

# Sequential city growth: Theory and evidence from the US

SHENG Kerong<sup>1</sup>, \*SUN Wei<sup>2,3</sup>, FAN Jie<sup>2,3</sup>

1. Business School, Shandong University of Technology, Zibo 255012, Shandong, China;

2. Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing 100101, China;

3. Key Laboratory of Regional Sustainable Development Modeling, CAS, Beijing 100101, China

**Abstract:** City growth patterns are attracting more attention in urban geography studies. This paper examines how cities develop and grow in the upper tail of size distribution in a large-scale economy based on a theoretical model under new economic geography framework and the empirical evidence from the US. The results show that cities grow in a sequential pattern. Cities with the best economic conditions are the first to grow fastest until they reach a critical size, then their growth rates slow down and the smaller cities farther down in the urban hierarchy become the fastest-growing ones in sequence. This paper also reveals three related features of urban system. First, the city size distribution evolves from low-level balanced to primate and finally high-level balanced pattern in an inverted U-shaped path. Second, there exist persistent discontinuities, or gaps, between city size classes. Third, city size in the upper tail exhibits conditional convergence characteristics. This paper could not only contribute to enhancing the understanding of urbanization process and city size distribution dynamics, but also be widely used in making effective policies and scientific urban planning.

**Keywords:** sequential city growth; discontinuities; city size distribution; conditional convergence

## 1 Introduction

Ever since Auerbach discovered the regularity in 1913 that the size distribution of cities is well approximated by a Pareto distribution, the study of city size distribution has attracted sustained interest of researchers (Berry, 1961). From the late 1950s, the focus of this research has gradually shifted to examinations of the economic mechanisms underlying the city size distribution. Explanations fall into two types: those that are based on the location choice behavior of microeconomic agents or urban system models, and those that focus on the relationship between the regional development conditions and city size distribution characteristics. The former are represented by the central place theory developed by Christaller (1966) and urban system theory developed by Henderson (1974; 1991). Turning

**Received:** 2014-01-26 **Accepted:** 2014-02-18

**Foundation:** National Natural Science Foundation of China, No.41230632; Key Project for the Strategic Science Plan in IGSNRR, CAS, No.2012ZD006

**Author:** Sheng Kerong (1977–), Associate Professor, specialized in economic geography. E-mail: shengkerong@163.com

\***Corresponding author:** Sun Wei (1975–), Associate Professor, E-mail: sunw@igsnr.ac.cn

to latter, Krugman and Livas (1996), Alonso (2001), Ades and Glaeser (1995) have examined the impacts of trade conditions and political factors on regional mega cities; Zhou (1986), Gu (2008), Lu (2002) and others have studied the effects of regional economic base, as well as natural and geographical conditions on the structural characteristics of urban systems. However, these studies have focused on the interpretation of static or short-run evolutionary characteristics of city size distribution. The study of how cities grow and consequently how urban systems evolve in the long run is critical for either in-depth understanding of the urbanization process or the making and implementation of regional policies.

It was not until recently that researchers began to place attention to the study of city growth dynamics. Henderson and Venables (2009) have developed a model of city formation in which urban agglomerations grow in a sequential pattern: the initially largest city will grow firstly, and then the cities with higher ranks will grow successively. In their model, two issues regarding the deviation of equilibrium city size from efficient size in the competitive economy and the impact of government fiscal policies on urban system efficiency were also analyzed. Cuberes (2011) undertook a thorough empirical analysis by taking into account a longer time period and a greater number of countries, and he showed the average rank of the fastest-growing cities increased over time gradually, which provides empirical evidence for the sequential pattern of urban growth. However, these studies are mainly based on neo-classical economics and neutral space theory, and cannot tell in which locations city will grow first, nor can they explain the differences of city sizes and the hierarchical characteristics of city size distribution (Fujita and Mori, 1997; Fujita *et al.*, 1999).

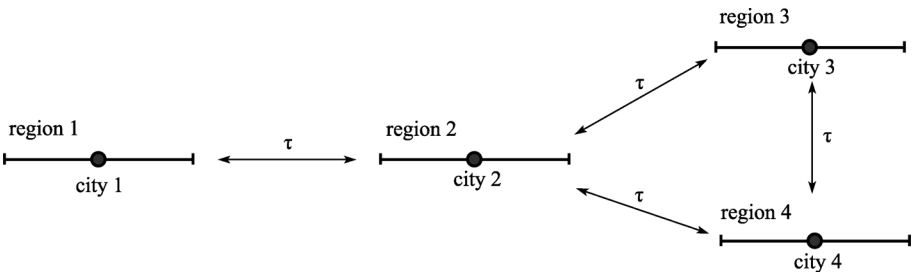
This paper uses a general equilibrium model of industrial location and city growth and US data to examine the relative growth of cities, the evolution dynamics of city size distribution, as well as the underlying driving mechanisms. We show that cities grow in a sequential pattern in the process of urbanization. Cities with the largest market potentials are the first to grow fastest. Only when these cities reach a critical size do the second-level cities in urban hierarchy become the fastest-growing ones. And the third-level cities take the lead when the second-level cities too reach a critical size, and so on. This paper also reveals three other characteristics in city size distribution evolution consistent with sequential city growth pattern: the inverted U-shaped pattern of first increasing and then decreasing urban concentration, the persistent discontinuities in city size distribution, and the conditional convergence in city size in the upper tail. This paper can not only enhance our understanding of urbanization process and city size distribution dynamics, but also provide a new perspective for understanding the spatial structure in regional development.

## 2 The model

### 2.1 The spatial economy

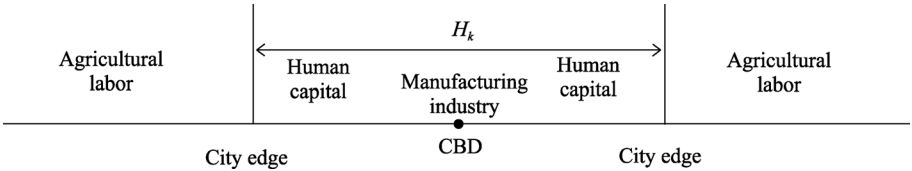
Our model is based on Forslid and Ottaviano (2003) version of Krugman (1991). We simulate the growth patterns of major cities in the upper tail of size distribution in a large-scale economy (Figure 1). In our model, the space of the economy is composed of four one-dimensional regions, denoted by region 1, region 2, region 3 and region 4, and two factors of production, human capital and labor with total endowments of  $H$  and  $L$  respectively. We assume the shares of labor in regions 1, 2, 3 and 4, i.e.  $L_1$ ,  $L_2$ ,  $L_3$  and  $L_4$ , are  $1/4$ ,  $1/4$ ,

$1/4+\varepsilon$  and  $1/4-\varepsilon$  respectively. The economy consists of two sectors, agricultural sector (A) and manufacturing sector (X). The agricultural goods are homogenous, traded without costs and produced under constant returns with labor as the only input. For simplicity, it is assumed that the production of unit agricultural product requires unit labor, so that, restricted by agricultural production conditions, agricultural labor is evenly distributed in each region. The manufacturing sector is characterized by firm-specific increasing returns to scale, and its cost function consists of fixed cost and variable cost. It is assumed that the fixed cost includes only human capital, and variable cost includes only labor. Choosing units and letting one enterprise uses only one unit of human capital as the fixed cost, the number of human capital in each region, i.e.  $H_k$ , is equal to that of the firms  $n_k$  in the manufacturing sector. The spatial aggregation of the manufacturing sector leads to the emergence of cities. In this paper, it is assumed that the enterprises can only aggregate in the center of each region, and then the city (city 1, city 2, city 3 and city 4) is located in the regional center.



**Figure 1** The spatial configuration of economic activities

The city has a monocentric spatial structure pioneered by Alonso (1964): all production in a city occurs at a point, the central business district (CBD); extending from the CBD is a stretch of residence, where each human capital lives on a lot of unit size and commutes to the CBD (and back) at a constant cost per unit of distance  $f$  (Figure 2). The residents living with a distance  $d$  from the CBD have to pay a commuting cost  $fd$  plus a land rent  $g(d)$ . Assuming that the urban land belongs to all the residents of the city, land rents are finally evenly distributed to each resident. Thus the cost of living of resident is land rent plus commuting cost minus land rent subsidy. Free mobility of residents in the city  $k$  ( $k = 1, 2, 3, 4$ ) requires all residents to pay the same cost of living  $\beta n_k$ , where  $\beta=f/4$ .<sup>1</sup> As we will see in subsequent analysis, living cost is a major constraint for city size expansion.



**Figure 2** The urban spatial structure

<sup>1</sup> (1) In equilibrium of location choice, land rent plus commuting cost per person equal the commuting cost of the edge worker,  $g(d)+fd=g(n_k/2)+fn_k/2$ . And the land rent the edge worker paid is zero, so  $g(d)=f(n_k/2-d)$ . (2) The total rent of urban land is  $2\int_0^{n_k/2} g(x)dx=fn_k^2/4$ , and the rent subsidy per person is  $fn_k/4$ . Thus, the living cost is  $g(d)+fd-fn_k/4=fn_k/4$ .

Every consumer, labor or human capital, shares the same quasi-linear utility tastes (Pflüger, 2004). Taking into account the fact that the living cost of human capital living in cities is higher than that of rural labors by  $bn_k$ , each human capital's in city  $k$  is characterized by:

$$U_k = \gamma \ln C_x + C_A - \beta n_k, \quad C_x = \left( \sum_{j=1}^4 \sum_{i=1}^{n_j} c_{ji}^{(\sigma-1)/\sigma} \right)^{\sigma/(\sigma-1)}, \quad \gamma > 0, \sigma > 1 \quad (1)$$

where  $C_x$  represents the composite index of differentiated manufacturing goods,  $c_{ji}$  represents the quantity consumed of variety  $i$  in region  $j$ ,  $n_j$  is the number of varieties produced in region  $j$ ,  $C_A$  represents the consumption of agricultural good, and  $\sigma$  represents the price elasticity of demand and the elasticity of substitution between any two varieties in the manufacturing sector.

The agricultural sector has a perfectly competitive market structure. The agricultural goods are chosen as numéraire, thus the price of the agricultural goods and labor wage is unity. In the long run, human capital is mobile inter-regionally, while labor force can only move freely between two sectors in the same region. All products of manufacturing industry have the same production technology such that cost function of the representative function is given by  $R + a_m x$ , where  $R$  and  $a_m$  are, respectively, the compensation of one unit of human capital and marginal labor requirements. The product of the manufacturing sector has a Dixit-Stiglitz monopolistic competitive market structure (Dixit and Stiglitz, 1977). Provided the number of varieties is large, profit maximizing prices are constant markups on marginal cost, and the market price of product  $i$  is  $p_i \equiv p = a_m \sigma / (\sigma - 1)$ . The manufactured goods need no transportation cost in the same region, while the trade between regions is inhibited by Samuelson iceberg costs. For example, when one unit product of the modern sector is shipped from region 1 to region 3, only  $1/(2\tau)$  arrives to region 3, while the prices charged in region 3 increases by  $(2\tau)$  times, where  $\tau (\tau > 1)$  is the parameter to characterize the transportation cost.

## 2.2 Short-run equilibrium

In the short run, the spatial configurations of human capital and labor are both given, and the regional development environment (price elasticity of demand for manufactured goods, population size and structure, transportation cost, urban congestion effect, etc.) has no changes. Using the conditions of market clearing and zero profit, the nominal wage rate of human capital in each city is solved:

$$R_1 = (p - a_m)[(L_1 + n_1)x_{11} + (L_2 + n_2)\tau x_{12} + 2(L_3 + n_3)\tau x_{13} + 2(L_4 + n_4)\tau x_{14}] \quad (2)$$

$$R_2 = (p - a_m)[(L_1 + n_1)\tau x_{21} + (L_2 + n_2)x_{22} + (L_3 + n_3)\tau x_{23} + (L_4 + n_4)\tau x_{24}] \quad (3)$$

$$R_3 = (p - a_m)[2(L_1 + n_1)\tau x_{31} + (L_2 + n_2)\tau x_{32} + (L_3 + n_3)x_{33} + (L_4 + n_4)\tau x_{34}] \quad (4)$$

$$R_4 = (p - a_m)[2(L_1 + n_1)\tau x_{41} + (L_2 + n_2)\tau x_{42} + (L_3 + n_3)\tau x_{43} + (L_4 + n_4)x_{44}] \quad (5)$$

where  $x_{jk} = ap_{jk}^{-\sigma} P_k^{\sigma-1}$  represents the quantity of a variety consumed in region  $k$  produced in region  $j$ , of which  $p_{jk}$  represents the delivered price of a variety in region  $j$  imported from region  $k$ . For example, the price charged in region 4 for a variety produced in region 2 is  $p_{24} = \tau p$ , and that charged in region 1 from region 3 is  $p_{31} = 2\tau p$ .  $P_k$  is the price index for the

manufactured goods in region  $k$ , given by:

$$P_1 = (n_1 p^{1-\sigma} + n_2 (\tau p)^{1-\sigma} + n_3 (2\tau p)^{1-\sigma} + n_4 (2\tau p)^{1-\sigma})^{1/(1-\sigma)} \quad (6)$$

$$P_2 = (n_1 (\tau p)^{1-\sigma} + n_2 p^{1-\sigma} + n_3 (\tau p)^{1-\sigma} + n_4 (\tau p)^{1-\sigma})^{1/(1-\sigma)} \quad (7)$$

$$P_3 = (n_1 (2\tau p)^{1-\sigma} + n_2 (\tau p)^{1-\sigma} + n_3 p^{1-\sigma} + n_4 (\tau p)^{1-\sigma})^{1/(1-\sigma)} \quad (8)$$

$$P_4 = (n_1 (2\tau p)^{1-\sigma} + n_2 (\tau p)^{1-\sigma} + n_3 (\tau p)^{1-\sigma} + n_4 p^{1-\sigma})^{1/(1-\sigma)} \quad (9)$$

Next, the indirect utility function of human capital in each city can be derived as follows:

$$V_1 = -\gamma \ln(P_1) + R_1 + \gamma(\ln \gamma - 1) - \beta n_1 \quad (10)$$

$$V_2 = -\gamma \ln(P_2) + R_2 + \gamma(\ln \gamma - 1) - \beta n_2 \quad (11)$$

$$V_3 = -\gamma \ln(P_3) + R_3 + \gamma(\ln \gamma - 1) - \beta n_3 \quad (12)$$

$$V_4 = -\gamma \ln(P_4) + R_4 + \gamma(\ln \gamma - 1) - \beta n_4 \quad (13)$$

In the short run equilibrium, according to equations (10), (11), (12) and (13), the utility of human capital in city  $k$  depends on the nominal wage rate  $R_k$ , the price index of manufactured goods  $P_k$  and city size  $n_k$ . As can be seen from the wage equations (2), (3), (4) and (5), the larger the market size ( $L_k + n_k$ ) of the city in region  $k$  is, the higher the wage rate of human capital paid by the manufacturing firms will be. From the living cost index equations (6), (7), (8) and (9), the larger the size ( $n_k$ ) of city  $k$  is, the lower the price index for consumers living in city  $k$  will be. Such demand and supply linkages will increase the utility of human capital for large cities, constituting the agglomerating force of economic activity. On the other hand, the larger the size ( $n_k$ ) of city  $k$  is, the higher the living cost of residents living in city  $k$  will be, thereby leading to a decline in the utility of human capital. This congestion effect and the agricultural labor market dispersed in the four areas constitute the spreading force of economic activity.

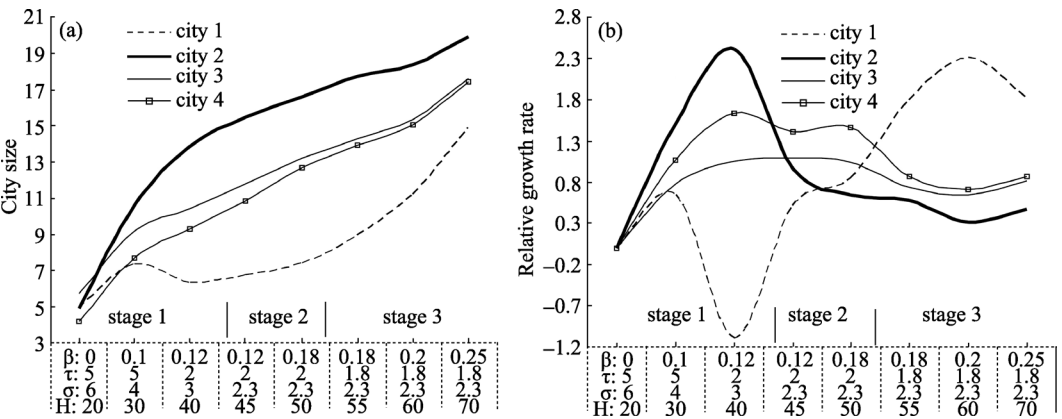
### 2.3 Long-run equilibrium

In the long run, the structural economic changes in regional development are the driving force for the evolution of city size distribution (Krugman, 1991; Brakman *et al.*, 2001). The structural changes constantly alters the relative strength of agglomerating forces and spreading forces, leading to the inter-regional flows of human capital, and consequently the dynamic evolution of industrial location and city size distribution. The following section simulates the urban growth and the dynamic pattern of the city size distribution evolution in the regional development process.

During the transition towards an industrialized society in regional development process, economic parameters changed dramatically in four aspects. (1) Transportation cost declined dramatically. In particular, the construction of railways greatly increased the importance of economies of bundled transportation, and significantly reduced the transaction cost of manufactured products; we stimulate this process by decreasing parameter  $t$  from 5 to 1.8. (2) The manufacturing is transformed from handicraft to large-scale machine production. By the mid-19th century when the Industrial Revolution came to an end in Europe and the US, large-scale machine production dominated the manufacturing sector; we capture this change by decreasing the parameter  $\sigma$  to 2.3 from 6. (3) Population size and the share of urban population increase greatly. We stimulate this by fixing the amount of labor at 50, while in-

creasing the scale of human capital gradually to 70 from 20. (4) Negative feedbacks such as congestion grow significantly in large cities. Congestion mainly refers to traffic congestion and soaring land prices. In the model, we stimulated this change by increasing the parameter  $\beta$  gradually from 0 to 0.25.

To stimulate the relationship between economic parameters and city growth, we simplify the analysis by choosing units of manufactured products and letting  $a_m = (\sigma - 1)/\sigma$ , so that the price of the manufactured product is normalized to unity. At the same time, the parameter  $\varepsilon$  which denotes the difference of agricultural labor between regions 3 and 4 is fixed at 0.02, and parameter  $\gamma$  denoting the preference for the differentiated manufactured goods is fixed at 3. Figure 3 illustrates the dynamics of city growth in the upper tail as a function of the economic parameters  $\beta$ ,  $t$ ,  $\sigma$  and  $H$ , in which, panel (a) depicts the evolution of long-run equilibrium city sizes, and panel (b) describes the evolutionary process of the relative city growth (ratio between urban growth rate and average growth rate). Numerical simulation reveals that city growth in the regional development process can be divided into three stylized stages. Stage 1 represents the transition from traditional society to industrialized society; in panel (a), the population sizes of cities 1, 2, 3 and 4 increase from 4.97, 4.97, 5.79 and 4.26 to 6.38, 13.88, 10.43 and 9.31, respectively. Stage 2 represents the industrialization period; in panel (a), the population sizes of the four cities increase to 7.43, 16.64, 13.22 and 12.71, respectively. Stage 3 represents the post-industrialization period; the sizes of the four cities increase to 14.93, 19.95, 17.65 and 17.47, respectively. The model has the following four testable implications that will be explored in the next section.



**Figure 3** Numerical stimulations of city growth

First, cities grow in a sequential order. In stage 1, due to a substantial decline in transportation cost and significant increase in economies of scale in the manufacturing sector, the city with the largest market potential (city 2 in the model) is the first to grow rapidly and eventually become the primate city in the spatial economy. In stage 2, congestion starts to dominate the agglomerating force in the primate city; the growth rate of the primate city declines, the cities with better location conditions (city 3 and city 4) become the fastest-growing cities. In stage 3, the growth rates of the medium-sized cities (city 3 and city 4) decrease, while the size of smaller city (city 1) expands rapidly. Therefore, the model implies that the average rank of the fastest growing cities should increase with improvement of

urbanization level.

Second, city size distribution evolves in an inverted U-shaped pattern. In traditional society, due to high transportation cost and relative unimportance of returns to scale, the cities are small, and urban population are evenly distributed across the various cities, leading to a low-level balanced distribution of city size. During the period of transition toward a modern society, the urban landscape is changed dramatically with the rapid growth of certain cities (e.g. city 2), and city size distribution is transformed from a balanced into a primate pattern. As time passes, the medium and small cities in the urban hierarchy sequentially become the fastest-growing cities, indicating the urban population in the lower tail expands, and the city size distribution then exhibits a balanced pattern at a high level. The model predicts that the city size distribution evolve along an inverted U-shaped path from low-level balanced to primate and finally high-level pattern.

Third, the structure of urban system exhibits persistent discontinuity. Cities grow in sequential order, which indicates that cities' growth is non-parallel, and the city size distribution should exhibit discontinuity and long-term deviation from the rank-size regularity. At the same time, there are differences in development conditions (such as the location in the model) between cities. Cities with similar development conditions (city 3 and city 4 in the model) tend to have similar population sizes, and form specific clusters in city size distribution. As a result, the model implies that distinct clusters should be identified in the size distribution.

Fourth, the sizes of cities in the upper tail converge conditionally. Because of the sequential characteristics of urban growth, the growth rate of the primate city slows down at some point, and the higher-ranked cities take the lead in growth. As a result, the size difference between the primate city (city 2), the medium-sized cities (city 3 and city 4) and the small city (city 1) is predicted to decrease over time. However, the difference in city development conditions exists in the long run, the cities with different development conditions will eventually converge to different population sizes. The model predicts that the size of cities (at least) in the upper tail will show conditional convergence characteristics over time.

### 3 Empirical evidence

#### 3.1 Data sources

The paper exploits the US data to investigate the evolutionary characteristics of cities in the upper tail of size distribution. There are three reasons why the experience of the US is of interest. First, the US has been transformed from a predominantly rural, agricultural nation into an urbanized, industrial one over the last two centuries, and the US case can be used to investigate the evolutionary dynamics of city size distribution during the complete urbanization process. Second, the US is a large-scale economy and has a well-developed market system, so that we can investigate thoroughly the city growth patterns under decentralized market economy (Tan and Li, 2010). Third, the historical population data of the US cities can be easily collected, and the accuracy can also be checked through multiple channels. The data used in this study comes from Jan Lahmeyer (<http://www.populstat.info/>), who collects historical population data of the major US cities from 1820 to 2000 on a decadal frequency.

This study focuses on cities in the upper tail of size distribution. Two methods are used to

determine the sample cities: the relative cutoff value method, by which cities with population size above 60% of the average in each year are collected, and the absolute cutoff value method, by which a total of 89 cities with inhabitants of more than 200,000 in 2000 are selected. These two methods are complementary, of which the city samples obtained by the absolute cutoff value method contain those obtained by the relative cutoff value method. Next we test the sequential city growth pattern and the accompanying distribution dynamics, clustering characteristics and conditional convergence predicted by the model.

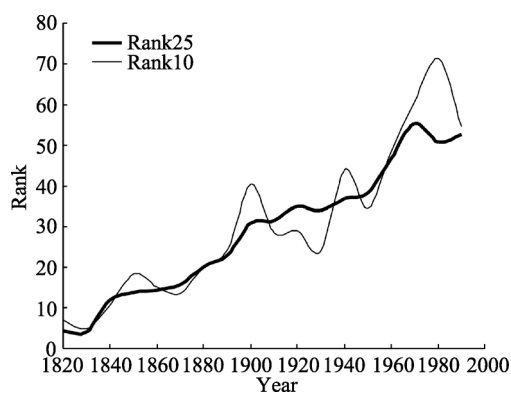
### 3.2 Sequential city growth

We follow the methodology used by Cuberes (2011) to address this issue. The data of many cities trace back to 1820, when many of the existing cities in recent urban landscape did not exist, and many cities with population records were much smaller. In order to avoid the estimation and interpretation problem that cities are too small to lead to high growth rate, the sample from relative cutoff value method is used.

First, based on the population growth rates of sample cities in each historical stage, a statistical description analysis is carried out. The Jarque-Bera (JB) test reveals that, except for 1820 and 1830 in which the JB coefficients are not statistically different from zero, the observations in all the other years reject city growth rates distributed normally. The skewness test also reveals that the distribution of city growth rates in all years is positively skewed. The test indicates that city growth is non-parallel, and the peaks of the growth rate in each decade are concentrated in a small number of cities.

Next, the average rank of the fastest-growing cities and its evolution path are analyzed. (1) We begin by ranking the cities by size—in terms of population—at every decade, with the largest city having rank 1, the second largest city having rank 2, and so on. (2) The cities are then sorted in a decreasing order by the growth rate, and the average rank of the cities with growth rates at the top 25% and 10% are calculated and denoted as Rank 25 and Rank 10, respectively. (3) This routine is repeated for every decade and thus ends up with a time series of Rank 25 and Rank 10 (Figure 4). It can be seen from this figure that Rank 25 and Rank 10 present a clear, positive trend in the 1820–1990 time interval.

Here we run the regression analysis to determine the relationship of Rank 25 and Rank 10 with time variable ( $t$ ). As mentioned above, an important feature of our database is that the



**Figure 4** Evolution of Rank25 and Rank10 for US urban system, 1820–2000

number of observations grows over time. To account for this, we include the logarithm of the number of cities without missing data ( $\ln(\text{number})$ ) as a control variable. In some specifications we also include the squared term of time variable ( $t^2$ ) as a control in order to better capture the nonlinearity of the change of Rank 25 and Rank 10. Table 1 shows the estimates with ordinary least square method. In the specifications, the coefficients of time variable and its squared term which fail to be rejected at the 5% level are both positive. The results show that the average rank of the fast-



est-growing cities rises over time, i.e. urban growth shows a sequential pattern.

**Table 1** Regressions of Rank25 and Rank10 on time for US urban system, 1820–2000

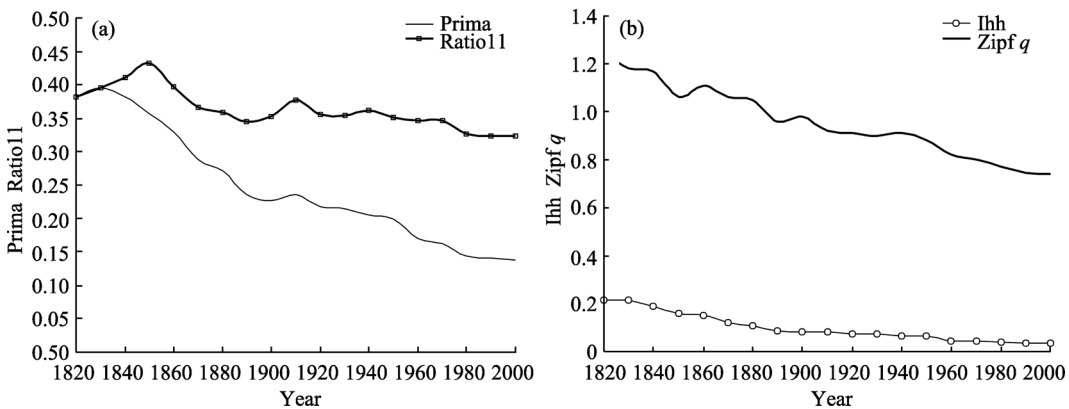
	Rank25			Rank10		
	(I)	(II)	(III)	(IV)	(V)	(VI)
$t$	2.873*** (0.206)		2.076** (0.703)	3.263*** (0.581)		−0.943 (1.721)
$t^2$		0.143*** (0.012)	0.042 (0.036)		0.177*** (0.025)	0.223** (0.087)
$\ln(\text{number})$	1.212** (0.518)	4.134*** (0.407)	1.959** (0.807)	0.822 (1.464)	3.740*** (0.792)	4.729** (1.977)
$R^2$	0.969	0.956	0.97	0.830	0.887	0.884
$DW$ 值	1.728	1.172	1.86	1.325	1.651	1.631

Note: Robust standard errors in the parentheses; \*\* and \*\*\* denote significance at 5% and 1% level, respectively.

3.3 Inverted U-shaped evolution path

Primary ratio is used to describe the concentration of urban population in the largest cities. Primate ratio (Prima) is defined as the share of the largest city New York ( $Pop_1$ ) in the total urban population,  $Prima = Pop_1 / (Pop_2 + Pop_3 + \dots + Pop_m)$ . In order to consider the relative size of New York in the top 11 largest cities, we construct the variable  $Ratio11 = Pop_1 / (Pop_2 + Pop_3 + \dots + Pop_{11})$ , where  $Pop_2, Pop_3, \dots, Pop_m$  refers to the size of the cities from Rank 2 to Rank  $m$ .

The Herfindahl-Hirschman index and Zipf  $q$  index are used to describe the city size distribution as a whole. The Herfindahl-Hirschman index is defined as the sum of squares of urban population share, i.e.  $Ihh = \sum (SP_j)^2$ , where  $SP_j$  is the share of city  $j$  in total urban population. The index of the Zipf  $q$  is estimated from the equation  $\ln(Pop_j) = \ln(cont) - q \ln(Rank_j) + u_j$ , where  $Pop_j$  and  $Rank_j$  are the size and rank of city  $j$  respectively,  $cont$  is the constant, and  $u_j$  is a standard error term. The sample from relative cutoff value method is used to address this issue. It plots the evolution of primary ratio, Herfindahl-Hirschman index and Zipf  $q$  during 1820–2000. Figure 5 clearly reveals that since about 1840, all the indices show a downward trend, and the Zipf  $q$  index declined from 1.25 in 1820 to 0.74 in



**Figure 5** The evolution of Prima, Ratio11, Ihh and Zipf  $q$  for US urban system, 1820–2000

2000. Considering the substantial expansion of the US urban system after 1820, it can be inferred that after 1820 the US urban system began to transit from primate to high-level balanced distribution.

The regression analysis of primacy ratio, Zipf  $q$  index and Herfindahl-Hirschman index with time variable ( $t$ ) further confirms the evolutionary trend of the US urban system (Table 2). In the econometric analysis, the number of cities in each decade is controlled. And to avoid the potential problems caused by the limited dependent variable— $Ihh$  can only vary between 0.213 and 0.035—the inverse of  $Ihh$ , i.e.  $IvIhh$ , is taken as a dependent variable for regression. The coefficients of time variable ( $t$ ) is negative and statistically significant in specifications (VII), (VIII) and (X), while positive and statistically significant in regression (IX), confirming that the primary ratio, and Zipf  $q$  index show a downward trend over time. The econometric results also reveal that the coefficient of  $\ln(number)$  is negative and statistically significant. These provide evidence for the decline of the primate city in urban population and the improvement of urban system.

**Table 2** Regressions of Prima, Ratio11,  $Ihh$  and Zipf  $q$  for US urban system, 1820–2000

	Prima	Ratio11	$IvIhh$	Zipf $q$
	(VII)	(VIII)	(IX)	(X)
$t$	−0.007** (0.003)	−0.004*** (0.001)	2.157*** (0.312)	−0.016*** (0.004)
$\ln(number)$	−0.065*** (0.019)		−5.702** (2.275)	−0.078** (0.032)
$cont$	0.547*** (0.049)	0.403*** (0.007)	16.262** (5.892)	1.399*** (0.075)
$R^2$	0.968	0.699	0.953	0.976

Note: Robust standard errors in the parentheses; \*\* and \*\*\*denote significance at 5% and 1% level, respectively.

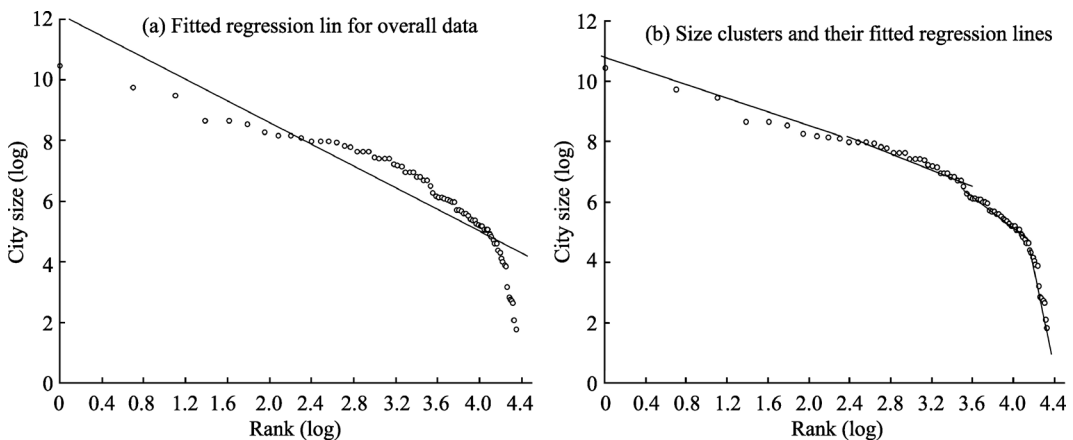
These evidences show that since about 1820, the US urban system entered the stage of transition towards a high-level balanced urbanization, i.e. the concentration of the US urban population reached a peak in the early 19th century. The US development history reveals that in the colonial period the settlements were small and evenly distributed; from the 18th to the early 19th century, New York and Philadelphia were the fastest-growing cities, and the US city size distribution was transformed from low-level balanced to primate distribution (Walton and Rockoff, 2009). Therefore, in the transition from traditional society to modern society, the evolution of city size distribution shows inverted U-shaped characteristics, i.e. from low-level balanced distribution to primate distribution and finally high-level balanced distribution.

### 3.4 Discontinuities in urban structure

The kernel density estimation method developed by Silverman (1981) is used to study the discontinuous pattern of city size distribution. We focus our analysis on the samples in 1900 and 1950 with data collected using absolute cutoff value method. To reveal the hierarchical clustering characteristics in a straightforward way, linear regression models relating log rank to log city size are fitted: one for the overall database, and one for each identified size class. A note of caution is that the regression slope is an estimate of Zipf  $q$  (Garmestani *et al.*,

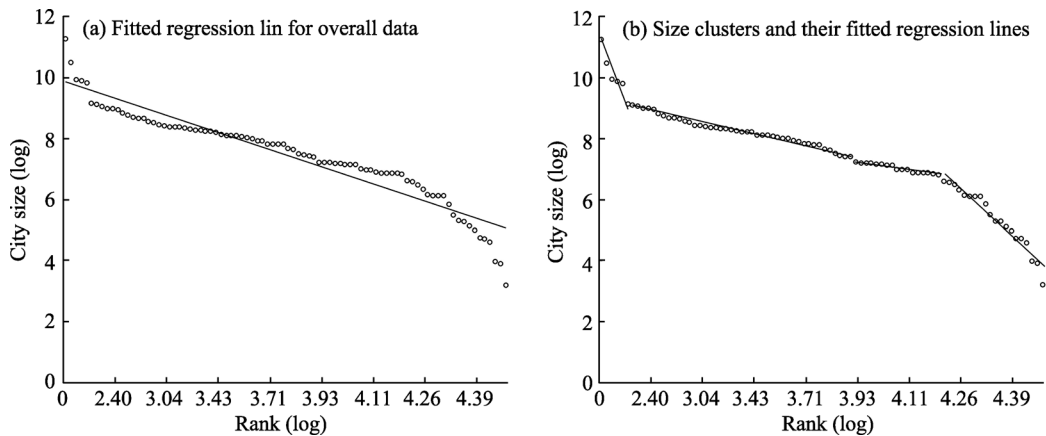
2005; Garmestani *et al.*, 2008).

There are 77 city samples in 1900 that are self-organized into 4 groups with discrete sizes (Figure 6). The regression slope for the overall rank-size city data is  $-1.788$ , while the coefficient of determination  $R^2$  is only 0.889. Power law provides good fits for each of the individual size classes. The first group consisting of the 10 largest cities results in a least squares line that explains 0.986 of the variation in the 10 log city size and has an estimated slope of  $-1.079$  with an associated standard error of 0.065. The second group of 24 cities has a fitted line explaining 97.7% of the variation in those log city sizes. The slope for this group is  $-1.320$  with a standard error of 0.062. The regression line for the third group of 27 cities explains 0.991 of the variation in log city size and has an estimated slope of  $-2.531$  with a standard error of 0.073. The final group containing the 16 smallest cities results in a regression line that explains 0.972 of the variation in log city size with an estimated slope of  $-13.53$  and a standard error of 0.874. The four size classes are differentiated by different slopes, e.g. the regression slope of the third group is more than that of the second group by 19 times standard errors, but less than that of the fourth group by 157 times standard errors, revealing the hierarchical cluster structure of the city size distribution.



**Figure 6** Hierarchical clusters in the US city size distribution in 1900

There are 83 city samples in 1950 which are self-organized into four discrete groups (Figure 7). The regression line for the overall database has an estimated slope of  $-1.394$ , but explains only 0.866 of the variation in log city size. The first group consisting of the 6 largest cities results in a least squares line that explains 0.972 of the log city size variation and has an estimated slope of  $-1.051$  with a standard error of 0.131. The regression line for the second group of 40 cities explains 0.973 of the variation in log city size with an estimated slope of  $-0.803$  and a standard error of 0.031. The third group containing 17 cities has a fitted line that explains 0.978 of the variation in those log city sizes; the slope for this group is  $-1.642$  with a standard error of 0.108. The fourth group consisting of the 20 smallest cities has a fitted line explaining 0.965 of the variation in those log city sizes; the slope for this group is  $-12.01$  with a standard error of 0.823. The estimated slope for the third group is more than that of the second group by 27 times standard errors, but less than that of the fourth group by 96 times standard errors, revealing the persistent discontinuous characteristics of the urban structure.



**Figure 7** Hierarchical clusters in the US city size distribution in 1950

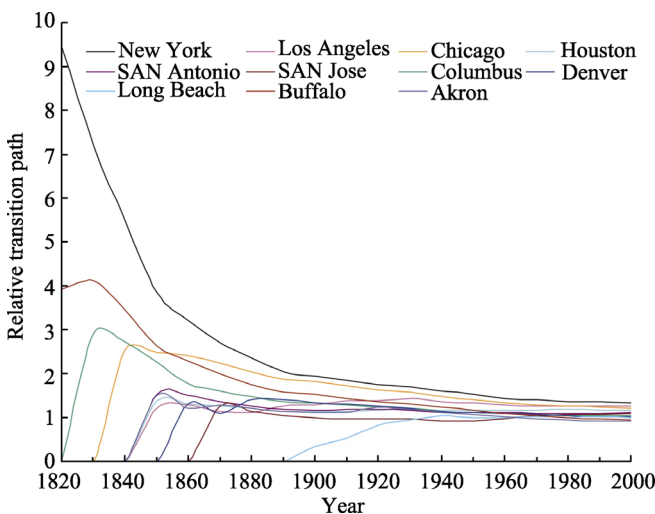
### 3.5 Conditional convergence in city size

In order to analyze the transitional behavior of city size, we apply the  $\log(t)$  test developed by Phillips and Sul (2007). The regression model of the  $\log(t)$  test is given by:

$$\log(H_1/H_t) - 2\log(\log(t)) = \beta_0 + \beta_1 \log(t) + u_t \quad (14)$$

where  $H_1/H_t$  is the cross-sectional variance ratio;  $H_t$  is the transition distance, denoted by  $H_t = N^{-1} \sum_{i=1}^N (h_{jt} - 1)^2$ , of which  $h_{jt} = \log(S_{jt}) / \left( N^{-1} \sum_{i=1}^N \log(S_{jt}) \right)$ , and  $S_{jt}$  is the population size of city  $j$ .

According to Phillips and Sul, the variable  $h_{jt}$ , also called relative transition path, traces out the trajectory of city  $j$  to the average. For intuitive purposes, we start by plotting the relative transition paths for some selected cities in the upper tail of the US urban system. A total of 11 cities, namely New York, Los Angeles, Chicago, Houston, San Antonio, San Jose, Columbus, Denver, Long Beach, Buffalo and Akron, are selected, and the relative transition



**Figure 8** Relative transition paths for selected cities in the upper tail of the US urban system, 1820–2000

paths of these representative cities since 1820 are shown in Figure 8. Figure 8 reveals that the relative differences in the city size declines over time, indicating a convergence trend in the upper tail of the size distribution.

Next, the formal econometric test based on Equation (14) is conducted. The null hypothesis of the  $\log(t)$  test is that city sizes converge over time. In Equation (14), the  $t$ -statistic of the estimated coefficient of  $\log(t)$ ,  $\beta_1$ , plays a key role in determining whether the urban growth is

convergent or divergent. If the  $t$ -statistic is less than  $-1.65$ , the null hypothesis of convergence in city size can be rejected at the 5% level. Conversely, if the  $t$ -statistic of the estimated  $\beta_1$  is larger than  $-1.65$ , the null hypothesis can not be rejected. According to Phillips and Sul, if the null hypothesis of convergence is rejected, it should also be investigated whether or not the club convergence exists.

The log ( $t$ ) test requires a balanced panel data. We use the sample obtained by the absolute cutoff value method to address this issue. And we carry out a little data transformation due to data missing problems by assigning a population of 1 to the cities that did not exist in each period. The transformation means that these cities have a zero log-population in the periods in which they did not exist. As suggested by Phillips and Sul, the time series data is trimmed based on  $r=0.3$ , i.e. the first 3% of the sample is discarded. The data trimming focuses attention on the latter part of the city sample after 1860. The empirical log  $t$  regression is carried out based on Equation (14). The  $t$ -statistic of the coefficient  $\beta_1$  is estimated as 6.743 at the 1% significant level. Combined with the persistent gaps among city sizes, it implies conditional convergence in the city growth in the upper tail of the US urban system.

#### 4 Conclusions and policy suggestions

In this study, we study the city growth pattern and the related evolution dynamics of city size distribution using a theoretical model and the US data. Our main findings can be summarized as follows.

First, cities grow in a sequential order. The cities with the best economic conditions grow fastest early in the process of urbanization; their growth rates decline when they reach a critical size owing to the escalating diseconomies, and the fastest growth can be found in smaller cities further down in the urban hierarchy. Second, city size distribution evolves along an inverted U-shaped path. In the transition process towards industrialized society, city size distribution evolves from low-level balanced to primate and finally high-level balanced pattern. Third, the structure of urban system is discontinuous. Because of the differences of city growth and economic conditions, the city size distribution will exhibit persistent clustering characteristics and deviation from rank-size rule. Fourth, city size converges conditionally in the upper tail of city size distribution. Because of the facts of sequential city growth and persistent difference in city development conditions, the sizes of cities in the upper tail appear to be conditionally converging over time.

Understanding of the city formation and urban system dynamics is critical to effective policy formulation in developing countries that face rapid urbanization. Disputes such as “focus on large cities” and “focus on small cities” exist throughout a long term in China, and plague policy makers who must determine when and where to improve urban infrastructure investments. Based on the above conclusions, policy suggestions are recommended for developing countries, especially China, which is currently undergoing rapid urbanization.

First, the types of cities supported in priority should be determined based on the stages of regional development. In the early stage of regional development, a priority should be to support the development of large cities, and as regional development enters into late stages, the policy focus should be shifted to the lower-ranked medium-sized cities and finally small cities. At the same time, taking into account large regional differences in China’s urbaniza-

tion (Zhou, 1986; Gu, 2008), the emphasis of urbanization policies in different regions and city types with support priorities should be different, so as to improve the economic efficiency of governmental public investment.

Second, the clustering structure and inverted U-shaped evolutionary path of city size distribution should be fully considered in urban planning practice. This study reveals the sequential pattern of urban growth, and suggests the prediction of the city population with different ranks should not only be based on the development stage of urbanization, but the location conditions of city development, so as to enhance the prediction accuracy of urban population size and effectiveness of urban planning practice.

## References

- Ades A F, Glaeser E L, 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics*, 110(1): 195–227.
- Alonso V O, 2001. Large metropolises in the third world: An explanation. *Urban Studies*, 38(8): 1359–1371.
- Alonso W, 1964. *Location and Land Use*. Harvard University Press.
- Berry B J, 1961. City size distributions and economic development. *Economic Development and Cultural Change*, 9(4): 573–588.
- Brakman S, Garretsen H, Marrewijk C V, 2001. *An Introduction to Geographical Economics: Trade, Location and Growth*. Cambridge: Cambridge University Press.
- Christaller W, 1966. *Central Places in Southern Germany*. Prentice Hall.
- Cuberes D, 2011. Sequential city growth: Empirical evidence. *Journal of Urban Economics*, 69(2): 229–239.
- Dixit A K, Stiglitz J E, 1977. Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3): 297–308.
- Forslid R, Ottaviano G I P, 2003. An analytically solvable core-periphery model. *Journal of Economic Geography*, 3(3): 229–240.
- Fujita M, Krugman P, Moria T, 1999. On the evolution of hierarchical urban systems. *European Economic Review*, 43(2): 209–251.
- Fujita M, Mori T, 1997. Structural stability and evolution of urban systems. *Regional Science and Urban Economics*, 27(4): 399–442.
- Garmestani A S, Allen C R, Bessey K M, 2005. Time-series analysis of clusters in city size distributions. *Urban Studies*, 42(9): 1507–1515.
- Garmestani A S, Allen C R, Gallagher C M, 2008. Power laws, discontinuities and regional city size distributions. *Journal of Economic Behavior & Organization*, 68(1): 209–216.
- Gu Chaolin, Yu Taofang, Li Wangming *et al.*, 2008. *The Pattern, Process and Mechanism of China's Urbanization*. Beijing: Science Press. (in Chinese)
- Henderson J V, 1974. The sizes and types of cities. *American Economic Review*, 6(4): 640–656.
- Henderson J V, 1991. *Urban Development: Theory, Fact and Illusion*. Oxford University Press.
- Henderson J V, Venables A J, 2009. The dynamics of city formation. *Review of Economic Dynamics*, 12(2): 233–254.
- Krugman P, 1991. Increasing returns and economic geography. *Journal of Political Economy*, 99(3): 483–499.
- Krugman P, Livas E R, 1996. Trade policy and the third world metropolis. *Journal of Development Economics*, 49(1): 137–150.
- Lu Yuqi, 2002. The mechanism of the model of dual-nuclei structure. *Acta Geographica Sinica*, 57(1): 85–95. (in Chinese)
- Pflüger M P, 2004. A simple, analytically solvable Chamberlinian agglomeration model. *Regional Science and Urban Economics*, 34(5): 565–573.
- Phillips P C B, SuL D, 2007. Transition modeling and econometric convergence tests. *Econometrica*, 75(6): 1771–1855.
- Silverman B W, 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society*, 43(1): 97–99.
- Tan Minghong, Li Xiubin, 2010. The evolution of the urban system of the United States in the 20th century and its implications for China. *Acta Geographica Sinica*, 65(12): 1488–1495. (in Chinese)
- Walton G M, Rockoff H, 2009. *History of the American economy*. South-Western College Pub.
- Zhou Yixing, Yang Qi, 1986. Review of evolution of city hierarchy in China and their regional classification. *Acta Geographica Sinica*, 41(2): 97–111. (in Chinese)