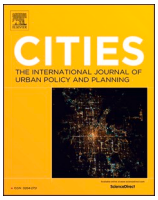
Cities 132 (2023) 104079

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities



Parametric, semiparametric and nonparametric models of urban growth

Rafael González-Val

Departamento de Análisis Económico, Universidad de Zaragoza, Facultad de Economía y Empresa, Gran Vía 2, 50005 Zaragoza, Spain
Institut d'Economia de Barcelona (IEB), Facultat d'Economia i Empresa, Universitat de Barcelona, John M. Keynes 1-11, 08034 Barcelona, Spain

ARTICLE INFO

JEL classification:
C12
C14
R11

Keywords:
Urban growth
Gibrat's law
Parametric models
Semiparametric models
Nonparametric models

ABSTRACT

This paper discusses parametric, nonparametric, and semiparametric models of urban growth. To illustrate differences across approaches, we test Gibrat's law in the long run, using the three methods and three different datasets: Spanish capital cities and regions (1900–2011, annual data) and US MSAs (1900–2000, decennial data). Our results reveal that the estimation of the relationship between growth and initial size can significantly vary across methods. We suggest and encourage the use of semiparametric methods in future research of urban growth.

1. Introduction

Urban growth models have a long tradition. Many theories have been proposed to try to explain why some cities (or regions) attract more people than others. Following Davis and Weinstein (2002), these theoretical explanations can be grouped into three main theories: the existence of increasing returns to scale, the importance of locational fundamentals, and the absence of both (random growth).

Each of these theories has different implications for the understanding of city growth. The existence of increasing returns suggests the presence of endogenous mechanisms in city growth that can lead to multiple equilibria, depending on initial conditions of income or population. Seminal articles discussing the endogenous character of city growth and proposing theoretical models of urban growth are Fujita (1976), Eaton and Eckstein (1997), and Black and Henderson (1999), among many others. In contrast, a body of literature argues that city growth is mainly driven by exogenous geographical characteristics (i.e., locational fundamentals). According to this theory, the presence of a natural harbour, a specific climate or access to the sea, among many other physical characteristics, can determine cities' populations. The third theory postulates that urban growth is a random variable. In that case, the growth process of cities tends to be multiplicative and independent of their initial size, a proposition that became known in urban economics as Gibrat's law.

What all these theories have in common is that when it comes to empirically testing them, most studies rely on linear growth models. In other words, although the theory allows for nonlinear behaviours of some variables, the empirical models usually do not. Some authors have tried to overcome this limitation by using polynomial specifications (e.g., Black & Henderson, 2003; Wheeler, 2003) or threshold regression models (Bosker et al., 2007; Davis & Weinstein, 2008; González-Val & Olmo, 2015).

In testing Gibrat's law, this strand of the literature has adopted nonparametric methods since the early 2000s (Eeckhout, 2004; Ioannides & Overman, 2003). Numerous empirical studies have tested its validity for city-size distributions, arriving at a majority consensus, although not absolute, that explains the growth of cities relatively well and tends to hold in the long term.

This paper examines the properties of these two traditional methods: linear and nonlinear growth regressions. Then, we propose using of a new methodology – the semiparametric method – which combines the better of the two traditional approaches. To illustrate the usefulness of this approach, we conduct an empirical examination of Gibrat's law in the long term using three datasets. Although we focus on Gibrat's law, the semiparametric model allows for the linear inclusion of city control variables and, thus, can be extended to other models testing increasing returns or locational fundamentals.

The remainder of the paper is organised as follows. In Section 2, we describe the different methodologies. Section 3 presents the population data used. Section 4 shows the main results, and Section 5 concludes.

E-mail address: rafaelg@unizar.es.

https://doi.org/10.1016/j.cities.2022.104079
Received 8 September 2021; Received in revised form 6 August 2022; Accepted 30 October 2022
Available online 11 November 2022

2. Methodology

Let Pop_{it} denote the population of city i at time t. We define the relative size of the ith city, s_{it}, as the quotient derived from dividing the city's population by the total population of the country, $s_{it} = Pop_{it}/Country\ population_t$.[1] Let $Growth_{it}$ be the logarithmic growth rate of the relative size of city i at time t, $Growth_{it} = \ln s_{it} - \ln s_{it-1}$. Then, we can define g_{it} as the normalised growth rate (subtracting the mean and dividing by the standard deviation).[2] This normalisation is a common practice in the literature (e.g., Desmet & Rappaport, 2017; Devadoss & Luckstead, 2015; Eeckhout, 2004; Giesen & Südekum, 2011; González-Val, 2010; Ioannides & Overman, 2003; Luckstead & Devadoss, 2014). In a long-term context of increasing populations over time, normalising with the contemporaneous average growth and standard deviation avoids some periods overpowering others on account of absolute population growth.[3] In the end, we are not interested in whether cities grow more or less in gross terms; normalised growth rates allow us to focus on whether cities' growth is higher or lower than the contemporary average growth. Moreover, as we will explain below, the normalisation makes the visual interpretation of the results easier.

A first way to test the relationship between growth and initial relative size is to run a simple linear growth regression (equivalent to a standard unconditional β-convergence regression):

$$g_{it} = \mu + \beta ln s_{it-1} + u_{it}, \tag{1}$$

where u_{it} is a random variable representing the random shocks that the growth rate may suffer, which we shall suppose are identically and independently distributed for all cities, with $E(u_{it}) = 0$ and $Var(u_{it}) = \sigma^2$ $\forall$ i, t. It is well established in the literature (Favaro & Pumain, 2011; Sutton, 1997) that results are strongly dependent on the assumptions made about u_{it} and the constant term because inadequate specifications can potentially invalidate the results. Under the pure form of Gibrat's law, city growth is a stochastic variable (Eeckhout, 2004; Gabaix, 1999), so note that the only explicative variable in Eq. (1) is initial city size.

Let us call this model a pure Gibrat's parametric regression. If $\beta = 0$ and u_{it} is an iid error term, Gibrat's law holds and we find that growth is independent of the initial size, with an average growth rate μ, whereas a significant β (positive or negative) would indicate a rejection of Gibrat's law. The literature has modified this pure Gibrat's parametric regression in two main ways. First, adding more explicative variables and controls. Second, changing the parametric model to a nonparametric one (but still *unconditional*). We examine both of these in detail.

The first way to modify the pure Gibrat's model in Eq. (1) is to keep the parametric specification but adding additional explicative variables besides the initial city size. In his survey of Gibrat's law in the firm size literature, Sutton (1997) explained how in the 1980s there was a rise of new literature with two main themes: econometric issues (e.g., the specification of an appropriate functional relationship or the problem of heteroscedasticity) and the criticism with the "only stochastic" models. Both themes can also justify the modification of the pure Gibrat's model for city sizes.

First, as explained above, inadequate specifications regarding the error and constant terms can lead to biased results (Favaro & Pumain, 2011), so authors deal with these problems by incorporating an autoregressive error term correction for serial correlation (Chesher, 1979) or by using dynamic panel data methods and unit-root tests (Pesaran, 2007) and, thus, modifying the baseline model of Eq. (1). These methods are based on the following equation (Favaro & Pumain, 2011):

$$g_{it} = \mu + \beta ln s_{it-1} + \phi_i + \delta_t + u_{it}, \tag{2}$$

where ϕ_i are city fixed effects and δ_t are time fixed effects. Note that the interpretation of this model is different from the pure Gibrat's model; in case $\beta = 0$ and the error term is iid, the conclusion would be that growth is independent of city size, but allowing for differences in the mean growth across cities, captured by the city fixed effects ϕ_i.

Second, although Gibrat's law is based on pure stochastic growth of cities, we might be interested in controlling for deterministic components of growth to separate both determinants and obtain more accurate estimates of the random part of growth. Skouras (2009) highlighted that there is substantial empirical evidence that some extremely persistent city characteristics are correlated with growth rates, including physical geographical attributes (i.e. first nature causes in the NEG terminology) such as coastal proximity, weather, or availability of natural resources, as well as human-made amenities (second nature causes), such as market potential, population education levels, and city infrastructure (Glaeser et al., 1995; Black & Henderson, 2003; Gabaix & Ioannides, 2004). Let x_{it}^T be the vector of these city-specific characteristics such as weather variables, location features (e.g., access to the sea), or any other city characteristic that may influence population growth; then, the model specification changes to:

$$g_{it} = \mu + \beta ln s_{it-1} + x_{it}^T \theta + \phi_i + \delta_t + u_{it}. \tag{3}$$

This growth equation is similar to those used in many studies on urban growth; see, for instance, Glaeser et al. (1995), Glaeser and Shapiro (2003), or Black and Henderson (2003). Moreover, the literature provides a theoretical economic background for this kind of linear growth model of city population growth; see the model of urban growth put forward by Glaeser et al. (1995) and further explicated by Glaeser and Shapiro (2003). Again, Gibrat's law is tested with the β coefficient, but the independence between growth and the initial population is conditional not only to the city and time effects but also to the non-random distributed vector of city-characteristics x_{it}.

However, one possible concern that remains in all these parametric models (Eqs. (1), (2) and (3)) is that the relationship (conditional or unconditional) between growth and size is restricted to be linear. As mentioned above, some authors (Eeckhout, 2004; Ioannides & Overman, 2003) suggested the use of nonparametric models to deal with this issue, while other authors (Wheeler, 2003) adopted an alternative solution, proposing parametric growth regressions but including polynomial specifications of the initial size:

$$g_{it} = \mu + \sum_{j=1}^{k} \beta_j (ln s_{it-1})^j + x_{it}^T \theta + \phi_i + \delta_t + u_{it}. \tag{4}$$

Eq. (4) allows for different polynomial functions; we consider values of k from 1 to 3 in order to capture any nonlinearity in the relationship between growth and initial size.[4] In the case of $k = 2$, we fit a quadratic function, while a cubic term ($k = 3$) is included in some specifications to control for possible increasing growth at high levels of relative size. However, although the β_j coefficients can help to detect nonlinearities in the relationship between both variables, a parametric growth regression is not the best way to address such nonlinear relationships.

Therefore, nonparametric growth regressions have become popular in the field. Among others, Ioannides and Overman (2004) have highlighted the advantages of the nonparametric approach over the standard parametric approach. Mainly, nonparametric methods do not impose any structure on underlying relationships that may be nonlinear and

[1] From a long-term temporal perspective of steady state distributions, it is necessary to use a relative measure of size (Gabaix & Ioannides, 2004), especially when we consider all the growth rates jointly in a pool.

[2] All the analysis was repeated using non-normalised growth rates and results were similar. These results are available from the author upon request.

[3] From the statistical point of view, Gabaix (1999) used Geometric Brownian Motion to describe city size processes but, even under the Geometric Brownian Motion assumption, the sample mean and standard deviation can be time varying (Ioannides & Overman, 2003, p. 133; Anderson & Ge, 2005, p. 769).

[4] Wheeler (2003) considered polynomials of initial size up to $k = 5$, but here for illustrative purposes we set at most $k = 3$.

may change over time (no need to restrict the relationship to being stationary); this is especially important when long periods are considered.

The kernel regression version of the pure Gibrat's model (Eq. (1)) consists of taking the following specification (Eeckhout, 2004; Giesen & Südekum, 2011; Ioannides & Overman, 2003):

$$g_i = m(lns_i) + \varepsilon_i, \tag{5}$$

where g_i is again the normalised growth rate at time t and $\ln s_i$ the logarithm of the ith city's initial relative size (at time $t-1$). The difference with the model in Eq. (1) is that here, instead of making assumptions about the functional relationship m, $m(\ln s)$ is estimated as a local mean around point $\ln s$ and is smoothed using a kernel, which is a symmetrical, weighting, and continuous function in $\ln s$. Thus, this nonparametric estimate lets growth vary with the initial population over the entire distribution. This specification has two relevant implications. First, note that temporal subscripts are omitted; this kernel regression can be run for a cross-section of growth rates (Eeckhout, 2004), as well as for a pool of growth rates from different periods (Giesen & Südekum, 2011; González-Val, 2010; Ioannides & Overman, 2003), and no temporal controls are included in any case. Note that, even when several periods are considered jointly in a pool, the panel dimension of the data is not exploited. Second, as in the model in Eq. (1), this kernel regression estimates the *unconditional* relationship between growth and size. Not only are temporal controls omitted but also fixed effects and any other control variables.

To estimate $m(\ln s)$, usually the Nadaraya-Watson method is typically used as it appears in Härdle (1990, Chapter 3), considering a kernel (in our case, an Epanechnikov) and a bandwidth h that determines the scale of the smoothing of the kernel density estimation.[5] The result of estimating Eq. (5) is a graph showing how estimated growth varies with size. As the growth rates are normalised, if the growth was independent of the initial population, the nonparametric estimate would be a straight line on the zero value and values different from zero would involve deviations from the mean.[6]

It is obvious that both approaches, parametric (Eqs. (1), (2), (3) and (4)) and nonparametric (Eq. (5)) growth regressions, have their drawbacks. To overcome the limitations of traditional approaches, Durlauf (2001) suggested the use of semiparametric methods. This alternative approach allows us to tackle the possible nonlinear effect of initial size on urban growth in a flexible way because the model is a mixture of both parametric and nonparametric growth regressions. For instance, the standard correlation index and the coefficients from parametric regressions give us only an aggregate average relationship between growth and size, and this relationship is restricted by the fact that it must remain unchanged throughout the entire distribution of city sizes. In contrast, the semiparametric estimate allows growth to vary with city size over the entire distribution (as in the nonparametric approach, Eq. (5)) but allows for the linear effects of other conditioning variables, such as city and time fixed effects and any other control variable we may include.

Therefore, the estimated semiparametric relationship between growth and size can be more accurate than that obtained using parametric models (because the relationship is allowed to vary over the distribution of sizes) and nonparametric models (because additional linear control variables are included). Nevertheless, to our knowledge, this empirical approach remains unexplored in the urban growth literature. In the related literature, Barrios and Strobl (2009), Lessmann (2014), and Díez-Minguela et al. (2020) have applied this methodology to the study of regional inequalities, and Basile (2008) used a semiparametric spatial Durbin model to analyse regional economic growth in Europe.

We perform the semiparametric analysis using Baltagi and Li's (2002) fixed-effects semiparametric regression estimator. Keeping the previous notation, a panel-data semiparametric model is:

$$g_{it} = m(lns_{it-1}) + x_{it}^T\theta + \phi_i + \delta_t + u_{it}, \tag{6}$$

where $m(.)$ is a smooth and continuous, possibly nonlinear, unknown function of $\ln s_{it-1}$. Like the parametric growth model (Eq. (3)), x_{it} can include any time-variant city characteristic at the city level. The model has a parametric ($x_{it}^T\theta + \phi_i + \delta_t$) and a nonparametric ($m(\ln s_{it-1})$) part. Baltagi and Li's approach is a two-step methodology. First, having estimated θ and δ, the fixed effects ϕ_i are fitted to estimate the error component residual:

$$\widehat{u}_{it} = g_{it} - x_{it}^T\widehat{\theta} - \widehat{\phi}_i - \widehat{\delta}_t = m(lns_{it-1}) + \varepsilon_{it}.$$

θ and δ are estimated after taking the first difference of the model in Eq. (6), applying a procedure similar to that whereby variables can be partialled out of an OLS regression. Then, the curve $m(.)$ can be fit by regressing $\widehat{u}_{it}$ on $\ln s_{it-1}$ by using some standard nonparametric regression estimator.[7] Note that in the nonparametric part of the model, the curve $m(.)$ is fitted to the linear prediction of the residuals, $\widehat{u}_{it}$, instead of to the gross growth rates (g_{it}). Therefore, contrary to Eq. (5), here we estimate the *conditional* independence between growth and initial size: how growth varies with city size but excluding time effects and the effect on the growth of observed and unobserved characteristics that can vary at the city level.

Finally, it can also be discussed whether models in Eqs. (4) and (6) really are so different, because if we increase the order of polynomials k to a large enough number, the estimates from Eq. (4) should be sufficiently close to those obtained from Eq. (6) (Newey, 1997). However, as Libois & Verardi (2013) explain, although the most efficient and unbiased estimator could be the fixed-effects estimator with the appropriate polynomial specification (model (4)), this specification is generally unknown. Therefore, as the true data-generating process is unknown, one might rely in high values of k arguing that a sufficiently flexible polynomial fit would be preferable to a semiparametric one. However, Libois & Verardi (2013) show that this is not the case, providing some empirical examples. Not only semi-parametric models provide a better fit for complex data-generating process than high order polynomial specifications; they can also help identify the relevant parametric form and help applied researchers avoid some trial and error.

3. Data

We aim to apply the different methods explained in the previous section to study urban growth from a long-term temporal perspective. The three models can also be applied to cross-sectional data (although the semiparametric model would require a slight change in the model specification[8]), but here we favour the use of panel data models. Panel data can model both the common and individual behaviours of groups, containing more information, more variability and more efficiency than pure time-series data or cross-sectional data.

Thus, we need long time series of city sizes. Although the nonparametric kernel regression model in Eq. (5) is estimated based on pooled data, Eqs. (1), (2), (3) and (4) and Eq. (6) are panel data models that require a high frequency in the time dimension. Therefore, ideally, our

[5] Although some authors set a fixed value of the bandwidth (e.g., $h = 0.5$ in Eeckhout (2004)), here the bandwidth is set using a rule-of-thumb.

[6] Here we only focus on the relationship between mean growth and initial size, although strictly speaking, random growth implies that the growth rate has a distribution function with both mean and variance independent of the initial size (Gabaix & Ioannides, 2004).

[7] The semiparametric models are estimated using the 'xtsemipar' Stata package. See Libois & Verardi (2013) for more details.

[8] A semiparametric model can be estimated for cross-sectional data using Robinson's (1988) approach.

data set should contain yearly observations to fully exploit the advantages of the models. However, long time series of year-by-year city populations are hard to obtain, and studies about city sizes are usually based on decennial census data.

As far as we know, only a few studies consider long-term annual city populations. Sharma (2003) used a sample of 100 Indian cities for the 1901–1991 period; Bosker et al. (2008) constructed a dataset of 62 West-German cities for the 1925–1999 period; Ronsse and Standaert (2017) constructed a dataset of 2681 Belgian municipalities for the 1880–1970 period; and González-Val and Silvestre (2020) built a dataset of 49 capital cities in Spain from 1900 to 2011. Sharma (2003), Bosker et al. (2008), and González-Val and Silvestre (2020) consider only the largest cities, while Ronsse and Standaert's (2017) sample includes all Belgian municipalities.

In this paper, we use González-Val and Silvestre's (2020) dataset. It consists of annual data for the 49 capital cities in Spain for the 1900–2011 period. Yearly data was estimated using information from the decennial censuses and historical reports of deaths and births. The geographical unit of reference is the municipality. Municipalities are the smallest spatial units (local governments); they are the administratively defined 'legal' cities. They are the lowest spatial subdivision in Spain; in terms of the current European Union's standard classification of European regions, municipalities are the LAU 2/NUTS 5 regions. Table 1 summarises the descriptive statistics of the samples at census dates.[9] As shown in Panel A, the number of capitals, 49, remained unchanged throughout the period considered; thus, the number of cities is fixed.[10] Each city is the administrative capital of one NUTS 3 region, and city population data account for changes in municipal boundaries over time.

For illustrative purposes, we also use the regional data provided in the González-Val and Silvestre (2020) dataset. For the same period, annual information about the 49 Spanish provinces (NUTS 3 regions) is available (Panel B in Table 1). The two provinces corresponding to the Canary Islands are joined together (due to the census reporting criteria during the first decades of the twentieth century), and the Spanish enclaves of Ceuta and Melilla in northern Africa are excluded. In comparison to capital cities, provinces comprise the country's total land area and, therefore, the entire population. Therefore, we expect to observe important differences in population growth between cities and regions.

Finally, as a robustness check, we also considerer decennial census data of the United States (US) Metropolitan Statistical Areas (MSAs) from 1900 to 2000 (Panel C in Table 1). The database is similar to that employed by Sánchez-Vidal et al. (2014) to test sequential city growth in the US. In line with Ioannides and Overman (2003), for the period from 1900 to 1950, data from Bogue's (1953) Standard Metropolitan Areas is used. These are based on the definition of Standard Metropolitan Areas for 1950, used to reconstruct the population for the period 1900 to 1940. The series is completed for the period 1950 to 2000 by taking data from the US Census Bureau. Although the Spanish samples have a fixed number of elements (49), in the US, the number of MSAs increases over time from 94 in 1900 to 274 in 2000.

As most studies use decennial census data, the US MSA dataset will help us to show that even if the frequency in the time dimension is low (ten years between censuses), if the period considered is long enough, the semiparametric approach can still provide more interesting results than parametric or nonparametric growth regressions.

Note that our data sets only include population data. Although the specification of some of the models includes a vector of observed city characteristics (x_{it}) that can influence city growth, in our empirical exercise we only include city and time fixed effects for two reasons. First, city time-invariant characteristics that are usually included in the urban growth models using dummy variables (such as climate, access to port or river, belonging to one of the main industrial belts in the case of US cities, etc.) cannot be included because of the city fixed effects. Nevertheless, this is not a problem because the effect of these time-invariant characteristics will be absorbed by the fixed effects. Second, in the case of time-variant characteristics, data availability is a problem because annual information (or decennial for the US case) of any economic or demographic variable for the whole twentieth century is difficult to find. Our results will show below that by only including city and time fixed effects, results significantly differ across methodologies, so for illustrative purposes and to allow comparisons between the results using the different data sets, additional control variables are omitted.

Table 1
Descriptive statistics.

A. Spanish capital cities. 1900–2011 (yearly data)

Year	Cities	Mean	Standard deviation	Minimum	Maximum
1900	49	65,664.8	110,879.8	7736	576,538
1910	49	72,817.9	123,931.8	8144	659,775
1920	49	85,821	155,040.9	8167	848,383
1930	49	107,635.7	210,762.5	10,588	1,137,943
1940	49	132,657.7	239,583.7	13,441	1,326,674
1950	49	155,533	290,974.9	17,297	1,645,215
1960	49	188,433.2	380,485	19,589	2,259,931
1970	49	241,565.1	499,939.5	23,030	3,146,071
1981	49	276,811.8	506,634.6	28,225	3,188,297
1991	49	287,243.6	489,878.5	31,068	3,084,673
2001	49	277,107.5	461,329.9	31,158	2,938,723
2011	49	300,962.6	499,881.1	35,660	3,198,645

B. Spanish regions (NUTS 3). 1900–2011 (yearly data)

Year	Provinces	Mean	Standard deviation	Minimum	Maximum
1900	49	379,477.7	183,176.7	96,385	1,054,541
1910	49	406,671.6	201,741.8	97,181	1,141,733
1920	49	434,757.2	237,776.2	98,668	1,349,282
1930	49	480,895.2	305,922.9	104,176	1,800,638
1940	49	528,121.9	348,645.2	112,876	1,931,875
1950	49	570,954.2	407,671.8	118,012	2,232,119
1960	49	621,034.7	526,897	138,934	2,877,966
1970	49	690,284	743,043.7	114,956	3,929,194
1981	49	767,683.6	909,249.6	98,803	4,726,986
1991	49	800,035.7	945,652.3	94,130	4,935,642
2001	49	830,805.2	1,007,302	90,717	5,423,384
2011	49	952,062.8	1,184,318	94,610	6,421,874

C. US MSAs. 1900–2000 (decennial data)

Year	Cities	Mean	Standard deviation	Minimum	Maximum
1900	94	312,626.7	682,220.7	52,577	5,597,481
1910	117	343,491.2	833,460.6	50,731	7,757,308
1920	124	408,940.5	987,907.7	54,664	9,386,725
1930	130	498,633	1,225,093	50,872	11,896,737
1940	133	529,359.6	1,294,329	51,782	12,760,857
1950	135	635,803.3	1,467,167	56,141	14,191,901
1960	220	555,444.7	1,399,082	51,616	15,620,434
1970	226	640,225.3	1,538,806	53,766	16,206,841
1980	274	643,653.5	1,618,349	58,460	18,905,705
1990	274	719,220.9	1,776,348	56,735	19,549,649
2000	274	819,336.3	1,974,707	57,813	21,199,865

Notes: Descriptive statistics in census years.

4. Results

In this section, we perform an analysis of Gibrat's law using the different models described above. Before starting, we must acknowledge

[9] For the Spanish data, the censuses were conducted in years ending in zero, between 1900 and 1970, and one, from 1981 onwards.

[10] For the Canary Islands, only one of the two capitals is included (Santa Cruz de Tenerife).

Table 2
Parametric growth regressions.

A. Spanish capital cities, 1900–2011 (yearly data)				
	(1)	(2)	(3)	(4)
ln(Relative size)	−0.019	−0.361	−3.781***	−3.909**
	(0.038)	(0.237)	(0.643)	(1.645)
ln(Relative size) ^{2}			−0.296***	−0.321
			(0.060)	(0.371)
ln(Relative size) ^{3}				−0.002
				(0.025)
City fixed effects	N	Y	Y	Y
Time fixed effects	N	Y	Y	Y
Observations	5439	5439	5439	5439
R^2	0.000	0.067	0.086	0.086

B. Spanish regions (NUTS 3), 1900–2011 (yearly data)				
	(1)	(2)	(3)	(4)
ln(Relative size)	0.635***	0.142	−0.943	−11.982***
	(0.086)	(0.268)	(1.216)	(3.648)
ln(Relative size) ^{2}			−0.125	−2.924***
			(0.136)	(0.975)
ln(Relative size) ^{3}				−0.224***
				(0.080)
Region fixed effects	N	Y	Y	Y
Time fixed effects	N	Y	Y	Y
Observations	5439	5439	5439	5439
R^2	0.193	0.394	0.396	0.412

C. US MSAs, 1900–2000 (decennial data)				
	(1)	(2)	(3)	(4)
ln(Relative size)	−0.029	−0.563***	−0.833	−8.069***
	(0.036)	(0.132)	(0.971)	(2.357)
ln(Relative size) ^{2}			−0.022	−1.319***
			(0.078)	(0.433)
ln(Relative size) ^{3}				−0.075***
				(0.026)
City fixed effects	N	Y	Y	Y
Time fixed effects	N	Y	Y	Y
Observations	1727	1727	1727	1727
R^2	0.001	0.578	0.578	0.583

Notes: All models include a constant. Coefficient (robust standard errors). Standard errors clustered by city/region. Significant at the *10 %, **5 %, ***1 % level.

that this is not a model selection exercise; the true data generating process is unknown, not all the models are nested, estimation methods are different across models, and standard information criteria cannot be applied.[11] We aim to illustrate how semiparametric methods can provide new insights into the workings of urban growth compared to the other empirical models (both parametric and nonparametric) used by authors to date.

Table 2 shows the results of the OLS estimation of the different versions of the parametric growth model. The first column reports the pure Gibrat's model in Eq. (1), a simple bivariate regression, finding a negative but not significant impact of initial relative size on the growth of Spanish and US cities (Panels A and C) and a positive and significant effect of size on growth in the case of the Spanish regions (Panel B). Column (1) in Table 2 is equivalent to the parametric growth regression estimated by Eaton and Eckstein (1997), and Fig. 1(a), (c), and (e) shows these estimated lines, along with the data points. Apparently, these initial unconditional linear regressions are not far from the real behaviour of the data; the cloud of dots is around the zero value for the Spanish and US cities, while in the case of regions, the increasing relationship seems obvious.

However, the inclusion of city/region and time fixed effects changes the results. In column (2), we show the results of the estimation of the model in Eq. (2); note that while the estimated coefficient for the Spanish cities (Panel A) remains negative and not significant, that of the regions' sample (Panel B) has now changed to not significant, and the effect of initial size on growth is significant and negative for the US cities. Fig. 1(b), (d), and (f) represents these estimates. Again, the slope of the fitted line is the estimated coefficient for the initial size as shown in column (2) and represents the linear relationship between growth and size after keeping all other variables (namely, the city/region and time fixed effects) constant.

In columns (3) and (4), the results of the estimation of the polynomial specifications (Eq. (4)) are displayed. In column (3), we include both the initial relative size and its square term to capture any nonlinearity in its relationship to population growth. The estimated coefficients are significant only in the case of the Spanish cities (Panel A), whereas, for the Spanish regions and US MSAs (Panels B and C, respectively), neither of the two coefficients is significant. For the Spanish cities, we found that both initial size and its square are negative and significant, supporting a quadratic function: a decreasing concave relationship between growth and city size.

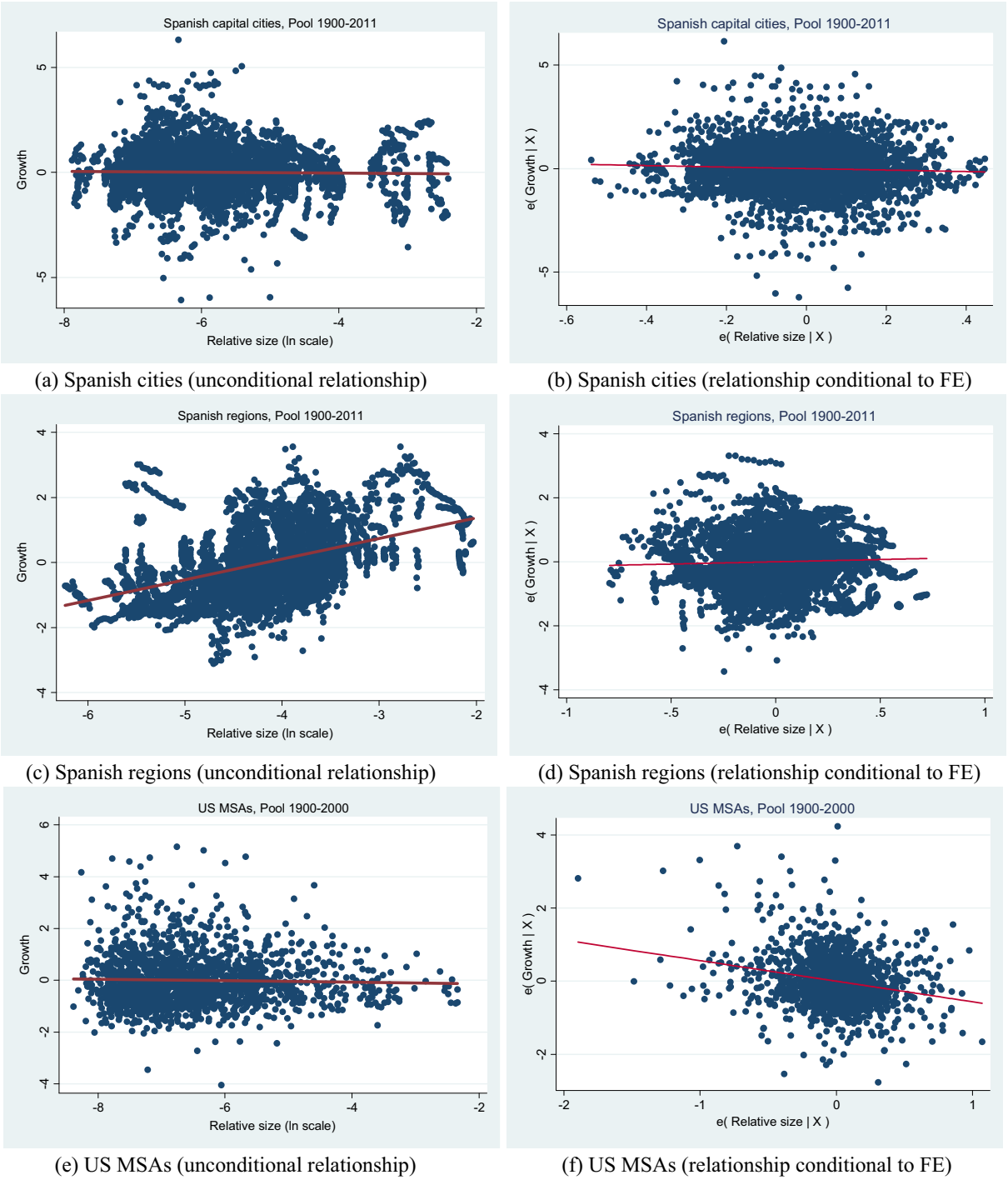
Nevertheless, when we consider a third-degree polynomial function (column 4), we obtain robust evidence of a nonlinear relationship between growth and size. For the Spanish regions (Panel B) and US cities (Panel C), the three coefficients are negative and significant, while in the case of the Spanish capital cities, the only coefficient significant (and negative) is that of the initial size; both the square and the cubic term are not significant.

Overall, the results from these regressions do not seem quite robust since there are important changes in the significance of the parameters depending on whether fixed effects or any polynomial function is included in the specification. Why? A look at the low values of R^2 (especially for the Spanish cities, Panel A) suggests that we are probably omitting important variables that can influence growth. As mentioned above, we choose to keep the specifications as simple as possible (the vector x_{it}^T of city characteristics is not included in any case) to allow comparisons in the results across the different datasets.[12] However, changes in the significance of the parameters of the different polynomial functions estimated tell another story: they indicate the presence of some kind of nonlinearity in the three cases that a linear regression is not able to properly capture.

Therefore, we next estimate kernel growth regressions (Eq. (5)) using a pool of all growth-initial size pairs. As mentioned above, nowadays, this is the common approach in this literature (e.g., Eeckhout, 2004; Giesen & Südekum, 2011; González-Val, 2010; Ioannides & Overman, 2003; Luckstead & Devadoss, 2014). Fig. 2 shows the results for the three samples. For each dataset, we show two sets of graphs. The left graphs (Fig. 2(a), (c), and (e)) display the fitted curve and the data points to allow comparisons with the corresponding panels in Fig. 1(a), (c), and (e) (the y-axis scale is the same as that in Fig. 1), while the right graphs (Fig. 2(b), (d), and (f)) zoom in to values of growth close to zero to highlight nonlinear patterns. The estimated nonlinear relationship is different in each case; we obtain an inverted U-shaped curve for the Spanish capitals (Fig. 2(b)), a clearly increasing relationship between size and growth for the Spanish regions (Fig. 2(d) also shows a small decrease in growth for the largest top-populated provinces) and random growth for the US MSAs (Fig. 2(f), with an estimated growth not different from zero for most of the distribution of city sizes (only the largest MSAs show significant negative growth). Some of these patterns do not seem to match well with the clouds of dots in the graphs in Fig. 2

[11] The fit provided by parametric and nonparametric models can be compared, but using non-standard methods. Nevertheless, a model selection analysis is beyond the scope of this paper.

[12] Furthermore, in the Spanish case, it is not easy to find annual city time-variant variables for the whole twentieth century.



(a) Spanish cities (unconditional relationship)

(b) Spanish cities (relationship conditional to FE)

(c) Spanish regions (unconditional relationship)

(d) Spanish regions (relationship conditional to FE)

(e) US MSAs (unconditional relationship)

(f) US MSAs (relationship conditional to FE)

Fig. 1. Growth versus initial relative size.

(a), (c), and (e), but note that in Fig. 2(b), (d), and (f), the y-axis scale is more reduced, and take into account that the clouds contain 5439 dots for the Spanish cities/regions and 1727 observations for the US case. Thus, many dots are concentrated in a very narrow range of values, making the visual inspection of the graphs, including data points, difficult.

In summary, the standard nonparametric method leads us to conclude that Gibrat's law is rejected for both Spanish cities and regions (we observe a convergence pattern in cities, while regions exhibit clear divergence), but random growth holds for the US MSAs (except for the largest cities). However, as explained above, these kernel regressions estimate the *unconditional* relationship for a pool of all growth-initial size pairs, so omitted variable bias could be present. Moreover, we learned from the estimated linear models in Table 2 and from Fig. 1(b), (d), and (f) that the inclusion of fixed effects significantly altered the results.

City fixed effects are relevant because they represent the individual city-specific growth, thus allowing for heterogeneity in growth rates across cities, and time fixed effects are even more important because they capture the influence of some temporal events on growth, such as pandemics (e.g., Spanish flu) or wars (e.g., Spanish Civil War, WWI and WWII). Moreover, the mainstream in the literature argues that random growth (or Gibrat's law) corresponds to the steady state (a long-run average), but to reach that situation, temporal episodes of different

(a) Spanish cities (b) Spanish cities (zoom)

(c) Spanish regions (d) Spanish regions (zoom)

(e) US MSAs (f) US MSAs (zoom)

Fig. 2. Nonparametric estimates of growth.

growth patterns across some cities are possible: '*the casual impression of the authors is that in some decades, large cities grow faster than small cities, but in other decades, small cities grow faster*' (Gabaix & Ioannides, 2004, p. 2353). These temporal trends are also captured by the time fixed effects.

The semiparametric approach allows us to include the fixed effects and, at the same time, perform a nonparametric estimate of the relationship between growth and size. The model in Eq. (6) has two parts: the linear part of the model, which in our case includes only city/region and time fixed effects, and the nonparametric part. Thus, the model has two outcomes: a table with the linear estimates (not shown in our case because it only includes the fixed effects) and a graph displaying the nonlinear relationship between the linear fitted residuals of growth and initial relative size.

Fig. 3 shows the semiparametric estimates of growth. Recall that now the y-axis variable is the linear prediction of the residuals ($\widehat{u}_{it}$) in Eq. (6) instead of the gross growth rates. In other words, we filter growth rates, excluding all variation due to city/region and time fixed effects. Results are quite different from those obtained using the unconditional

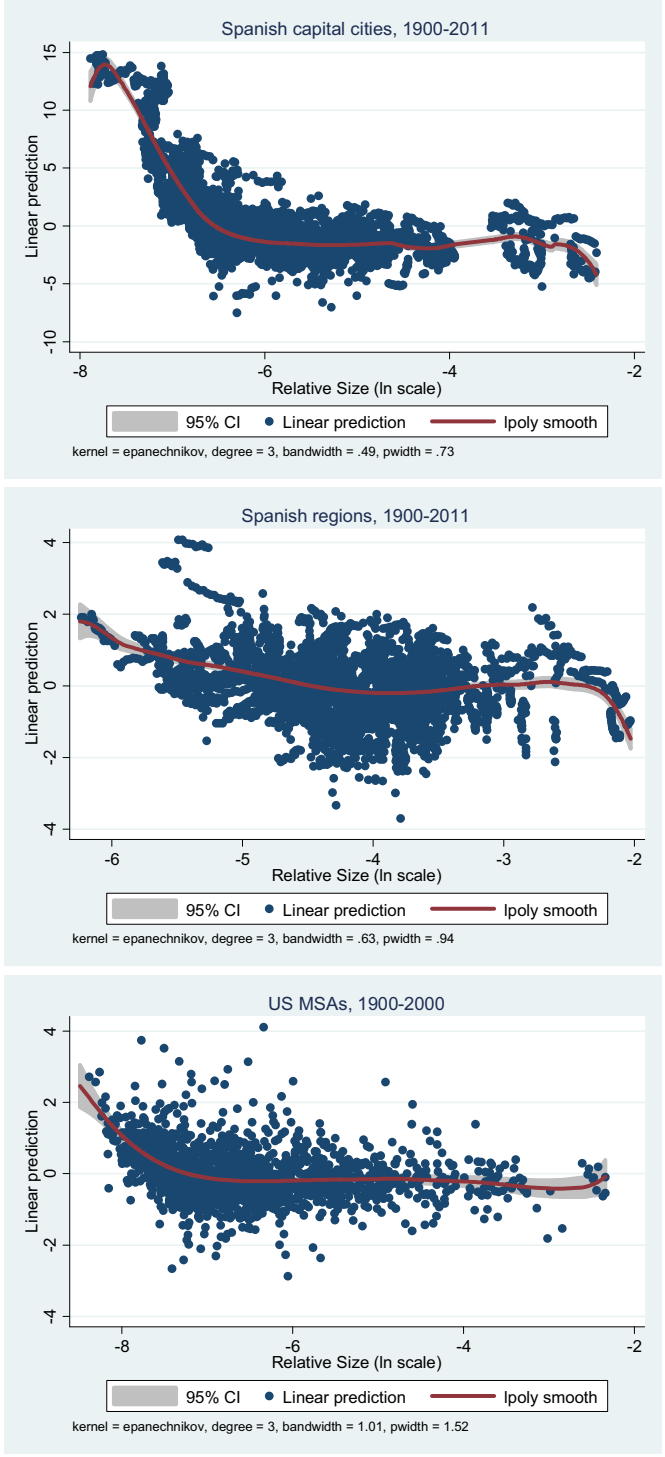


Fig. 3. Semiparametric estimates of growth.

nonparametric method in Fig. 2 or the equivalent parametric growth model including fixed effects (column 2 in Table 2). In the case of Spanish cities, we still conclude that Gibrat's law does not hold (estimates are different from zero), but the inverted U-shaped curve has disappeared, although the convergence pattern remains: cities with small relative sizes have higher growth rates, while most of the medium-sized and large cities show lower than average growth rates.

For the Spanish regions, the change in the relationship is striking. While in Fig. 2(d) the pattern was clearly divergence across regions, Fig. 3 shows that after adding the fixed effects, growth is not different from zero for most of the relative sizes. Only for the smallest and largest units do we observe higher and lower than average growth, respectively, pointing to some convergence for these extreme values of the distribution of relative sizes. Finally, for the US MSAs, the results do not change much because random growth still holds for most of the distribution. The difference is that in Fig. 2(f), it is the largest cities that show deviations from average growth, but in Fig. 3, the smallest cities show higher growth than the mean.

Overall, considering the three datasets, the semiparametric estimates give more support to Gibrat's law (random growth) than the simple nonparametric kernel regression, which means that including the influence of city/region and time fixed effects on growth helps us to discover the true underlying relationship between growth and initial size.

Moreover, the fit provided by the semiparametric and the parametric models can be different for all variables, including those in the linear part of the model, not only for the variable estimated in the nonparametric part of the model (namely, the initial relative size). In our case, the linear part only includes the time fixed effects although Eq. (6) allows for the inclusion of any observed city characteristics (x_{it}). Fig. 4 shows that even in this simple scenario in which only time fixed effects are included, strong differences can be observed between the coefficients estimates obtained by the semiparametric (linear part of the model) and the parametric models. It compares the estimated coefficients for all the time fixed effects and the corresponding 95 % confidence bands using the parametric version of the polynomial model in Eq.(4) (column 4 in Table 2) and the semiparametric model. Estimated values using these two methods are significantly different for the last years in the sample for both the Spanish regions and US MSAs, and for most of the years in the case of the Spanish cities.

5. Conclusions

This paper examines two traditional methods in the study of urban growth (parametric and nonparametric models) and proposes using a different methodology, a semiparametric model, which is a mixture of the other two. This approach takes the best of parametric and nonparametric models: it allows the linear inclusion of control variables and fixed effects and, for our variable of interest, performs a nonparametric estimate of the influence of this regressor on the dependent variable.

To illustrate the usefulness of this new approach, we test Gibrat's law in the long run, using the three methods and three different datasets: Spanish capital cities and regions (1900–2011, annual data) and US MSAs (1900–2000, decennial data). Our results reveal that the estimation of the relationship between growth and initial size significantly changes across methods. Although we only include city and time fixed effects in the linear part of the model, our results show important differences with the results obtained using unconditional nonparametric regressions, the latter being the approach popularised in the empirical literature in the last decades. Nevertheless, we must be cautious in interpretating of the results because the models are not nested and it is not easy to interpret which part of the differences in model specifications results in the differences in the estimates.

However, as far as we know, to date, these methods have been almost unused in the urban growth literature. We hope this paper helps to disseminate the advantages of this technique, and we strongly encourage the use of semiparametric methods in future research. Our results contribute to the literature addressing Gibrat's law for city sizes but also have potential applications to the literature on firm size, in which Gibrat's law has a long tradition (Sutton, 1997). Furthermore, the semiparametric method described here can be used with any panel or cross-sectional data model of urban growth. For instance, in the past, influential studies have investigated the effect of market potential (Black & Henderson, 2003) or manufacturing (Glaeser et al., 1995) on urban growth using linear regression models, and these results and those from

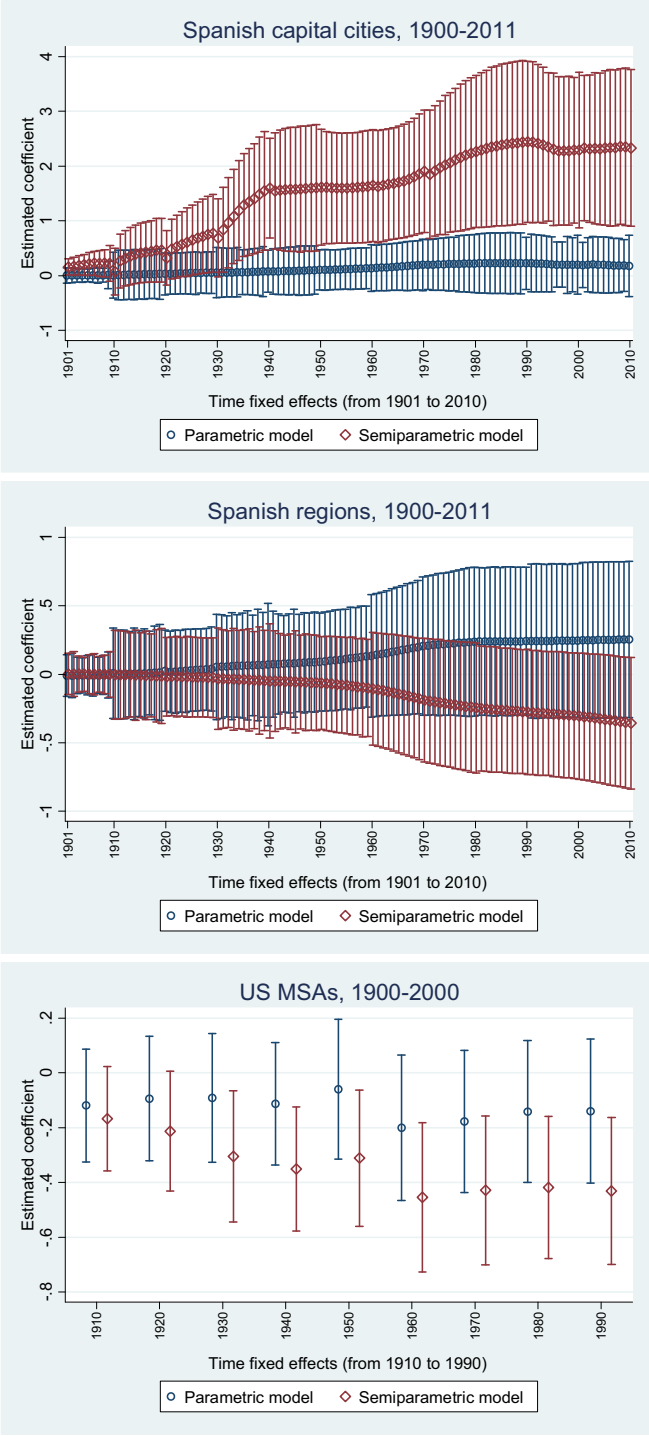


Fig. 4. Estimated coefficients of the time fixed effects by model.

many other studies could be improved by using semiparametric regressions because agglomeration economies and nonlinear effects can be better captured by a semiparametric model than by a standard parametric growth regression. We truly believe that urban growth literature can benefit in the future from the use of this tool.

CRediT authorship contribution statement

Rafael González-Val: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationshipsthat could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The author acknowledges the financial support of the Spanish Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación, MCIN/AEI/10.13039/501100011033 (projects ECO2017-82246-P and PID2020-114354RA-I00), DGA (project S39_20R, ADETRE research group), and ERDF.

References

Anderson, G., & Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics, 35*, 756–776.
Baltagi, B. H., & Li, D. (2002). Series estimation of partially linear panel data models with fixed effects. *Annals of Economics and Finance, 3*, 103–116.
Barrios, S., & Strobl, E. (2009). The dynamics of regional inequalities. *Regional Science and Urban Economics, 39*(5), 575–591.
Basile, R. (2008). Regional economic growth in Europe: A semiparametric spatial dependence approach. *Papers in Regional Science, 87*(4), 527–544.
Black, D., & Henderson, V. (1999). A theory of urban growth. *Journal of Political Economy, 107*(2), 252–284.
Black, D., & Henderson, V. (2003). Urban evolution in the USA. *Journal of Economic Geography, 3*(4), 343–372.
Bogue, D. (1953). *Population growth in standard Metropolitan Areas: 1900-1950*. Washington, DC: Housing and Home Finance Agency.
Bosker, E. M., Brakman, S., Garretsen, H., & Schramm, M. (2007). Looking for multiple equilibria when geography matters: German city growth and the WWII shock. *Journal of Urban Economics, 61*, 152–169.
Bosker, E. M., Brakman, S., Garretsen, H., & 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.
Chesher, A. (1979). Testing the law of proportionate effect. *Journal of Industrial Economics, 27*(4), 403–411.
Davis, D. R., & Weinstein, D. E. (2002). Bones, bombs, and break points: The geography of economic activity. *The American Economic Review, 92*(5), 1269–1289.
Davis, D. R., & Weinstein, D. E. (2008). A search for multiple equilibria in urban industrial structure. *Journal of Regional Science, 48*(1), 29–65.
Desmet, K., & Rappaport, J. (2017). The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. *Journal of Urban Economics, 98*, 50–68.
Devadoss, S., & Luckstead, J. (2015). Growth process of U.S. Small cities. *Economics Letters, 135*, 12–14.
Díez-Minguela, A., González-Val, R., Martinez-Galarraga, J., Sanchis, M. T., & Tirado, D. A. (2020). The long-term relationship between economic development and regional inequality: South-West Europe, 1860–2010. *Papers in Regional Science, 99*(3), 479–508.
Durlauf, S. N. (2001). Manifesto for a growth econometrics. *Journal of Econometrics, 100*(1), 65–69.
Eaton, J., & Eckstein, Z. (1997). Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics, 27*(4–5), 443–474.
Eeckhout, J. (2004). Gibrat's law for (All) cities. *American Economic Review, American Economic Association, 94*(5), 1429–1451.
Favaro, J.-M., & Pumain, D. (2011). Gibrat revisited: An urban growth model incorporating spatial interaction and innovation cycles. *Geographical Analysis, 43*, 261–286.
Fujita, M. (1976). Spatial patterns of urban growth: Optimum and market. *Journal of Urban Economics, 3*(3), 209–241.
Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quaterly Journal of Economics, 114*(3), 739–767.
Gabaix, X., & Ioannides, Y. M. (2004). The evolution of city size distributions. In J. V. Henderson, & J. F. Thisse (Eds.), *Handbook of urban and regional economics* (Vol. 4, pp. 2341–2378). Amsterdam: Elsevier Science.
Giesen, K., & Südekum, J. (2011). Zipf's law for cities in the regions and the country. *Journal of Economic Geography, 11*(4), 667–686.
Glaeser, E. L., Scheinkman, J. A., & Shleifer, A. (1995). Economic growth in a cross-section of cities. *Journal of Monetary Economics, 36*, 117–143.
Glaeser, E. L., & Shapiro, J. (2003). Urban growth in the 1990s: Is city living back? *Journal of Regional Science, 43*(1), 139–165.
González-Val, R. (2010). The evolution of US city size distribution from a long term perspective (1900–2000). *Journal of Regional Science, 50*(5), 952–972.
González-Val, R., & Olmo, J. (2015). Growth in a cross-section of cities: Location, increasing returns or random Growth? *Spatial Economic Analysis, 10*(2), 230–261.

González-Val, R., & Silvestre, J. (2020). An annual estimate of spatially disaggregated populations: Spain, 1900–2011. *The Annals of Regional Science, 65*, 491–508.
Härdle, W. (1990). *Applied nonparametric regression. Econometric society monographs*. Cambridge: Cambridge University Press.
Ioannides, Y. M., & Overman, H. G. (2003). Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics, 33*, 127–137.
Ioannides, Y. M., & Overman, H. G. (2004). Spatial evolution of the US urban system. *Journal of Economic Geography, 4*(2), 131–156.
Lessmann, C. (2014). Spatial inequality and development. Is there an inverted-U relationship. *Journal of Development Economics, 106*, 35–51.
Libois, F., & Verardi, V. (2013). Semiparametric fixed-effects estimator. *The Stata Journal, 13*(2), 329–336.
Luckstead, J., & Devadoss, S. (2014). Do the world's largest cities follow Zipf's and Gibrat's laws? *Economics Letters, 125*, 182–186.
Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics, 79*(1), 147–168.
Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics, 22*, 265–312.
Robinson, P. M. (1988). Root-N consistent semiparametric regression. *Econometrica, 56*(4), 931–954.
Ronsse, S., & Standaert, S. (2017). Combining growth and level data: An estimation of the population of belgian municipalities between 1880 and 1970. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 50*(4), 218–226.
Sánchez-Vidal, M., González-Val, R., & Viladecans-Marsal, E. (2014). Sequential city growth in the US: Does age matter? *Regional Science and Urban Economics, 44*, 29–37.
Sharma, S. (2003). Persistence and stability in city growth. *Journal of Urban Economics, 53*(2), 300–320.
Skouras, S. (2009). *Explaining Zipf's Law for US Cities (December 1, 2009)*. Available at SSRN: https://doi.org/10.2139/ssrn.1527497. or http://ssrn.com/abstract=1527497.
Sutton, J. (1997). Gibrat's legacy. *Journal of Economic Literature., 35*(1), 40–59.
Wheeler, C. H. (2003). Evidence on agglomeration economies, diseconomies, and growth. *Journal of Applied Econometrics, 18*(1), 79–104.