1 offers basic descriptive statistics of the data used to build the employment centers.

Employment at the tract level is constructed by adding commuting inflows from the "Census 2000 Special Tabulation Product 64", which provides tract to tract commuting volume data[3]. A key element in computing LISA statistics is how space is formally represented, how the tracts are spatially organized. The choice of the spatial weights matrix (W) is very important and it is often treated with skepticism to the extent that it represents an arbitrary decision. We adopt the "queen contiguity" criterion, which considers neighbors all the surrounding polygons with which an observation shares either an edge or a vertex, and build one matrix for each MSA[4]. This is a common and easy to understand criterion that yields plausible and robust results[5]. This setup results in a database of 844 centers distributed across all the 359 MSAs existing in 2000[6]. The first row (LISA) in the lower panel of Table 1 shows some basic statistics about this sample while the second one offers information about a complementary dataset. In order to check the robustness of our results with respect to the methodology followed (the LISA approach), we have repeated the empirical application starting from an alternative method to define employment centers; in particular, we have replicated that of the classic and pioneer work of Giuliano and Small (1991) (GS hereafter). All this is described in detail in Section 6. Finally, income per capita and population, used in Section 4, as well as the geographical data comes from the Census Bureau, although we obtain it through the National Historical GIS (NHGIS, Minnesota Population Center, 2004).

4 Empirical results for employment centers

Up to this point we have briefly described the main findings about city size distribution (Section 2) as well as the procedure adopted for identifying employment centers and data (Section 3). Both sections prepare and contextualize the empirical analysis carried out, which constitutes the contents of this fourth section. This represents the core of the paper and it is subdivided in three subsections. In the first one we use a traditional tool from urban economics, namely the so-called *Zipf plots*. In the second one, we carry out statistical tests to find

[3] This product is the only way we have found to obtain employment estimates for a small area unit such as the Census tract for the entire US. 2000 figures were released in 2004 and, unfortunately, the 2010 numbers have not been published yet.

[4] All the spatial analysis tasks in this study were performed using the open source library PySAL (Rey and Anselin, 2007). For more information regarding this project, please visit http://pysal.org.

[5] In this regard, we replicated the analysis shown below with different specifications of W (including distance based and k-nearest neighbors) and found no significant changes in the final conclusions.

[6] An online map of the two center datasets used in this paper (LISA and GS, explained in Section 6) is available for interactive exploration at `http://bit.ly/S6HWma`

	N	Min.	Median	Average	Max.
MSAs	359	17,665	86,484	270,754.46	7,591,617
Tracts	52,498	4	951	1,851.52	150,532
LISA	844	280	6,885	28,400.55	1,863,000
GS	1,480	292	8,418	41,460	1,961,000

NOTE: An online map of the two center datasets (LISA and GS) is available for interactive exploration at `http://bit.ly/S6HWma`

Table 1: Employment figures for original sources and the two center databases used

out which distribution (lognormal or double Pareto-lognormal) offers a better fit for the distribution of employment center sizes. By using a simple methodology, the third one presents correlations among certain variables associated with employment centers or their respective MSA.

4.1 Zipf plots

Figure 1 shows traditional Zipf plots[7] in which the logarithm of rank (corrected by subtracting $\frac{1}{2}$, as suggested in Gabaix and Ibragimov, 2011) on the vertical axis is displayed against the logarithm of the number of workers in that center (Fig. 1, a), against the logarithm of MSA employment size (total number of workers in that MSA, Fig. 1, b) and against the logarithm of MSA population size (Fig. 1, c) on the horizontal axis. The estimation results can be found, respectively, in the three panels of Table 2.

There are important differences between both types of size distributions (employment centers and MSAs). In the first place, the degree of fit, or R^2, for employment centers is 0.8881, notably smaller than that for MSAs (0.9667 and 0.9716). In the second place, but more relevant, the estimated Pareto exponent for centers is surprisingly low, in particular 0.6381. It is well known that such a coefficient may be seen as an indicator of the degree of eveness of the distribution, in a way that the larger (smaller) the number, the more accentuated the equality (inequality) in the distribution. We can compare this value of the Pareto exponent of 0.6381 for employment centers to what it is usual in the broad literature of city size distributions (which also includes MSAs size distributions). To that end we use the exhaustive information provided by Nitsch (2005): according to the meta-analysis carried out in relation to the empirical performance of Zipf's law, the average Pareto coefficient is 1.09, well

[7] A clear antedecent of what is done in this subsection can be found in Anderson and Bogart (2001). In that work it is studied whether Zipf's law is fulfilled for the urban employment centers of four metropolitan areas of the US (Cleveland, Indianapolis, Portland and St. Louis), finding a supportive evidence for it in Cleveland and Indianapolis.

		Rk - 1/2
Employment center size	b	-0.6381***
	Std. Error	0.0078
	R^2	0.8881
MSA employment size	b	-0.8901***
	Std. Error	0.0088
	R^2	0.9667
MSA population size	b	-0.9368***
	Std. Error	0.0085
	R^2	0.9716

NOTE: "*" implies significant at the 10%; "**" at the 5%; and "***" at the 1% levels.

Table 2: Pareto exponents

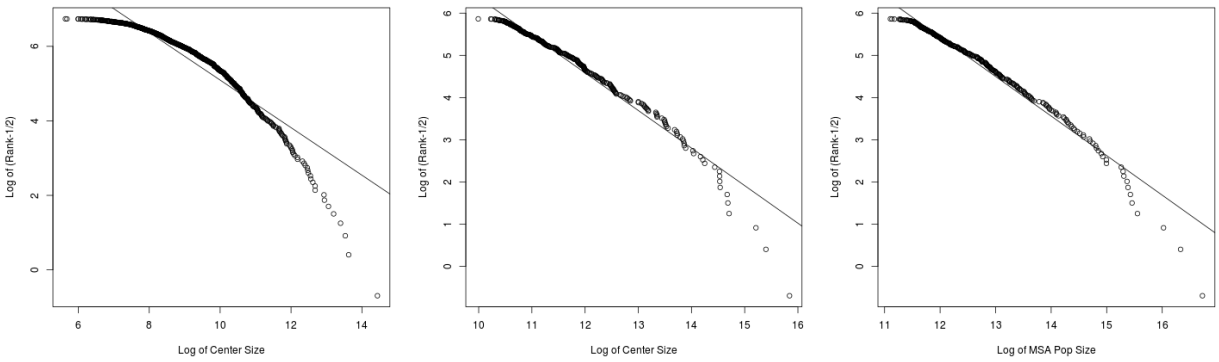
above the one found here, and the probability for such estimate to be between 0.8 and 1.2 goes up to 64%.

Although we are not trying to validate it, we can conclude that using the data from our sample, the closest case to Zipf's Law in this context occurs when we use population data at the MSA level; it is closely followed by the case of the MSA employment and the results appear very far when we analyze the distribution of employment center sizes. Furthermore, the maximum degree of eveness of the three distributions analyzed takes place for the case of the MSA populations ($b = 0.9368$); this decreases a little bit for the MSA employment ($b = 0.8901$); and experiences a sharp drop when we move on to the employment center level ($b = 0.6381$). This phenomenon of employment centers being distributed more unevenly than cities, requires further empirical evidence to be confirmed. Nevertheless, as a first approximation to the issue, this section clearly defines a fundamental difference between the two spatial units.

4.2 Statistical tests

A more rigorous approach to the quantification of the size distribution of employment centers involves formally testing for the best describing statistical distribution. In this section we follow recent insights in the literature of city size distributions and choose as candidates the two distributions that have been found to describe best city sizes (Giesen et al., 2010): the lognormal (LN) and the double Pareto-lognormal (DPLN)[8]. In order to assess the validity of the fits

[8]We have also tried the Pareto distribution in Eq. (1) for the entire samples of LISA (844 observations) and GS (1480 observations, explained in Section 6), finding that amongst the Pareto, lognormal and double Pareto lognormal, the first is the worst on these cases. Thus for the entire samples treated, the Pareto distribution does not seem to be adequate. However, if we take the top 50 or 100 observations of either the LISA or GS samples, the situation is the opposite: then the double Pareto lognormal cannot be estimated and the Pareto distribution



(a) Employment center size (b) MSA employment size (c) MSA population size

Figure 1: Zipf plots

obtained by using the mentioned distributions, we use two standard statistical tests.

The Kolmogorov-Smirnov (KS) test is a standard tool designed to discriminate between density functions. It is used, for example, in Giesen et al. (2010) to assess the adequacy of their studied distributions. The KS test has low power for small sample sizes, but very high power and extremely high precision for large sample sizes, see, e.g., Razali and Wah (2011). Another standard test is the Wilcoxon test. It has the null hypothesis that two independent samples come from populations with distributions with the same medians. The power of this test for large sample sizes is lower than that of KS test and does not tend to reject the null hypothesis in that case as often as the last one. We perform both tests comparing the estimated theoretical distributions with the empirical data. In order to do so, we estimate the LN and DPLN distributions fitted to the center size data by Maximum Likehood Estimation (MLE). For the sake of robustness, we carried out both Kolmogorov-Smirnov and Wilcoxon tests. Table 3 displays the results of the tests for the current dataset (LISA, left panel) and those relating to an additional one (GS, right panel) discussed in Section 6. P-values are displayed on the upper part of table, jointly with the tests' statistics. In spite of the previous discussion, none of the tests reject the null at the 5% level in the LISA approach and, consequently, we can say the distribution of urban centers with high employment density are well represented by both functions.

Which one is best? In order to answer the question we calculate information

performs better than the lognormal. This result is in accordance with the classical literature of Pareto distribution for the upper tail. More details are available from the authors upon request.

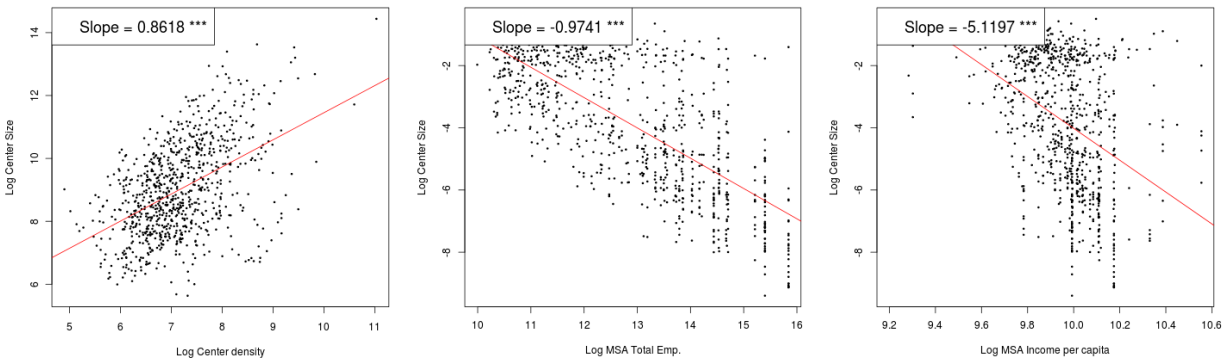
	LISA		GS	
	LN	DPLN	LN	DPLN
KS P-values	0.059	0.105	0.000	0.008
KS Statistic	0.047	0.043	0.058	0.045
Wilc. P-values	0.47	0.805	0.133	0.217
Wilc. z-statistic	-0.719	0.246	-1.503	1.236
Loglikelihood	-9,111.35	-9,101.86	-16,380.4	-16,354.1
AIC	18,227	18,212	32,765	32,716
BIC	18,236	18,231	32,775	32,737
Jeffrey's scale	Strong for DPLN		Strong for DPLN	

Table 3: Statistical tests for the two center databases

criteria, such as Akaike's information criterium (AIC), Bayesian information criterion (BIC) and Jeffrey's scale. Both the AIC and BIC incorporate the principle of parsimony in selection model as they penalize a higher number of parameters in the studied models (see, e.g., Burnham and Anderson, 1998). Thus, the model with lowest AIC and BIC is selected. In order to compare further the LN and the DPLN, since the double Pareto lognormal reduces to the lognormal taking appropriate limits, we can define the Bayes factor as $B \simeq \exp(S)$, where $S = \frac{1}{2}(BIC_{DPLN} - BIC_{LN})$. The value of B may be interpreted by using Jeffrey's scale (see Kass and Raftery, 1995). Results are reported in the remainder of Table 3. As can be seen from this table, the DPLN is systematically the preferred distribution according to all criteria, even when penalizing for having more parameters (four) than the LN (only two).

4.3 Some interesting correlations

So far, we have studied two relevant aspects in relation to employment centers: on the one hand, the degree of evenness of their size distribution measured by the magnitude of the Pareto exponent; on the other hand, the statistical distribution better describing the data. The main aim of this work, in a broader sense, is the empirical study of a geographical unit of reference hardly used in the literature: the urban employment centers. In this sense, since our data allow for it, we can delve deeper into the analysis and characterize these urban employment centers using simple instruments such as scatter plots and linear correlations. In this section we focus on a series of aspects related to the characteristics of centers and their relation with other socio-economic variables either at the center or at the MSA level.



(a) Density-Absolute size (b) MSA size-Relative size (c) MSA p/c Inc.-Relative size

Figure 2: Center log-size correlations

4.3.1 Center size

A first aspect to explore is the existence of patterns between two of the main center variables: size and density. In this context, we consider size as measured in absolute terms by the total number of employees working at a given center, and density as the ratio between size and the total area in squared kilometers of the center. Figure 2 (a) answers the question by plotting both variables in logarithms, a transformation used throughout the rest of the plots. It is straightforward to see the positive correlation between the two variables. Indeed, the coefficient of the fitted line is close to one and statistically significant at the 1% level, implying larger centers tend to be also denser, on average. As a natural extension, we try to find similar correlations between size of a center and employment size of the MSA where it is located in (total number of workers) and the average income per capita as dollars per resident in the MSA. The questions we are asking here are basically: *are larger centers located in larger MSAs?* and *Do richer MSAs tend to have larger employment centers?* In this case we cannot find significant relationships neither between center size and MSA size nor between center size and the MSA income per capita. Thus, the fact that a MSA is large or rich does not appear to be correlated with the absolute size of its employment centers.

It is also of relevance to consider the size of an employment center relative to the total employment of the MSA where it is located:

$$rs_c = \frac{s_c}{S_{MSA}} \tag{4}$$

where rs_c is the relative size of center c, s_c is its absolute size and S_{MSA}

is the total employment of the MSA. This measure gives us an idea about the agglomeration power of the employment center in the context of the MSA and how big of a role it plays as a cluster of workers in the local labor market. Figure 2 shows the relationship between the relative center size and the size of the MSA (b), and between the relative center size and the per capita income of the MSA (c), the only ones we have found statistically significative.

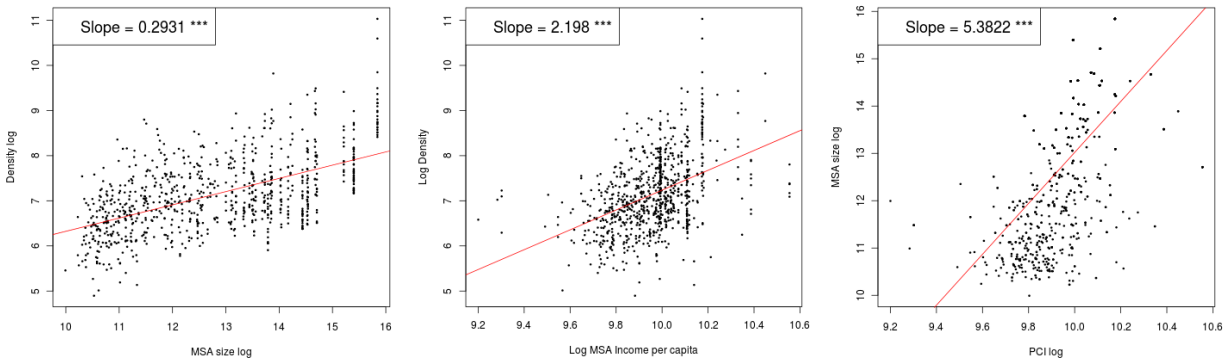
As we have seen, we cannot find any correlation between the *absolute* size of a center and the size of the MSA. However, as Figure 2 (b) shows, there is a strong pattern when we consider the *relative* size of the center. Indeed, this measure is negatively correlated with the size of the MSA[9]: smaller centers in relative terms tend to be located in larger MSAs. The explanation goes as follows: since the parameter for absolute center size and MSA size is indistinguishable from zero, when we confront relative size, that is absolute size divided by the MSA size, with the MSA size, the relation is logically inverse.

Figure 2 (c) depicts negative correlation between the relative size of the centers and the income per capita of the MSA. A priori, we did not anticipate either sign for the coefficient so the fact that it comes negative is greatly informing. Our preliminary interpretation of this finding is that it is likely that the smallest centers are located in large polycentric urban regions. The gap between relative size and per capita income is then closed by the fact that polycentric areas tend to be larger, and larger ones, richer.

4.3.2 Other socio-economic variables

We now turn into the analysis of other variables of interest in this context. First, we explore a bit more deeply the density of centers. Figure 3 displays the plots for the log of the center density against the log of the size of the MSA (a) and the log of the MSA's income per capita (b). Both of them show a significantly positive correlation, implying that denser centers tend to be located in larger and richer MSAs. Second, although the main focus of this paper lies in the analysis of employment centers, we consider also of relevance to show one more relationship in which none of the variables is measured at the center level. Figure 3 (c) displays the size of the MSA measured in terms of employment against the income per capita of the MSA. The correlation is clearly positive and highly significant.

[9] In this context, we also checked the relationship between relative size and density of the center. The coefficient is negative but barely significant at the 10% level. This aligns with the other more significant correlations: denser centers are located in large MSAs, which tend to be polycentric and thus to divide their workers between more than one center, resulting in a smaller proportion of the total employment allocated in one particular center, hence the smaller relative size.



(a) Density-MSA size (b) Density-P/c income (c) MSA size-P/c income

Figure 3: Other correlations

5 Theoretical underpinnings of the main empirical results

This is an eminently empirical work. However, our results can be related to theoretical principles and models of Urban Economics. These connections between results derived from the data and their theoretical counterparts are useful for understanding the true meaning of the first ones. The aim of this section is, therefore, to give a theoretical framework to the empirical results found in previous sections.

We have seen that the inequality in size distribution increases as we move from population to employment of a city and to employment centers. As Kopecky and Suen (2010) indicate, it is a stylized fact that employment is more concentrated than population, so a result like the one obtained should not be surprising. One possible theoretical explanation can be found in the model of Duranton and Puga (2005) where firms (and therefore the employment) have incentives, thanks to fundamentally technological progress in transport and telecommunications, to functionally separate their activity and become multi-plant firms. In this way it is obtained a small number of large business and headquarter centers (white-collar) which are separated from other centers, more abundant and smaller, in which the actual manufacturing production takes place (blue-collar).[10] This yields a distribution of the employment that is, by

[10] Certainly, the model of Duranton and Puga (2005) is more oriented to explain the differences across cities in an urban system than to understand the differences between a city center and its subcenters. Notwithstanding, Ota and Fujita (1993), in a previous article, are the first to apply this idea of front-office or front-unit and back-office or back-unit (in their terminology) that indeed generates divergence in the sizes of employment centers within a

definition, potentially more uneven than that of the population.

Another relevant result that, to the best of out knowledge, is new is the fact that bigger employment centers (in absolute terms) are, on average, denser. We interpret this result *à-la* New Economic Geography in terms of circular causation: larger centers, through the typical feedback mechanisms of demand (backward-linkage) or costs (forward-linkage) (see Krugman, 1991 or Fujita et al., 1999) result in higher densities, which in turn makes them larger.

The explanation of why smaller employment centers in relative terms are located, on average, in bigger and richer (with higher per capita income) MSAs combines theoretical and empirical arguments. We start from the hypothesis that the higher the number of employment centers in a city, the more likely is, ceteris paribus, that its relative size be smaller. In this context, the model of Fujita and Ogawa (1982) deduces, amongst other results, that the number of subcenters increases if the population and the income both increase. In turn, the model of Fujita and Krugman (1995) also concludes that an increase of population favors (sub)urbanization. In short, the theory predicts that bigger (with more population) and with increasing household income in time (richer) MSAs, tend to have a bigger number of employment centers and according to the previous hypothesis, their relative size is smaller. In addition, the empirical literature also confirms that increasing household income favors the appearance of more centers (see, e.g., the handbook of Bogart, 1998 or the survey article of Anas et al., 1998) and that the increase of the population makes that the number of centers grow.[11] It is possible that this evidence (bigger and richer MSAs have more reduced centers in relative terms) be favored by the stylized fact that bigger cities are more uneven in income ("big cities are characterized by big real inequality" Eeckhout et al., 2011), so that in these big cities the poor tend to locate in the center and the rich, much fewer people, go to the suburbs that are therefore smaller in relative and absolute terms. This last explanation, although plausible, is not more than a mere hypothesis which would require additional empirical work in order to confirm it.

We have also found out that, on average, denser centers are located on bigger MSAs with higher per capita income. The basic theoretical references which relate employment density with productivity and therefore with income level are Ciccone and Hall (1996) and Ciccone (2002). Both works start from a model based on two versions of spatial agglomeration: one based on spatial externalities and the other on non-tradable inputs produced with increasing returns. The two papers, the first with data of the US and the second with data of five european countries, also confirm from an empirical viewpoint the existence of agglomeration effects, in such a way that there is a direct relation between labor productivity and employment density, slightly greater in quantitative terms

city.

[11] McMillen and Smith (2003) manage to quantify the population thresholds necessary for a polycentric structure in 62 large US urban areas: 2.68 million inhabitants for the first subcenter and 6.74 for the second. In turn, Kopecky and Suen (2010) estimate that rising real income, falling prices of automobiles, and changes in the costs of travelling by both cars and public transportation can account for 60% of postwar US suburbanization.

for the case of the US. These agglomeration effects constitute also the basic foundation of the positive relation between the center density and city size.

Finally, the last correlation tells us that bigger MSAs are, on average, richer in terms of per capita income. If we accept the existence of the agglomeration economies, as the empirical evidence indicates (Puga, 2010), this is a directly derived outcome. Furthermore, it aligns well with established findings in the literature of urban economics (see Glaeser, 2011, for a very accessible overview) and urban complexity (Bettencourt et al., 2007, 2010).

6 An assessment on the robustness of the results

One might think that the results presented here are closely tied to the choice of the center identification methodology. Indeed, at the core of this analysis is a database on the size of centers, which ultimately depends on one specific technique. If this had unique features that produced very different results from those obtained using other of the suggested methods, the evidence presented in this work would be dubious. This issue is of special relevance because, as we have already mentioned, there is a notable lack of consensus about which is the most appropriate method to identify employment centers in urban areas. In order to mitigate this concern, we replicate the analysis using a database on the size of centers obtained applying a different center identification algorithm and compare it with the main conclusions derived from the previous sections.

One of the pioneering works on identification of urban centers and subcenters is Giuliano and Small (1991). As part of their objectives, they "*develop a method for systematically identifying employment subcenters*", which they apply to the Los Angeles region. The paper is one of the first of its kind and has inspired many researchers, which have widely cited it in other employment center-related investigations. The original definition they consider for a subcenter is:

> *(. . .) a continuous set of zones, each with density above some cutoff $\bar{D}$, that together have at least $\bar{E}$ total employment and for which all the immediately adjacent zones outside the subcenter have density below $\bar{D}$. (. . .)*
>
> Page 167

For their application to Los Angeles, they choose as cutoffs 10 employees per acre for $\bar{D}$ and 10,000 employees for $\bar{E}$. However, if we are to replicate this technique for all the 359 MSAs, we cannot keep a fixed threshold for every city, as it would not be appropriate for smaller ones (the entire remaining sample except for New York in fact). The approach we adopt to deal with this is to keep fixed the proportions and to relativize the cutoffs for each MSA: using our own data in 1990, we calculate what share of the total employment of Los Angeles' 10,000 workers represents and what is the ratio of 10 employees per

acre over the global employment density of the region. This leads to $\bar{D}$ being 4.76 times the global density, calculated as the total number of workers in the MSA over its area, and $\bar{E}$ equivalent to 0.215% of the total employment.

The second row in Table 1 shows a descriptive summary of our Giuliano and Small database of center sizes. Although it includes significantly more observations than the one built on the LISA method (1,480 against 844), it is interesting to note that the bounds remain fairly similar: the minimum size is 292 for GS and 280 for LISA, while the maximum values are 1,961,000 and 1,863,000, respectively. Median and average sizes are larger in the case of GS. Overall, it appears that the whole distribution for GS is somewhat shifted rightwards from that of the LISA results. Our interpretation of this phenomenon relates to the different mechanisms at work to select areas as employment centers. By comparing the actual value of the statistic to a simulated distribution of random values, obtained through a permutation approach, and using this measure of statistical significance to select significant employment centers, the LISA technique controls for the rare but actual possibility of high values in a random setting and "discounts" for it, resulting in a more conservative output. On the contrary, GS only uses a fixed cutoff per MSA, ignoring the fact that some of the areas above it might appear even if a completely random spatial process underlied the observed data. This results in a more liberal rule that identifies more areas as part of an employment center. It remains to be seen however whether this apparent shift is accompanied by a structural change of the underlying statistical distribution, that is, if all the additional areas that are picked by GS follow a particular pattern (e.g. as new smaller and independent centers, as part only of the largest centers, etc.) that is strong enough to modify the underlying statistical distribution, or if by contrast these are added in a proportionate fashion, maintaining the structure of the data (and the best describing distribution) similar in both cases. Although this might be a topic for future work, our results confirm that GS and LISA conclusions are very similar.

Equivalent results for GS about the statistical tests to evaluate the fit of the LN and DPLN are reported on the right hand side of Table 3. Overall, the main conclusions reached in Section 4.2 hold: both distributions appear to represent the data reasonably well and the DPLN arises as the preferred one, mirroring recent advances in the city size distributions literature. It is necessary to note that for GS, in this case the KS test rejects both of them. However, as we have mentioned before, this test over-rejects when the sample size is not small and, since GS has an even larger number of observations than the LISA sample, it is along these lines that we interpret this result. When the alternative less-powered Wilcoxon test is considered, we cannot reject any of the two null hypothesis. Although we do not report the figure, the corresponding Zipf plot sheds a b coefficient equal to 0.6022, similar the 0.6381 obtained with LISA. These findings allow us to answer in part the question raised in the previous paragraph: in fact, despite the differences between the two methods and the more liberal nature of GS at selecting areas as part of an employment center, the global structure of the data and their distribution remains fairly robust

	LISA	**GS**
Density - Abs. size	0.8618	0.6832
MSA size - Rel. size	-0.9741	-0.7869
MSA p/c Income - Rel. size	-5.1197	-3.5631
Density - MSA size	0.2931	0.4516
Density - p/c Income	2.198	3.0037
MSA size - p/c Income	5.3822	4.505

NOTE: All the coefficients are statistically significant at the 1% level.

Table 4: Robustness comparison

and well accommodated by the two distributions proposed, particularly by the DPLN.

The last exercise we carry out to verify the robustness of the results in Section 4 is to assess whether the correlations we report remain significant and with the same signs in the case of GS as in LISA. Table 4 presents a comparison of the slope coefficients we obtain when we replicate the plots from Figures 2 (upper panel in Table 4) and 3 (second part). For both methods, all of them are significant at the 1% level. Although there exist slight quantitative differences, in every case, the magnitudes of the values and their signs remain the same, indicating that the conclusions we reached in Section 4.3 are also robust to the choice of the center identification methodology.

7 Conclusions

The present study has considered the distribution of the size of urban employment centers. This merges two strands of literature, namely urban employment centers and city size distributions. Using a database on center size for the 359 US MSAs in 2000, our analysis mirrors recent advances in city size distributions, evaluating Zipf's law as well as trying to discern what is the best statistical distribution to describe the data.

A first conclusion from this work emerges from its conceptual approach. Indeed, the choice of the employment center as the relevant unit of analysis, a decision fairly novel, may be seen as one more step in a chain that begins with the nation (international economics) as the unit of reference and evolves refining the scale down to the region (regional science/economics) and the city (urban economics). In order to better delineate the spatial extent of agglomeration economies, we consider that a finer than urban resolution is more appropriate. This work does not aim to estimate or quantify such agglomeration economies, for which there already exists a range of literature. Instead, it tackles a much simpler task: to the extent these do have a purely spatial reflection, the study

represents an exploration (not a test) to improve our understanding of these entities.

Regarding the actual results of the analysis, there are three main conclusions. The first one is a divergence from the literature on city size distributions: employment center sizes are more unevenly distributed than city sizes. The second result represents a meeting point: the two functions that better describe city size distributions, namely the lognormal and the double Pareto-lognormal, also offer a good fit for the case of centers, particularly the latter one. Third, we have presented a set of interesting statistically significant relationships: larger centers in absolute terms tend to be denser; larger centers in relative terms tend to be found in smaller and poorer MSAs; on the contrary the correlation between center density and the size and income per capita of the host MSA is positive; finally, we also see how larger MSAs experience higher income per capita.

In order to assess the robustness of these results and to confirm the validity of the conclusions obtained, we have replicated the entire analysis using a database on center sizes obtained from a different but well established identification methodology, namely that presented in Giuliano and Small (1991). Although the two databases present some divergences in terms of the number of centers and their sizes, the distribution remains remarkably stable and all the conclusions originally drawn are also robust to the method of choice.

References

Anas, A., Arnott, R., and Small, K. (1998). Urban spatial structure. *Journal of Economic Literature*, 36:1426–1464.

Anderson, N. and Bogart, T. (2001). The structure of sprawl: identifying and characterizing employment centers in polycentric metropolitan areas. *American Journal of Economics and Sociology*, 60(1):147–169.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, 27(2):93–115.

Batty, M. (2011). Defining city size. *Environment and Planning B: Planning and Design*, 38:753–756.

Baumont, C., Ertur, C., and Le Gallo, J. (2004). Spatial analysis of employment and population density: the case of the agglomeration of Dijon 1999. *Geographical Analysis*, 36(2):146–177.

Beaudry, C. and Schiffauerova, A. (2009). Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy*, 38(2):318–337.

Bettencourt, L., Lobo, J., Helbing, D., Kühnert, C., and West, G. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306.

Bettencourt, L., Lobo, J., Strumsky, D., and West, G. (2010). Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. *PLoS ONE*, 5(11):e13541.

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

Bogart, W. (1998). *The economics of cities and suburbs*. New Jersey: Prentice Hall.

Bosker, M., Brakman, S., Garretsen, H., and Schramm, M. (2008). A century of shocks: the evolution of the german city size distribution 1925-1999. *Regional Science and Urban Economics*, 38(4):330–347.

Burnham, K. and Anderson, D. (1998). *Model selection and inference*. New York: Springer-Verlag.

Cheshire, P. (1999). Trends in sizes and structure of urban areas. In Cheshire, P. and Mills, E., editors, *Handbook of Regional and Urban Economics*, volume 3, chapter 35. Elsevier, Amsterdam.

Ciccone, A. (2002). Agglomeration effects in Europe. *European Economic Review*, 46(2):213–227.

Ciccone, A. and Hall, R. (1996). Productivity and the density of economic activity. *American Economic Review*, 86(1):54–70.

Craig, S. and Ng, P. (2001). Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *Journal of Urban Economics*, 49(1):100–120.

de Groot, H., Poot, J., and Smit, M. (2009). *Handbook of regional growth and development theories*, chapter Agglomeration, innovation and regional development: theoretical perspectives and meta-analysis, pages 256–281. Cheltenham: Edward Elgar.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In Henderson, V. and Thisse, J., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 48, pages 2063–2117. Elsevier.

Duranton, G. and Puga, D. (2005). From sectoral to functional urban specialization. *Journal of Urban Economics*, 57(2):343–370.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 94(5):1429–1451.

Eeckhout, J., Pinheiro, R., and Schmidheiny, K. (2011). Spatial sorting. *Mimeo*.

Fujita, M. and Krugman, P. (1995). When is the economy monocentric: von Thünen and Chamberlin unified. *Regional Science and Urban Economics*, 25:505–528.

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

Fujita, M. and Ogawa, H. (1982). Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics*, 12(2):161–196.

Fujita, M. and Thisse, J. (2002). *Economics of agglomeration: cities, industrial location, and regional growth*. Cambridge University Press.

Gabaix, X. (1999). Zipf's law and the growth of cities. *American Economic Review*, 89:129–132.

Gabaix, X. and Ibragimov, R. (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 Ioannides, Y. (2004). The evolution of city size distributions. In Henderson, V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 53, pages 2341–2378. Elsevier.

Giesen, K., Zimmermann, A., and Suedekum, J. (2010). The size distribution across all cities-double Pareto lognormal strikes. *Journal of Urban Economics*, 68(2):129–137.

Giuliano, G. and Small, K. (1991). Subcenters in the Los Angeles region. *Regional Science and Urban Economics*, (21):163–182.

Glaeser, E. (2011). *Triumph of the city: how our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin Press.

González-Val, R. (2012). Zipf's law: main issues in empirical work. *Région et Development*, (36):147–164.

González-Val, R., Ramos, A., Sanz, F., and Vera-Cabello, M. (2013). Size distribution for all cities: which one is best? *Papers in Regional Science*, in press.

Griffith, D. and Wong, D. (2007). Modeling population density across major US cities: a polycentric spatial regression approach. *Journal of Geographical Systems*, 9(1):53–75.

Ioannides, Y. and Overman, H. (2003). Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics*, 33(2):127–137.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kopecky, K. and Suen, R. (2010). A quantitative analysis of suburbanization and the diffusion of the automobile. *International Economic Review*, 51(4):1003–1037.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy*, 99(4):483–499.

Krugman, P. (1996). Confronting the mystery of urban hierarchy. *Journal of the Japanese and the International Economies*, 10:399–418.

Lucas, R. (2001). Externalities and cities. *Review of Economic Dynamics*, 4:245–274.

Lucas, R. and Rossi-Hansberg, E. (2002). On the internal structure of cities. *Econometrica*, 4:1445–1476.

McMillen, D. (2001). Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473.

McMillen, D. (2004). Employment densities, spatial autocorrelation, and subcenters in large metropolitan areas. *Journal of Regional Science*, 44(2):225–244.

McMillen, D. and Smith, S. (2003). The number of subcenters in large urban areas. *Journal of Urban Economics*, 53(3):321–338.

Minnesota Population Center (2004). National Historical Geographic Information System: Pre-release Version 0.1.50 Willey Hall, 225 19th Ave S, Minneapolis, MN 55455: University of Minnesota. http://www.nhgis.org.

Mobley, L., Root, E., Anselin, L., Lozano-Gracia, N., and Koschinsky, J. (2006). Spatial analysis of elderly access to primary care services. *International Journal of Health Geographics*, 5(19):5–19.

Moretti, E. (2004). Human capital externalities in cities. In Henderson, V. and Thisse, J., editors, *Handbook of Regional and Urban Economics*, volume 4, chapter 51, pages 2243–2291. Elsevier.

Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, 57(1):86–100.

Ota, M. and Fujita, M. (1993). Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Regional Science and Urban Economics*, 23(6):695–729.

Paez, A., Uchida, T., and Miyamoto, K. (2001). Spatial association and heterogeneity issues in land price models. *Urban Studies*, 38(9):1493–1508.

Puga, D. (2010). The magnitude and causes of agglomeration economies. *Journal of Regional Science*, 50(1):203–219.

Razali, N. and Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2:21–33.

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

Rey, S. and Anselin, L. (2007). Pysal, a Python library of spatial analytical methods. *The Review of Regional Studies*, 37(1):5–27.

Rey, S. and Montouri, B. (1999). US regional income convergence: a spatial econometric perspective. *Regional Studies*, 33(2):143–156.

Riguelle, F., Thomas, I., and Verhetsel, A. (2007). Measuring urban polycentrism: a european case study and its implications. *Journal of Economic Geography*, 7(2):193–215.