Journal of Urban Economics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

The settlement of the United States, 1800–2000: The long transition towards Gibrat's law ☆

Klaus Desmet [a,*], Jordan Rappaport [b]

[a] Southern Methodist University, United States
[b] Federal Reserve Bank of Kansas City, United States

ARTICLE INFO

Article history:
Received 2 October 2014
Revised 4 March 2015
Available online xxxx

Keywords:
Gibrat's law
Long-run development
United States
1800–2000
Local growth
Convergence
Divergence

ABSTRACT

Gibrat's law, the orthogonality of growth with initial levels, has long been considered a stylized fact of local population growth. But throughout U.S. history, local population growth has significantly deviated from it. Across small locations, growth was strongly negatively correlated with initial population throughout the nineteenth and early twentieth centuries. This strong convergence gave way to moderate divergence beginning in the mid-twentieth century. Across intermediate and large locations, growth became moderately positively correlated with initial population starting in the late nineteenth century. This divergence eventually dissipated but never completely. A simple-one sector model combining the entry of new locations, a friction from population growth, and a decrease in the congestion arising from the supply of land closely matches these and a number of other evolving empirical relationships.

1. Introduction

Gibrat's law, the orthogonality of growth and initial levels, has long been considered a stylized fact of local population growth (Glaeser et al., 1995; Eaton and Eckstein, 1997; Ioannides and Overman, 2003). This orthogonality is often interpreted as implying that growth is random in the sense that the distribution of local population is not pinned down by exogenous determinants of productivity and quality of life. Orthogonal growth is also frequently cited as an asymptotic explanation for the observed log normal distribution of population across locations and, in the presence of a lower bound on city size, for the observed Zipf's distribution of cities (Eeckhout, 2004; Gabaix, 1999).

In this paper we analyze growth across all counties and metros throughout the entirety of U.S. history and reject that Gibrat's ever held. Instead we find that population growth was strongly negatively correlated with initial population among small locations throughout the nineteenth and early twentieth centuries. This strong convergence gave way to moderate divergence among small locations beginning in the mid-twentieth century, which persisted through 2000. Among intermediate and large locations, population growth also became moderately positively correlated with initial population beginning in the late nineteenth century. This divergence eventually ended among large locations but not among intermediate ones. The U.S. system of locations thus gradually transitioned towards Gibrat's law but never fully attained it.

We hypothesize that the observed convergence of small locations in the earlier period reflects the continual "entry" of new counties into the U.S. system of locations and their subsequent upward transitions to their long-run relative population levels. Over the two hundred years we study, the U.S. continental land area grew from less than 1 million square miles, primarily along the eastern seaboard, to over 7.5 million square miles, coast to coast. Correspondingly, some counties have been settled by Europeans considerably longer than others. We further hypothesize that the observed divergence in the later period represents a decrease in net congestion arising from a shift away from land-intensive production and an increase in the returns to agglomeration. Hansen and Prescott (2002) and Michaels et al. (2012) emphasize the decrease in land congestion associated with the structural transformation away from agriculture during the late nineteenth and early twentieth centuries. Gaspar and Glaeser (1998) and Desmet and Rossi-Hansberg (2009) emphasize

☆ Michael Connolly, Daniel Molling and Li Yi have provided excellent research assistance. We thank Marcus Berliant, Jesus Cañas, Ed Coulson, Marcel Fafchamps, Rafael González-Val, Stuart Rosenthal, Esteban Rossi-Hansberg, Matt Turner and Gavin Wright for helpful comments. We acknowledge the financial support of the Spanish Ministry of Science and Innovation (ECO-2011-27014) and of the Fundación Ramón Areces. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

* Corresponding author.

http://dx.doi.org/10.1016/j.jue.2015.03.004
0094-1190/

the increase in the return to agglomeration arising from the introduction of several general purpose technologies during the twentieth century.

The paper documents five salient empirical relationships consistent with these hypotheses. First, locations of similar age since entry exhibit similar growth patterns independent of calendar year. Growth of young locations is *always* characterized by strong convergence. Growth of old locations is *never* characterized by strong convergence. Second, the rapid growth of newly-entered locations quickly dies out: within 20 years for most and within 60 years for all. Third, convergence completely dissipates by 1940, approximately 20 years after the waning of location entry. Fourth, the subsequent divergence among small locations persists through the end of the twentieth century. Fifth, intermediate and large locations begin to diverge among each other around 1880. This divergence soon ends among the largest locations but persists through 2000 among the intermediate locations.

Informed by these salient relationships, we develop a simple one-sector general equilibrium model of a system of locations that evolves due to entry, decreasing net congestion, and chance. At first only a small share of locations are actually occupied. Over time, the remainder exogenously enter with low initial population. Frictions on positive population growth slow upward transitions and so cause growth from low levels to be characterized by convergence. Overlapping this extended period of entry, net congestion gradually diminishes and so long-run relative population levels become more sensitive to underlying differences in productivity. This introduces a force towards divergence. Once entry is complete and the degree of net congestion stabilizes, the system soon approaches Gibrat's law.

A simulation of the model with just a handful of free parameters tightly matches each of the five evolving empirical relationships described above as well as a number of other evolving relationships. Most model parameters are pinned down to match a specific empirical moment such as aggregate U.S. population and the number of active counties in each decennial census year. Only four parameters retain any freedom to endogenously match more than fifty empirical relationships. These relationships include the distribution of population levels in eleven benchmark years, the distribution of population growth rates over ten twenty-year intervals, the nonlinear correlation between initial population and population growth for ten twenty-year periods, the high persistence of population growth over nine adjacent pairs of twenty-year periods, and the sixty-year growth trajectories of three entering cohorts of locations. Many of these empirical relationships are also matched for young and old sub-samples of locations.

This tight match to a rich set of empirical relationships suggests that the model captures key contours that shaped the system of U.S. locations as it evolved. As with all economic models, ours abstracts from numerous important considerations and developments. Examples include canals, steamships, railroads, the Civil War, electrification, immigration, and the automobile. Each of these played pivotal roles in determining the specific geographical location of U.S. economic activity. But much of the essence that drove overall local growth patterns can be understood in our simple framework.

We are not the first to question Gibrat's law in the context of local development. Michaels et al. (2012) also document a strong positive correlation of growth with population from 1880 to 2000. We differ from them by emphasizing the role of entry and by focusing on the gradually evolving relation between growth and size over the last two centuries. In particular, we find that the changing age composition of locations is a key driver of the evolving balance of convergence and divergence dynamics since 1800. Other papers that have questioned Gibrat's law include Beeson and DeJong (2002), who document the especially rapid population growth by U.S. states following their admission to the union; Holmes and Lee (2010), who divide the U.S. into a grid of six-by-six mile squares and find an inverted-U relation between growth and size from 1990 to 2000; and Dittmar (2011), who shows that orthogonal growth across European cities emerged only in the modern period, after 1500.

These rejections imply that orthogonal population growth cannot be the proximate cause of the level distribution of population across U.S. locations, whether Zipf's or log normal. Consistent with this conclusion, we document that the distribution of population across U.S. counties was already log normal in 1790, an early date to achieve an asymptotic outcome. An interpretation more in line with our findings is that the distribution of local population depends on the unobserved distribution of local productivity and quality of life (Krugman, 1996; Davis and Weinstein, 2002; Rappaport and Sachs, 2003). In a frictionless setting, with a sufficient number of stochastic determinants of local population, the distribution of population will be log normal (Lee and Li, 2013). If these stochastic determinants evolve orthogonally, the population distribution will evolve orthogonally as well. But following entry and other large shocks, frictions can cause population to significantly differ from its long-run distribution and population growth to be correlated with initial population.

Our work is also closely related to the newer literature that analyzes the importance of the age of locations on their growth and size distributions. Sánchez-Vidal et al. (2014) document the faster growth of younger cities in the United States throughout the twentieth century, especially in the first decade following their incorporation. This parallels our finding that the fast-growing small counties throughout the nineteenth and early twentieth centuries were ones that had recently entered the U.S. system. An important endogeneity concern is that counties may enter and cities incorporate when their geographical location has recently been experiencing fast population growth. To address this possible bias, we construct counties with borders from forty years prior to the initial year from which growth is measured. Doing so effects nearly identical results. Giesen and Suedekum (2014) emphasize the effect of city age on the long-run distribution of city sizes. In particular, they document a positive correlation between the cities' population in 2000 and their age. Our model is characterized by a similar correlation, but only during the transition of locations towards their long-run relative population levels.

The rest of the paper is organized as follows: Section 2 describes the data. Section 3 documents the salient empirical relationships described above. Sections 4 and 5 lay out the model and calibrate it. Section 6 presents numerical results. A final section briefly concludes.

2. Data

Our dataset is built using data for county and county-equivalents as enumerated in the 1790 through 2000 decennial censuses (Haines, 2005). During the nineteenth and early twentieth centuries, the number of enumerated counties soared from just under 300 in 1790 to more than 3100 in 1940 (Table 1 column 1).[1]

County borders changed considerably over time. Hence, we use a "county longitudinal template" (CLT) augmented by a map guide to decennial censuses to combine enumerated counties as necessary to create geographically-consistent county equivalents over successive twenty-year-periods (Horan and Hargis, 1995; Thorndale and Dollarhide, 1987). For example, suppose county A

[1] Because our focus is on a system of locations among which there is reasonably high mobility, we exclude Hawaii and Alaska from our sample.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

Table 1
Alternative measures of locations and location entry. Total population is for all continental U.S. counties enumerated in each decennial census. Geographically-consistent counties combine enumerated counties so as to keep borders approximately constant over each twenty-year interval. The hybrid of counties and metros, which is the empirical baseline, additionally combines some geographically-consistent counties into metro areas based on criteria described in the text.

Year	Population (million)	(1)	(2)	(3)	(4)	(5)	(6)
		Counties enumerated in decennial census		Geographically-consistent counties		Hybrid: metros + geographically-consistent counties	
		Number	Change	Number	Change	Number	Change
1790	3.9	293		233		233	
1800	5.3	417	124	310	77	310	77
1820	9.7	762	345	545	235	544	234
1840	17.1	1286	524	870	325	865	321
1860	31.4	2086	800	1706	836	1692	827
1880	50.2	2582	496	2396	690	2369	677
1900	75.6	2841	259	2696	300	2655	286
1920	105.7	3082	241	3014	318	2950	295
1940	131.7	3116	34	3062	48	2982	32
1960	178.5	3121	5	3064	2	2853	−129
1980	225.2	3126	5	3069	5	2631	−222
2000	279.7	3126	0	3069	0	2363	−268

splits into counties B and C in 1850. In this case, we measure the growth of population between county A in 1840 and the combination of counties B and C in 1860. But for 1860–1880, we separately measure the growth of counties B and C. Additional details on how we handle county borders are included in Appendix A.

The resulting geographic adjustments require that we construct a separate dataset for each of the ten twenty-year periods we study (1800–1820, 1820–1840, ..., 1980–2000). For growth between 1800 and 1820, we use geographic borders from 1800; for growth between 1820 and 1840, we use geographic borders from 1820; etc. For the earlier of these ten datasets, the required joins reduce the number of observations by about one third (Table 1 column 3 versus column 1). For the twentieth century datasets, the reduction is relatively modest.[2]

When and where metro areas exist, we argue that they better correspond to geographical markets in which individuals both live and work. Therefore, we use the hybrid of metro areas and remaining geographically-consistent counties as our baseline set of observations (Table 1, column 5). To construct metros in 1960 and 1980, we combine counties using Office and Management and Budget (OMB) delineations respectively promulgated in 1963 and 1983. For years prior to that, we apply the criteria promulgated by the OMB in 1950 to population and economic conditions *at the start* of each twenty-year period (Gardner, 1999). As with the geographically-consistent counties, growth over any period is measured using the geographic borders of the initial year. Additional details are included in Appendix A.

3. Empirical results

This section documents the evolving relationship between population growth and initial population size across U.S. counties and metro areas including the continual rejection of Gibrat's law over 200 years.

We run two types of regressions for each of the ten twenty-year periods, 1800–1820 through 1980–2000. The first are kernel regressions which yield continuous nonlinear approximations of growth versus initial population (Desmet and Fafchamps, 2006). These take the form

$$(L_{i,t+20} - L_{i,t})/20 = \phi_t(L_{i,t}) + e_{i,t}.$$

The term $L_{i,t}$ is the log population for location i in year t. Fitted growth rates give sharp visual comparisons of the relationship between growth and initial population as it evolves over time.

The second regressions take the form of continuous, piecewise-linear splines,

$$(L_{i,t+20} - L_{i,t})/20 = \vec{\beta}_t \cdot \vec{L}_{i,t} + e_{i,t}.$$

The 1-by-k vector $\vec{L}_{i,t}$ includes a constant and a spline of population with $k-1$ segments. The mapping of log population into its vector form is such that the coefficient on each spline segment measures the *marginal* effect of an increase in population size on growth. If growth is orthogonal, coefficients of each of the spline segments should be close to zero.

Because of unobserved characteristics shared by nearby locations, residuals from these regressions are unlikely to be independent. Using a generalization of the Huber–White algorithm, reported standard errors are constructed to be robust to spatial correlation between county pairs with centroids within 200 km of each other (Conley, 1999; Rappaport, 2007).

3.1. Baseline and robustness

As a starting point, we analyze the relationship between size and growth using the baseline hybrid of metros and geographically-consistent counties and metro areas. For each of the ten twenty-year periods spanning 1800–2000, both the kernel and spline regressions reject the null hypothesis of orthogonal growth over a wide range of population levels.

The kernel regressions emphasize this visually (Fig. 1).[3] Displayed growth rates are normalized by subtracting out the aggregate growth rate of locations that were already active at the start of each twenty-year period. If Gibrat's law held, each of the fitted growth lines would be horizontal. Clearly they are not.

The fitted kernels illustrate several of the salient relationships described in the introduction. From 1800 to 1940, the correlation between growth and size for locations with log population up to about 10 is strongly negative. Thereafter, the correlation over this size range is positive. Beginning in 1880, growth is positively correlated with size for log population ranging from 10 to 12. From 1880 to 1920, this positive correlation extends up to log population above 13.

[2] A limitation of the CLT is that it formally tracks county changes back only to 1840. For the 1800 and 1820 datasets, we use the CLT's 1840 borders with a handful of modifications. Calculated growth rates that start in either of these years may be biased downward by failing to account for splits that cede significant land and population.

[3] For a color version of Fig. 1 and all subsequent figures, the reader is referred to the web version of the article.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

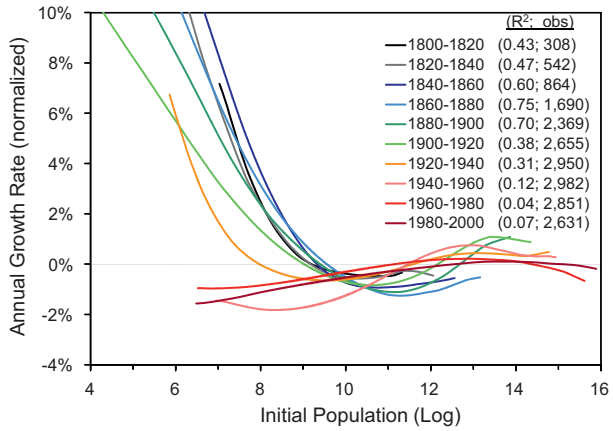

Fig. 1. Population growth versus initial population of metros and counties, 1800–2000. Figure shows fitted growth rates from kernel regressions of population growth on initial log population using the hybrid of metro and county observations. Fitted growth rates are normalized by subtracting the aggregate growth rate of locations active at the start of each 20-year period. In this and all fitted kernels below, the smallest 0.25% of locations and the 2 largest ones have been trimmed from the display.

The spline regressions give similar results (Table 2). For the time periods spanning 1800 to 1920, nearly all coefficients on increments of log population up through 10 are negative and statistically differ from 0 at the 0.05 level. Their magnitudes imply that a 1 log point increase in population is associated with slower annual growth of 1–5 percentage points per year. Beginning in 1880, all coefficients on increments of log population from 10 to 12 are positive, with most statistically differing from 0. Their magnitudes imply that a 1 log point increase in population is associated with faster annual growth of 0.3–1.3 percentage point.[4]

The kernel and spline regressions also suggest that there was some convergence among the largest locations during the mid-twentieth century. In the fitted kernels for the periods beginning in 1940 and 1960, the negative correlations are visible at high populations. The spline regression for each of these periods admits a statistically-significant negative coefficient on a high population bin. We interpret these negative correlations as reflecting measurement error in the sense that they pick up the redistribution of residents *within* labor markets. The forty years from 1940 to 1980 were a period of intense suburbanization. Using metro areas as observations avoids measuring reallocation as growth to the extent that initial geographic borders span the area in which the reallocation occurs. But to the extent that suburbanization spreads outward to counties not initially classified as being part of a metro, measured growth of a metro will be biased downward.

Fitted growth kernels that use only geographically-consistent counties, rather than the hybrid of these with metro areas, accentuate the negative correlation among large locations during these years (Fig. 2, Panel A). With these smaller geographies, all migration from the densely-settled counties in which a core city is located to surrounding suburban counties is measured as differential growth. Otherwise, the pattern of convergence and divergence is nearly identical to that using the hybrid set of observations.

An important endogeneity concern is if counties that were expected to grow fast were split into two or more smaller ones. This would induce the observed negative correlation among small counties during the nineteenth and early twentieth centuries. As a robustness check, we regress population growth on initial population for counties with borders lagged by forty years. In other words, we construct counties that are geographically consistent over sixty years and run the regression of growth on initial population over the final twenty years.[5] The resulting fitted kernels are indistinguishable from those using contemporary borders (panel B versus panel A).

A different possible critique is that the smaller counties exhibiting convergence from 1800 to 1940 were relatively unimportant to aggregate U.S. outcomes. On the contrary, nearly half of the U.S. population lived in counties with log population below 10 in 1800 (population below 22 thousand); more than a quarter still did so in 1900.

3.2. Transitional growth and location age

Our preferred explanation for the observed dynamics laid out above has two parts. First, we hypothesize that the inverse correlation between population size and growth among relatively small locations arises from the transition dynamics of newly-entered locations. During the nineteenth and early-twentieth centuries new locations continually entered the system, typically with low initial population. These newly-entered locations grew faster than average as they transitioned towards their long-run relative population levels. As entry waned over time, such transitional convergence died out. The counties that remained small through more recent time periods are those with low productivity and hence low long-run population levels.

Second, we hypothesize that the positive correlation between size and growth among medium and large locations, which began in the mid-nineteenth century, arises from some combination of a decrease in congestion coming from the fixed supply of land and an increase in agglomeration forces. One possibility is the structural transformation that lowered the importance of land for aggregate production (Michaels et al., 2012). Another is the impact of general purpose technologies, which increased the benefits from agglomeration (Desmet and Rossi-Hansberg, 2009). The resulting impetus toward population divergence applied to all locations, not just large ones. But among small locations, it was masked by the even stronger convergence forces. Then, with the end of location entry early in the twentieth century, this mask was removed and so divergence was observed across all locations.

Our hypotheses on the role of age suggest two predictions. First, growth among locations that are "young" should be characterized by convergence. Second, locations that are "old" should be characterized by divergence. Note that these predictions are not just restatements of the observed correlations between population growth and population size. With sufficiently low entry population and sufficiently high frictions to growth, all young locations will indeed be small. But not all small locations will be young. In particular, locations with low productivity will be small, regardless of age. We find strong evidence supporting each of these predictions.

As a first way of distinguishing between the dynamics of "old" and "young" locations, we compare the relationship between growth and size from 1840 to 1920 among the counties in the land area of the original thirteen colonies and counties in the remaining land area of the U.S. (Fig. 3).[6] We think of 1840 as the earliest starting year for which the first group of counties can be considered old and 1900 as the latest starting year for which the second group of counties can be considered young. Over the years spanning 1840–

[4] Bin increments are chosen such that those with a lower bound log population of 7 or less have a minimum of 40 observations; those with a lower bound of 8 through 10 have a minimum of 20 observations; and those with a lower bound of 11 or higher have a minimum of 10 observations.

[5] Because of the lag and the lower accuracy of the CLT prior to 1840, growth from 1880 to 1900 is the earliest for which we can run this.

[6] The land area of the thirteen colonies included Maine (originally part of Massachusetts), Tennessee and West Virginia (originally part of Virginia), and Kentucky (originally part of North Carolina).

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

Table 2
Population growth versus initial population of metros and counties, 1800–2000. Table shows results from regressing population growth on a linear spline of initial log population. Standard errors in parentheses are robust to spatial correlation. Bold type signifies coefficients that statistically differ from zero at the 0.05 level. The topmost coefficient row corresponds to the lowest log population bin, whatever its range. The bottommost coefficient row corresponds to the highest log population bin, whatever its range.

log(pop) bin:	(1) 1800–1820	(2) 1820–1840	(3) 1840–1860	(4) 1860–1880	(5) 1880–1900	(6) 1900–1920	(7) 1920–1940	(8) 1940–1960	(9) 1960–1980	(10) 1980–2000
Min to lowest lb	**−0.031**	**−0.047**	**−0.053**	**−0.047**	−0.007	**−0.023**	**−0.039**	0.004	0.000	0.001
	(0.003)	**(0.007)**	**(0.007)**	**(0.005)**	(0.024)	**(0.008)**	**(0.006)**	(0.004)	(0.004)	(0.002)
lpop 04to06					**−0.034**					
					(0.008)					
lpop 05to07				**−0.044**						
				(0.006)						
lpop 06to07					**−0.036**					
					(0.012)					
lpop 07to08				**−0.032**	**−0.026**	**−0.025**				
				(0.007)	**(0.007)**	**(0.007)**				
lpop 08to09		**−0.020**	**−0.039**	**−0.025**	**−0.022**	**−0.015**	−0.002	**−0.007**	0.002	**0.005**
		(0.005)	**(0.006)**	**(0.005)**	**(0.003)**	**(0.004)**	(0.002)	**(0.002)**	(0.003)	**(0.002)**
lpop 09to10	0.002	**−0.008**	−0.005	**−0.010**	**−0.010**	**−0.009**	0.001	**0.006**	**0.003**	**0.003**
	(0.004)	**(0.004)**	(0.003)	**(0.002)**	**(0.002)**	**(0.002)**	(0.001)	**(0.001)**	**(0.001)**	**(0.001)**
lpop 10to11	−0.001	**0.010**	0.000	0.001	0.003	**0.010**	**0.006**	**0.011**	**0.004**	0.001
	(0.007)	**(0.004)**	(0.003)	(0.002)	(0.002)	**(0.002)**	**(0.001)**	**(0.001)**	**(0.001)**	(0.001)
lpop 11to12				0.002	**0.009**	0.006	**0.006**	**0.013**	0.002	**0.004**
				(0.008)	**(0.004)**	(0.005)	**(0.002)**	**(0.003)**	(0.002)	**(0.002)**
lpop 12to13							−0.005	**−0.007**	0.000	0.000
							(0.003)	**(0.003)**	(0.002)	(0.002)
lpop 13to14									0.002	0.001
									(0.004)	(0.003)
Highest ub to max	0.002	−0.006	0.002	0.001	0.004	0.002	0.003	0.000	**−0.007**	−0.002
	(0.005)	(0.007)	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	**(0.002)**	(0.002)
Bins	4	5	5	8	9	7	7	7	8	8
N	304	538	860	1685	2354	2648	2943	2979	2849	2630
R^2	0.401	0.478	0.615	0.759	0.654	0.376	0.306	0.128	0.043	0.074

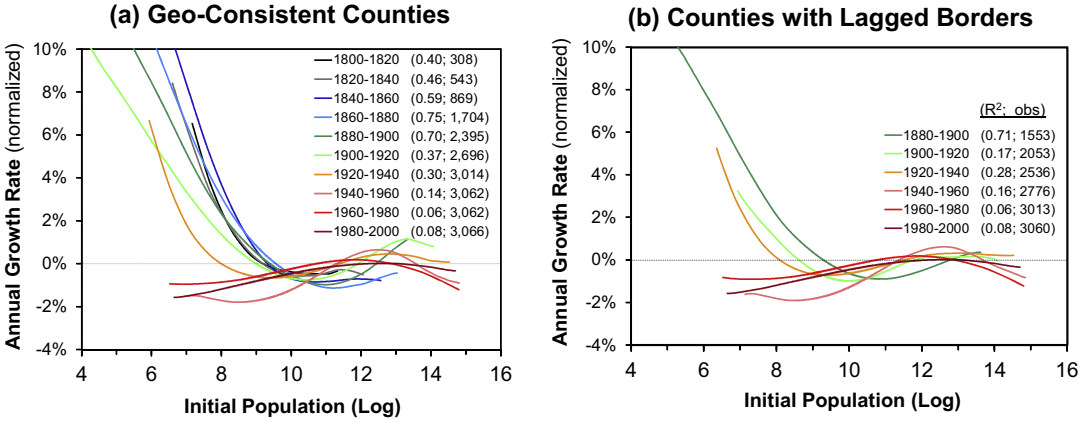

Fig. 2. Population growth versus initial population of counties. Left panel shows fitted growth rates from kernel regressions of population growth on initial log population using geographically-consistent counties as the unit of observation. Right panel shows the same using counties with borders from forty years earlier. The earliest growth interval for which these can be constructed is 1880.

1920, growth by the thirteen colonies counties was mostly flat, with some modest convergence among the smaller ones and some modest divergence among the larger ones (Panel A). In sharp contrast, growth among the remaining counties is characterized by strong convergence (Panel B).[7]

As a second way of distinguishing age, we use an algorithm to classify geographically-consistent counties as "young" and "old". A geographically-consistent county is classified as young if it is in a state or territory that was first enumerated as having at least two counties in the most recent 40 years (which implies that the state or territory entered the U.S. system in the previous 49 years). A geographically-consistent county is classified as old if at least 60 years have passed since any significant geographic changes occurred *and* no further geographic changes occur through 2000. A more detailed description is included in Appendix A.

[7] The divergence among the largest counties not in the thirteen colonies from 1880 to 1920 is driven primarily by those that were settled earliest and so should not be considered to be young. Such counties include Cook County (Chicago) and St. Louis, both of which entered the U.S. system by 1810 according to a criterion described in Appendix A.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

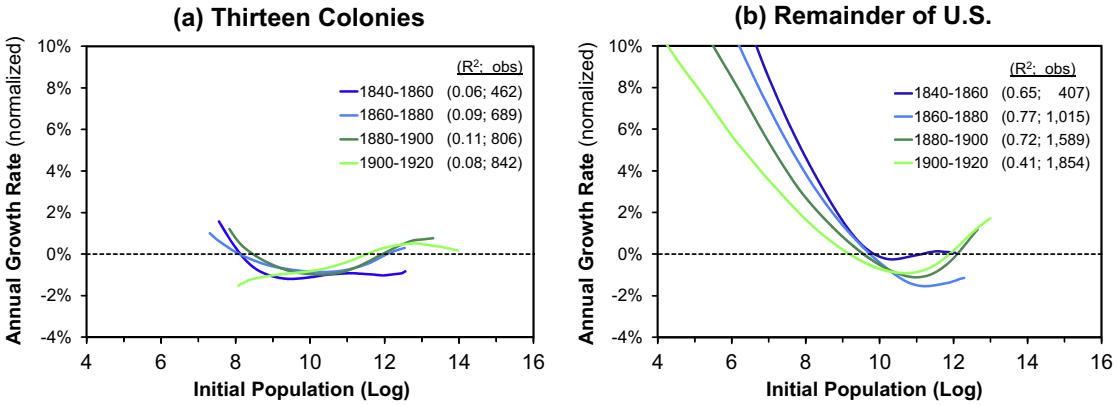

Fig. 3. Growth versus population, 1840–1920: thirteen colonies versus remainder of U.S. Figure shows fitted kernel regressions of population growth on initial log population for geographically-consistent counties in states descended from the original thirteen U.S. colonies versus those in the remaining states. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

The dynamics for this old-young split are similar to those for the split based on location in the thirteen colonies. Over the years spanning 1900–1960, counties classified as old moderately diverged at intermediate population levels (Fig. 4 Panel A). Over the years spanning 1840–1920, the counties classified as young strongly converged at small and intermediate population levels. This difference between old and young is not just driven by the older counties being on average larger. From 1900 to 1920, the only twenty-year period for which we are able to classify some counties in each age category, the young counties converged at intermediate population levels while the old counties diverged at the same levels (Panel C).

Growth trajectories of newly entering counties complement the description of these age-based dynamics. The transition hypothesis implies that growth trajectories by successive cohorts should depend on elapsed time since entry rather than on calendar year. Additionally, growth should be high initially and then decline with elapsed time. We classify a geographically-consistent county as having newly entered the U.S. system if it is in a state or territory that was first enumerated as having at least two counties in the current or previous decennial census (which implies that they entered the U.S. system in the previous 20 years). This is the same criteria for classifying a county as young except with the shorter elapsed time threshold. The relatively small number of counties

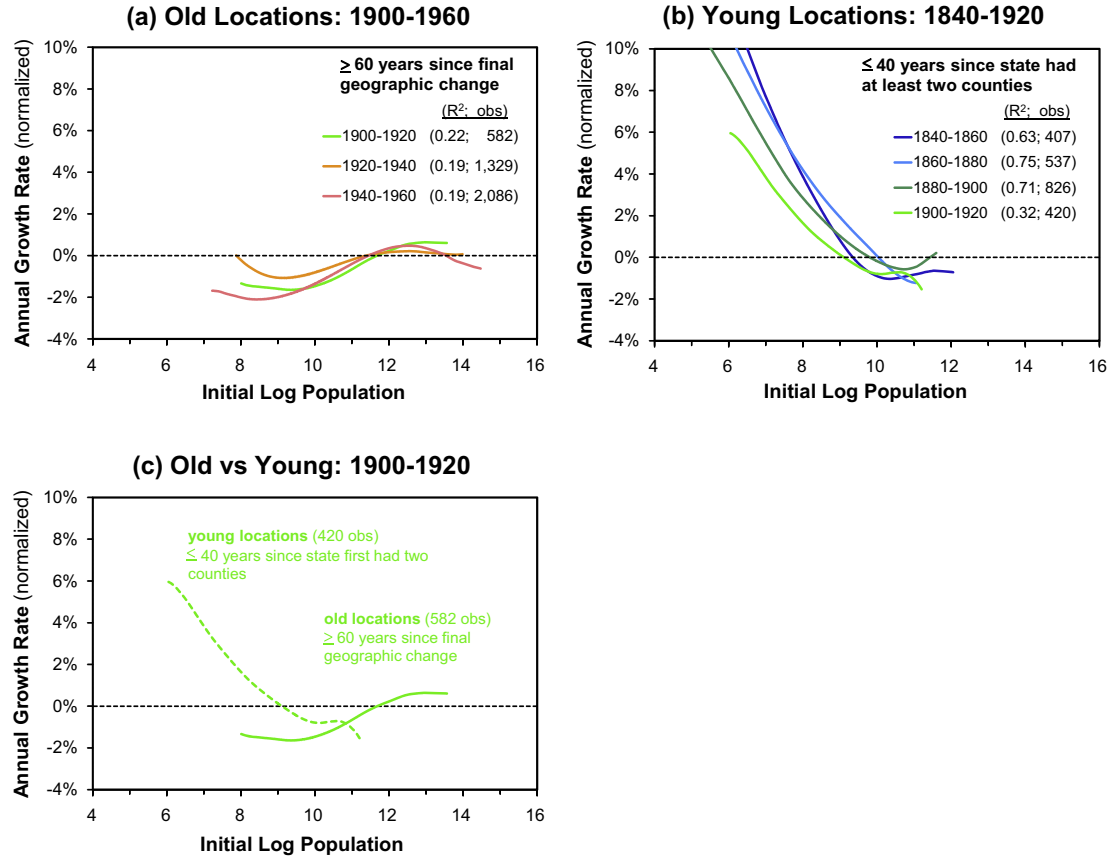

Fig. 4. Population growth versus initial population by location age. Figure shows fitted kernel regressions of population growth on initial log population for sub-samples of counties based on the number of years since they entered the system of U.S. locations. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

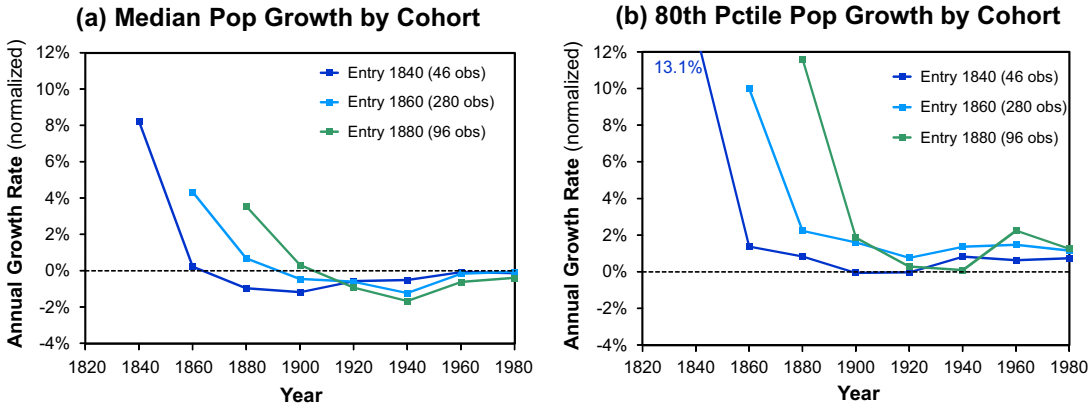

Fig. 5. Population growth by entry cohort. Figure shows growth trajectories for counties that entered the U.S. system of locations in the twenty years prior to each of 1840, 1860, and 1880. The displayed growth rates are the median and the 80th percentile for each cohort over each 20-year period.

that meet this tight entry criterion—552 over the near 200 year time span—emphasizes that many actual entries are not being classified as such. As a result, a sufficient number of entrants to construct growth trajectories is available only for 1840, 1860, and 1880.[8]

Both implications for growth trajectories are borne out empirically. Growth trajectories for the 1840, 1860, and 1880 entrant cohorts are characterized by nearly identical, high initial growth that dissipates over time (Fig. 5). For each cohort, median initial growth is at least 4% points above aggregate growth (Panel A). It approximately equals aggregate growth during the second twenty-year period following entry, and is below aggregate growth during the third and fourth twenty-year periods. The below-aggregate growth during the later periods following entry reflects that more recent entry cohorts "steal" population from older ones, which is consistent with the old-versus-young convergence results.

Transitions based on the 80th percentile growth rate within each cohort take somewhat longer: between 40 and 60 years (Panel B). This makes sense since the initially faster-growing locations are likely to be the ones that have the furthest to grow to attain their expected long-run relative population level. These faster growth trajectories also establish that the high convergence rates among small locations primarily arise from the early years of transitions. By the second twenty-year period, the 80th percentile growth rate is only 2% above aggregate growth. The high cross-sectional convergence observed over 140 years must therefore arise from successive waves of entry rather than from the fast growth of entering locations having a long duration.

A closely-related implication is that the distribution of population *levels* of younger locations should be shifted to the left relative to that of non-young locations, whereas the distribution of population *growth rates* of young locations should be shifted to the right relative to that of non-young locations. Additionally, the young population *level* distribution should be skewed to the left and the young *growth* distribution should be skewed to the right. All of these predictions are borne out in the data.[9]

3.3. Population growth persistence

The model we develop below critically assumes the existence of a growth friction that slows down transitions, and hence causes population growth to be persistent over time. For the most part, such persistence also characterized historical growth of U.S. counties and metro areas. Fig. 6 shows fitted splines from regressing growth during a "second" twenty-year interval on growth during a lagged "first" twenty-year interval. Among locations that experienced positive growth during the first interval, growth between the second and first intervals was always positively correlated. Throughout most of the twentieth century, growth was similarly persistent among locations that experienced negative growth during the first interval. But during the nineteenth century, population often bounced back from contractions.

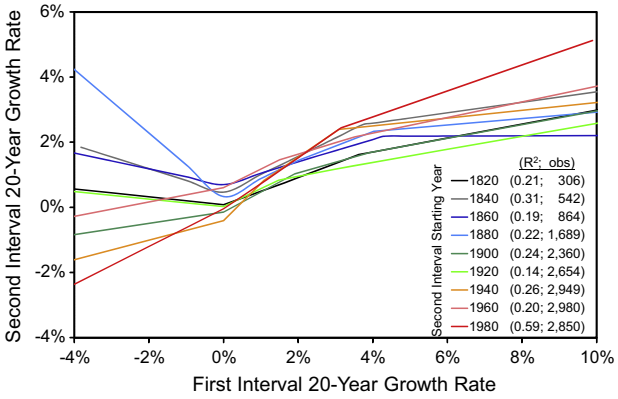

Fig. 6. Persistence of population growth. Fitted values from regressing county/metro population growth (not normalized) over twenty years on a four-way spline of population growth over the previous twenty years. Enumerated years are the start of the second twenty-year period.

The strong persistence of positive population growth among counties and metros contrasts with relatively weak persistence of positive population growth among U.S. municipalities from 1970 to 2000 (Glaeser and Gyourko, 2005). The difference reflects that positive population growth often takes place at the periphery of densely-settled locations and so may not be measured by the positive growth of existing municipalities.

4. Model

Informed by our empirical findings, we develop a simple one-sector general equilibrium model of a system of locations transitioning towards a balanced growth path. New locations exogenously enter the system over time with productivity that evolves stochastically. Agents are perfectly mobile, but a friction from positive population growth dampens a location's productivity. As the system of locations transitions towards balanced growth, local population growth rates approach Gibrat's law.

[8] We do not to construct growth trajectories for the 1820 entry cohort because of the lower accuracy of the CLT for that year.

[9] Appendix Figs. D.3 and D.2.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

4.1. Locations, endowments, and preferences

The economy consists of N potential locations indexed by i. At time t, a number $N_t \leqslant N$ of locations are active. The timing of each location's activation is exogenous and the order of entry is random. Once active, a location remains so forever after. Each is endowed with an identical amount of land, D. Aggregate population of the system of active locations, L_t, grows at an exogenous rate λ_t,

$$L_t = (1 + \lambda_t)L_{t-1}. \tag{1}$$

Individuals supply one unit of labor where they live implying that a location's population and its labor input are both given by $L_{i,t}$. For present purposes, land ownership need not be specified other than the requirement that agents' receipt of land income does not depend on where they live. Agents are freely mobile across locations. They maximize the present discounted value of utility, $\sum_{s=t}^{\infty}\beta^{s-t}u(c_s)$, where c_s denotes consumption in period s and $u(c_s)$ is the one-period utility flow.

4.2. Production

Perfectly competitive firms produce an identical non-storable good using Cobb-Douglas technology. Total output in each location is a function of its time-specific total factor productivity, its labor, and its land:

$$Y_{i,t} = Z_{i,t} \cdot L_{i,t}^{1-\alpha_t} \cdot D^{\alpha_t}.$$

Factors are paid their marginal product and so a location's wage is given by

$$w_{i,t} = (1 - \alpha_t) \cdot Z_{i,t} \cdot \left(\frac{L_{i,t}}{D}\right)^{-\alpha_t}. \tag{2}$$

Total factor productivity is the product of four components: a starting productivity draw, the accumulation of idiosyncratic shocks, an agglomeration factor, and a discount term arising from growth frictions:

$$Z_{i,t} = Z_i^0 \cdot \prod_{s=1}^{t} Z_{i,s}^{\text{idio}} \cdot Z_t^{\text{irs}}(L_{i,t}) \cdot Z^{\text{fric}}(\Delta L_{i,t}/L_{i,t-1}). \tag{3}$$

Potential locations draw their starting productivity from a log normal distribution, $\log Z_i^0 \sim \mathcal{N}(0, \sigma^0)$. In each period thereafter, both active and inactive locations draw an idiosyncratic shock $Z_{i,s}^{\text{idio}}$ from a second log normal distribution $\log Z_{i,s}^{\text{idio}} \sim \mathcal{N}(0, \sigma^{\text{idio}})$.[10] The agglomeration factor assumes that TFP endogenously increases as population increases with elasticity $\varepsilon_t \geqslant 0$,

$$Z_t^{\text{irs}}(L_{i,t}) = L_{i,t}^{\varepsilon_t}. \tag{4}$$

Net congestion, $\hat{\alpha}_t \equiv \alpha_t - \varepsilon_t$, is assumed to gradually decrease over time in order to generate the population divergence among older locations. This assumed decrease can equivalently arise from a decrease in the land income share, α_t, or an increase in the agglomerative elasticity, ε_t.

The growth discount can be rewritten in terms of a realized friction, $G(\cdot)$,

$$Z^{\text{fric}}(\Delta L_{i,t}/L_{i,t-1}) = 1 - G(\Delta L_{i,t}/L_{i,t-1}).$$

The realized friction equally lowers the wages of all workers in a location, not just those of migrants. This is meant to capture, in a reduced form way, either the negative externalities from the congestion of public goods (Dinkelman and Schulhofer-Wohl, 2013) or the slow buildup of capital in response to population inflows (Barro et al., 1995; Rappaport, 2005). In either case, wages are the same for all workers in a location, regardless of when they arrived. This assumption has the benefit of greatly simplifying the forward-looking nature of migration. Similar quantitative results can be attained in a forward-looking model in which utility flows converge slowly over time, and wages–measured in terms of a traded numeraire good–permanently differ across locations due to variations in productivity (Rappaport, 2004).[11] In the numerical implementation we use the following functional form,

$$G(\Delta L_{i,t}/L_{i,t-1}) = \begin{cases} \min(\xi_1(\Delta L_{i,t}/L_{i,t-1})^{\xi_2}, 1) & \text{if } \Delta L_{i,t} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\xi_1 \geqslant 0$ and $\xi_2 > 0$. This specification implies that $G \in [0, 1]$, and that its first derivative is weakly positive, $G'(\cdot) \geqslant 0$.

4.3. Equilibrium

In principle, the decision to migrate is forward-looking because it requires agents to compare the future time path of utility flows across all locations. In the present setup, however, the decision simplifies to a static one. Where an agent chooses to live does not affect her future utility flows. The growth friction affects all local residents rather than just migrants and so wages are equalized *within* locations. In addition, we assume that any factor income from land ownership does not depend on where an individual currently or previously lived. This allows us to abstract from the dynamic decision on whether to save via land ownership. In consequence, free mobility equalizes wages *across* locations. We are now ready to define the economy's *dynamic equilibrium*, which is a series of static equilibria that are linked through the laws of motion for local population and the number of locations.

Definition of Dynamic Equilibrium. A dynamic equilibrium is a sequence of a number of active locations $\{N_t\}$ and of location-specific variables $\{L_{i,t}, Z_{i,t}, w_{i,t}\}$ that satisfy: (i) free mobility, $w_{i,t} = w_t$; (ii) the sum of population across all locations equals aggregate population $\sum_1^{N_t} L_{i,t} = L_t$; (iii) labor market clearing; (iv) goods market clearing; (v) the exogenous laws of motion that determine the set of locations, N_t, and the aggregate population, L_t.

Conditions (i)–(iv) define each period's static equilibrium, whereas condition (v) states the laws of motion of locations and population that link the sequence of static equilibria, thus defining the dynamic equilibrium. The dynamic equilibrium is unique so long as agglomeration forces are not "too strong". Specifically,

Proposition 1. *The dynamic equilibrium is unique as long as* $\alpha_t > \varepsilon_t$.

Proof. Given the parameter restriction, $w_{i,t}$ will be strictly decreasing in $L_{i,t}$. For any common wage level, w_t, there will be a unique $L_{i,t}$ that satisfies it. Let $\bar{w}_t$ represent a one-period equilibrium common

[10] In the model developed by Giesen and Suedekum (2014), locations draw their productivity upon entering from a constant distribution, regardless of when their entry occurs. Thereafter, productivity evolves stochastically with a positive drift. As a result, older locations are more productive on average and so population and age are positively correlated.

[11] Rappaport, 2004 includes non-traded land services as a component of utility and assumes a growth friction that is borne solely by migrants. Following a positive shock to a location's productivity, the migration friction causes an extended transition during which the utility flow enjoyed by the location's residents exceeds that available elsewhere. The eventual equalization of utility flows is achieved in part by an increase in the price of land services, thereby allowing the location's wage denominated in terms of a traded numeraire good to exceed the wage available elsewhere. Consistent with this numerical result, considerable empirical evidence shows that income convergence occurs gradually over several decades (Barro and Sala-i-Martin, 1991; Gennaioli et al., 2014). In particular, real wages in the western U.S. remained substantially higher than in the remainder of the country during the nineteenth century (Easterlin, 1960; Rosenbloom, 1990; Mitchener and McLean, 1999).

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

wage. For any $w_t < \bar{w}_t$, the sum of local populations will exceed the assumed aggregate population. For any $w_t > \bar{w}_t$, the sum of local populations will fall short of the assumed aggregate population. This argument holds for each time period. □

To characterize the long-run behavior of our system of locations, we start by showing that in the absence of idiosyncratic shocks or in the absence of growth frictions, the economy converges to a balanced growth path, where the expected growth rates of all locations are identical. This is equivalent to the economy attaining Gribrat's law. We then discuss what happens when there are both idiosyncratic shocks *and* growth frictions.

Proposition 2. *Assume that $\alpha > \varepsilon$, all potential locations are active, and all parameters are constant. Then in the absence of idiosyncratic shocks or in the absence of growth frictions, the economy converges to a balanced growth path characterized by population in all locations growing at the same expected rate.*

Proof. See Appendix B. □

When there are both idiosyncratic shocks and growth frictions, as in the model, the system does not strictly converge to Gibrat's law. Rather, there remains a slight positive correlation between size and growth. For the intuition, consider a system with frictions but no shocks. In accordance with Proposition 2, this system will eventually converge to a balanced growth path in which population growth is identical across locations. Now let each location experience a one-time shock to its productivity. Because of the growth friction, locations that experience a positive shock will converge only gradually upward to their new balanced-growth relative population level. For general-equilibrium reasons, locations that experience a negative shock will converge only gradually downward to their new balanced-growth relative population level. Hence, during the transition, positively-growing locations will on average be slightly larger than negatively-growing ones. But for a wide range of calibrations, including the baseline one used below, this deviation from orthogonality proves to be second order.

5. Numerical calibration

The model depends on a few key parameters, summarized in Table 3. Most of these are parameterized based on predetermined criteria rather than on achieving a good fit to data. In what follows we discuss our main calibration choices.

Aggregate population growth is calibrated to match total U.S. population in each decennial year. Location entry is calibrated to match the number of geographically-consistent counties in each decennial year (Table 1 column 3).

The distribution of population depends closely on net congestion, $\hat{\alpha}_t \equiv \alpha_t - \varepsilon_t$. Given the equilibrium requirement that wages must be identical across locations, we can substitute (4) and (3) into (2) to obtain the key calibration equation

$$\log L_{i,t} = (1/\hat{\alpha}_t) \cdot (\log \widehat{Z}_{i,t}) + \text{constant}_t \tag{6}$$

where $\widehat{Z}_{i,t}(\cdot) \equiv Z_i^0 \cdot \prod_{s=1}^{t} Z_{i,s}^{\text{idio}} \cdot Z^{\text{fric}}(\cdot)$ can be thought of as TFP net of agglomerative forces. One important implication of (6) is that the dispersion of population across locations is inversely proportional to $\hat{\alpha}_t$. Hence, as $\hat{\alpha}_t$ decreases over time, population dispersion increases.

Contingent on the value of net congestion, the standard deviation of log time-zero TFP, σ_{Z^0}, is calibrated to match the observed standard deviation of log population in 1790. Because of this joint determination, the actual choice for $\hat{\alpha}_{1790}$ does not affect any outcome. Instead, an increase in net congestion (which implies less variation in population) is exactly offset by an increase in the dispersion of Z^0. We nevertheless use empirical estimates of land's historical factor income share to set $\hat{\alpha}_{1790}$ to 0.15 (see Appendix C).

In contrast, *proportional* changes to net congestion over time, as measured by $\Delta(\hat{\alpha}_t)/\hat{\alpha}_t$, are a key determinant of outcomes. In the baseline calibration, $\hat{\alpha}_t$ is assumed to slowly decline from 0.15 in 1840 to 0.10 in 1960 and otherwise remains constant. For a given initial standard deviation of location TFP, this one-third proportional decrease in net congestion causes the standard deviation of log population to proportionally increase by one-half (0.15/0.10). To the extent that the decrease in $\hat{\alpha}_t$ arises from a

Table 3
Model calibration. Sensitivity description is for a moderate change to a calibrated parameter.

Variable	Interpretation	Value	Source/rationale	Sensitivity
L_t	Aggregate population	U.S. aggregate	Decennial census	Low
N_t	Active locations	Geographically consistent counties beginning in 1790 (Table 1 column 3)	Decennial census; county longitudinal template (described in text)	High
$\hat{\alpha}_t \equiv \alpha_t - \varepsilon_t$	Net congestion (land factor share minus TFP elasticity)			
Initial level (t = 1790–1840)		0.15	Normalization; consistent with estimates	None
Proportional decrease during transition (t = 1840–1960)	Technological change: decreasing land intensity of production and/or increasing return to agglomeration	$\hat{\alpha}_{1960}/\hat{\alpha}_{1840} = 2/3$ decline is exponential	Size of decrease is calibrated to match observed divergence. Dates correspond to the acceleration and deceleration of the shift in population from rural to urban	High
Final level (t = 1990–2000)		0.10	Implied by above; consistent with estimates	
$F(Z_i^0)$	Distribution of starting TFP	Lognormal	Eeckhout (2004)	
$\sigma(\log z_i^0)$	Std dev of log (starting TFP)	0.834 = $\hat{\alpha}_{1790} \cdot \sigma(\log L_{i,empirical1790})$	Calibrated to match std dev of log(pop) in 1790	Moderate
$F(Z_{i,t}^{\text{idio}})$	Distribution of TFP shock	Lognormal	Eeckhout (2004)	
$\sigma(\log Z_{i,t}^{\text{idio}})$	Std dev of log(TFP shock)	0.004	Residually calibrated to match std dev of log(pop) in 2000	Low
ξ_1	Growth friction, level	such that G(·) = 0.06 at 4% growth	Calibrated (by grid search)	Low
ξ_2	Growth friction, convexity	0.84	Calibrated (by grid search)	Low
$\tilde{L}_t$	Pre-entry population	500	Calibrated (by grid search)	Moderate

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

decrease in the land share of factor income, α_t, a one-third decrease to an ending value of 0.10 is consistent with empirical evidence (Caselli and Coleman, 2001; Mundlak, 2001; see Appendix C). The start and end years for this transition correspond to a significant acceleration and then deceleration of the shift of the U.S. population from rural to urban locations.

The standard deviation of the idiosyncratic shock, $\sigma(Z_i^{\text{idio}})$, is calibrated so that, contingent on the assumed decrease in net congestion, the standard deviation of log population in 2000, $\sigma(\log L_{i,2000})$, matches its empirical value. Unsurprisingly, empirical population levels are considerably more diffuse in 2000 than in 1790. The assumed decrease in net congestion over time is able to achieve nearly three-quarters of the increase in population dispersion. The remaining increase in population dispersion is matched by the cumulative effect of the 210 annual idiosyncratic shocks beginning in 1791.[12]

A final set of three parameters jointly govern the friction that slows positive population growth. The parameters ξ_1 and ξ_2 specify the level and convexity of the growth friction. In addition, newly-entering locations can join the system with a population of $\widetilde{L}$ or less without incurring any frictional discount to their TFP. Thus $\widetilde{L}$ serves as a proxy for pre-entry population. Using a grid search over combinations of the three friction parameters to fit a large set of observed growth moments, the friction level, ξ_1, is set such that population growth of 4% decreases local TFP by 6%; the convexity of the friction, ξ_2, is set to 0.84 (moderately concave); and the pre-entry population proxy, $\widetilde{L}$, is set to 500 ($\log \widetilde{L} = 6.2$). These level and convexity parameters imply a degree of labor mobility consistent with empirical estimates in Gallin (2004) and Rappaport (2004).

6. Numerical results

The calibrated model tightly matches the evolving relationship between the growth and size of U.S. counties and metros throughout the nineteenth and twentieth centuries as well as a number of other evolving empirical relationships described above and below.

6.1. Simulated transitional growth

Paralleling the empirical analysis, we regress simulated population growth on a spline of simulated initial population. For each of the ten twenty-year intervals, we repeat this regression for each of 400 stochastic seeds.

The pattern of simulated growth over 200 years closely matches that of empirical growth (Fig. 7 versus Fig. 1). During the nineteenth century, simulated growth is characterized by significant convergence at low population levels as entering locations transition upward. Net congestion is assumed to begin declining in 1840. But the implied divergence is initially masked by upward transitions. Beginning in 1880, the divergence shows through at higher population levels. Beginning in 1940, it shows through at lower population levels as well.

Alternative simulations make clear that the end of convergence in 1940 reflects the dwindling of location entry during the 1920s and 1930s rather than the sharp declines in entry during earlier decades. After net congestion stabilizes in 1960, divergence rapidly dissipates. From 1980 to 2000, growth is approximately orthogonal.

The unusually tight match of simulated to empirical convergence and divergence is illustrated in a period-by-period comparison of fitted growth rates (Fig. 8). The modest discrepancies include that simulated convergence from 1880 to 1920 ends at a lower population, that simulated divergence weakens from 1960 to 2000 whereas empirical divergence strengthens, and that simulated divergence from 1940 to 1960 is spread more evenly across the size range than is empirical divergence.

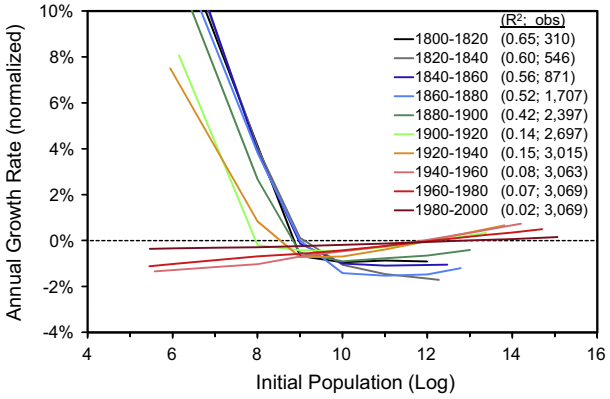

Fig. 7. Simulated population growth versus initial population, 1800–2000.

6.2. Simulated transitional growth and location age

The dependence of simulated convergence and divergence on location age is similar to the empirical dependence on age (Fig. 9 versus Fig. 4). Growth of young locations, those that entered the system in the previous 40 years, is primarily characterized by convergence (Panel B).[13] Growth of old locations, those that have been active for at least 60 years, is exclusively characterized by divergence (Panel A). This co-existence of divergence by old locations and convergence by new locations holds continuously from 1840 to 1940.

Simulated growth trajectories approximately match empirical ones (Fig. 10 versus Fig. 5). Both median and 80th-percentile simulated growth rates are moderately slower during the initial twenty years. And the duration of transitions is about 20 years shorter.

6.3. Simulated growth and level distributions

In addition to the five salient empirical relationships we emphasize in the introduction, the model approximately matches the evolving empirical distributions of population levels and growth rates, both when considering all locations and when splitting by age.

Fig. 11 compares the simulated and empirical growth distributions of all locations during four selected twenty-year periods. For the two nineteenth century time periods, both the simulated and empirical growth distributions are strongly skewed to the right. From 1920 to 1940, both growth distributions are moderately skewed to the right. From 1980 to 2000, the match is less tight. The simulated distribution is less dispersed around a mode that is shifted moderately to the right. Partly this reflects that the simulated small locations are no longer diverging whereas the empirical small locations continue to do so. The approximate symmetry of the simulated distribution also contrasts with some remaining right skew to the empirical distribution.[14]

[12] We could eliminate the idiosyncratic shocks, while still matching the increased dispersion in population, by assuming a slightly larger decrease in net congestion.

[13] The modest divergence by the largest young locations reflects that simulated upward transitions are a bit faster than empirical ones. In consequence, high-productivity locations grow sufficiently close to their balanced-growth population within 40 years for divergence to dominate during the subsequent 20 years.

[14] Appendix Fig. D.1 shows comparisons of simulated and empirical growth distributions for the remaining six twenty-year periods.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

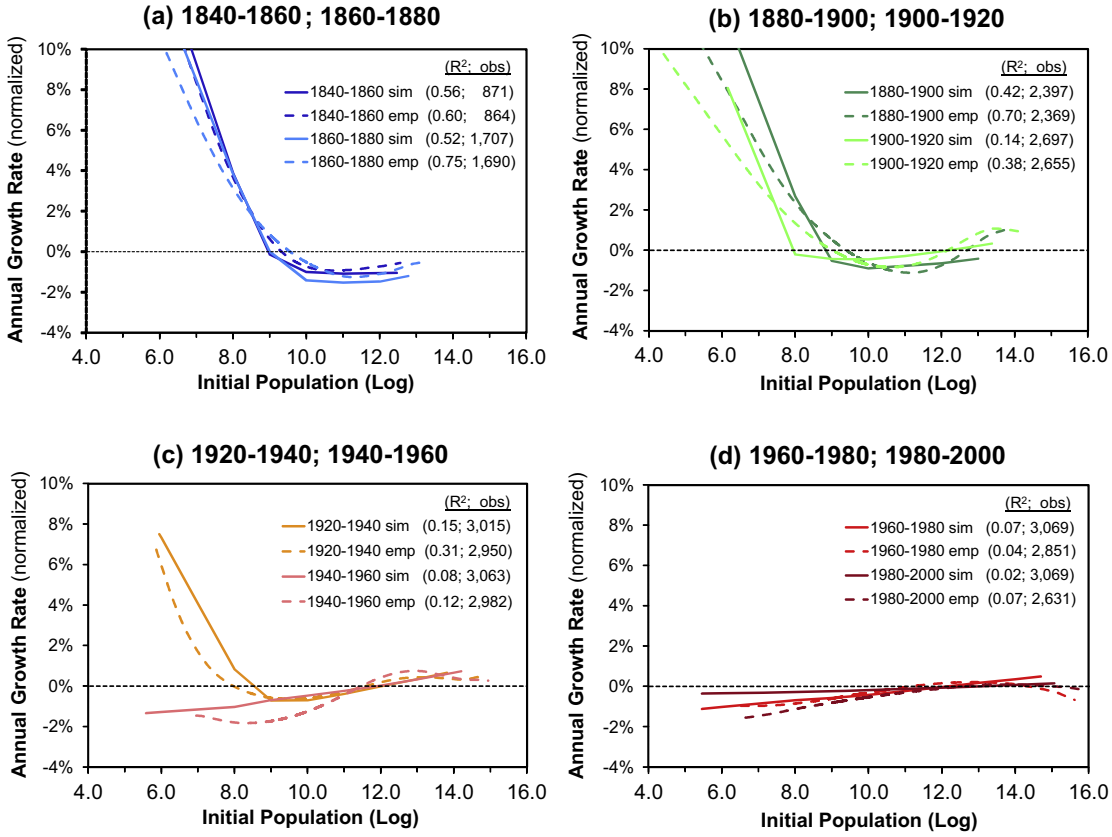

Fig. 8. Fitted population growth: simulated versus empirical. Population growth rates are normalized by subtracting the aggregate growth rate of locations existing locations at the start of each twenty-year period. *R*-squared values are shown in parentheses.

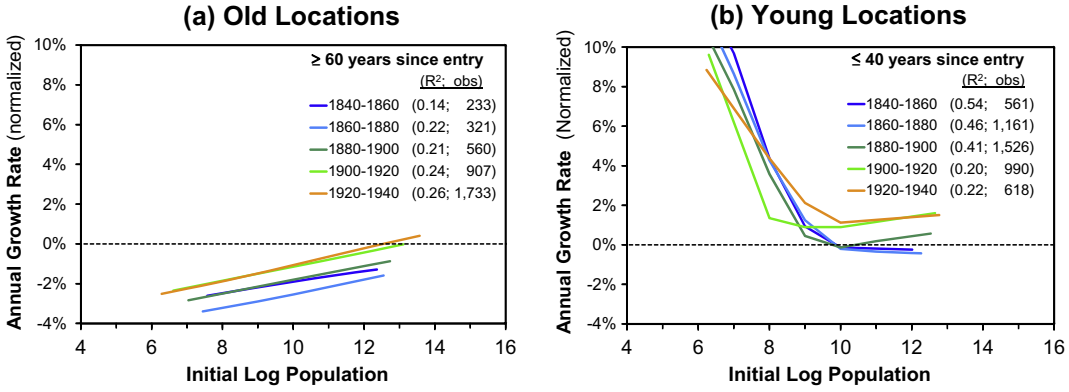

Fig. 9. Simulated population growth versus initial population by location age. Figure shows fitted spline regressions of population growth on initial log population for subsamples of locations based on the number of years since they entered the modeled system of locations. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

Fig. 12 compares the simulated and empirical level distributions of population in four representative years.[15] We start the simulation in 1790 with a frictionless equilibrium and so the simulated level distribution in that year is log normal by construction (Panel A). As it almost exactly overlays the 1790 empirical distribution, the latter must be log normal as well. Achieving log normality at such an early date would seem an unlikely asymptotic result.

The simulated level distribution for 1860 is characterized by a bulge in density at population levels moderately below the mean (Panel B). This captures the upward transition of young locations. Alternatively allowing for some positive correlation between entry size and productivity would dampen this hump. Doing so would also lessen the thickness of the right-tail *growth* rates during the nineteenth century (Fig. 11, Panels A and B). For 2000, the simulated population level distribution is very close to log normal as the transitions from entry and the decrease in net congestion are complete (Panel D). In contrast, the empirical distribution is modestly right skewed.

The relatively close matches between simulated and empirical population growth and level distributions, each split by age, are illustrated by figures in Appendix D.

[15] For comparability to the simulated results, the empirical distributions are for geographically-consistent counties rather than the hybrid of these with metro areas.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

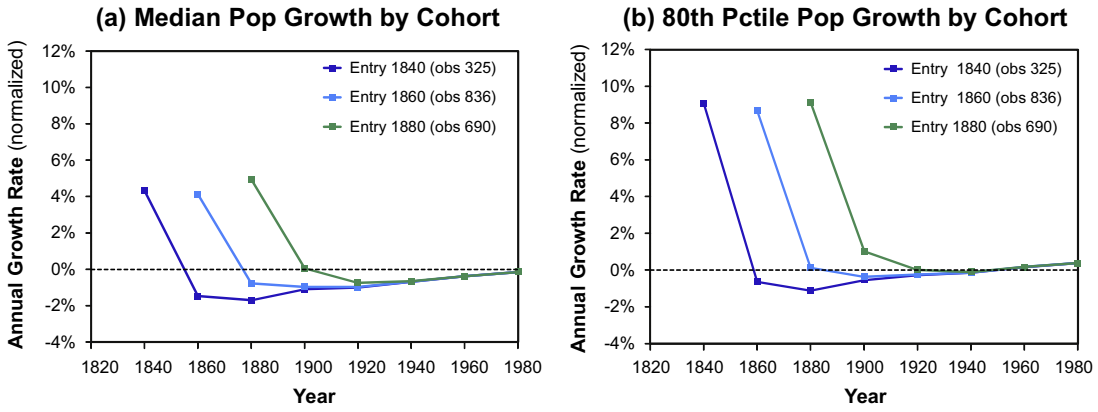

Fig. 10. Simulated population growth by entry cohort. Figure shows growth trajectories for counties that became active in the twenty years prior to each of the listed entry years. The displayed growth rates are the median and the 80th percentile for each cohort over each 20-year period.

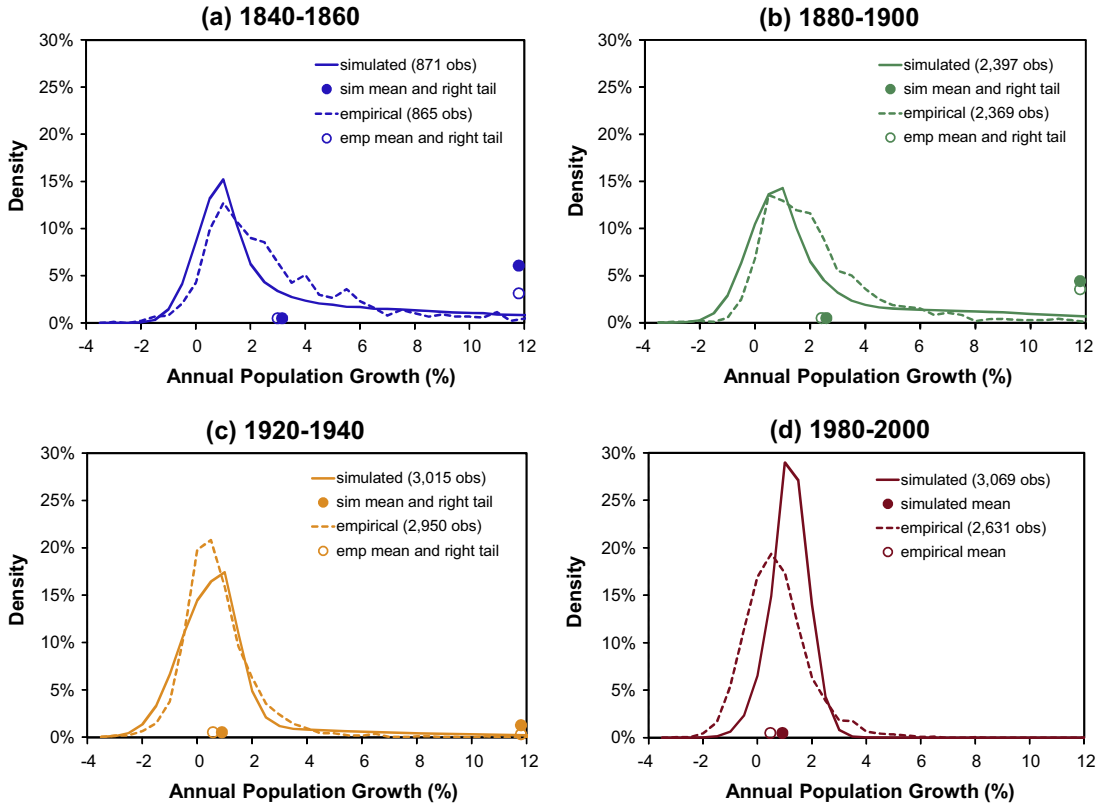

Fig. 11. Population growth distributions, simulated versus empirical. Growth rates are not normalized. The right tail is calculated as the cumulative density of locations with growth above 12%.

6.4. Alternative scenarios

The model's success in matching such a wide range of evolving empirical relationships over 200 years suggests that alternative scenarios can give insight into underlying mechanics. How would an early end to the U.S. westward expansion have affected population dynamics? What would U.S. development have looked like if frictions had been much smaller or larger? What would recent and future growth be if the decrease in net congestion had resumed?

6.4.1. Early end to location entry

The discussion so far has emphasized the key role of location entry in driving aggregate dynamics. Once entry ends, convergence ends soon after. Consider the alternative scenario under which entry matches the baseline until 1860 and then ends. In this case, convergence among small locations from upward transitions continues to dominate the divergence from decreasing net congestion from 1860 to 1880. But any residual convergence thereafter is completely masked by divergence. Essentially, the switch from net convergent growth to net divergent growth among smaller locations occurs 60 years earlier than under the baseline.

6.4.2. Growth friction

The U.S. is generally thought to have been characterized by relatively high labor mobility and so, implicitly, by low growth frictions. As an extreme, the model can be calibrated to be frictionless.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

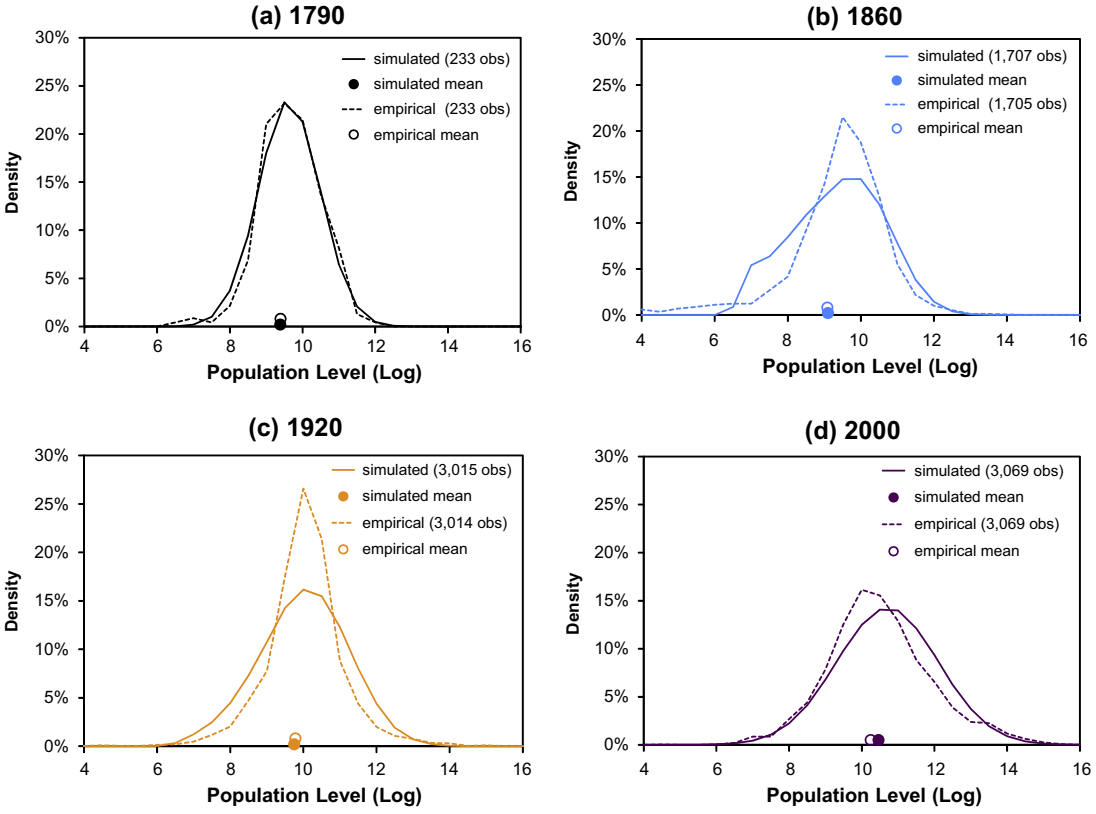

Fig. 12. Population level distributions, simulated versus empirical. Empirical distributions are for geographically-consistent counties rather than the hybrid of these and metro areas.

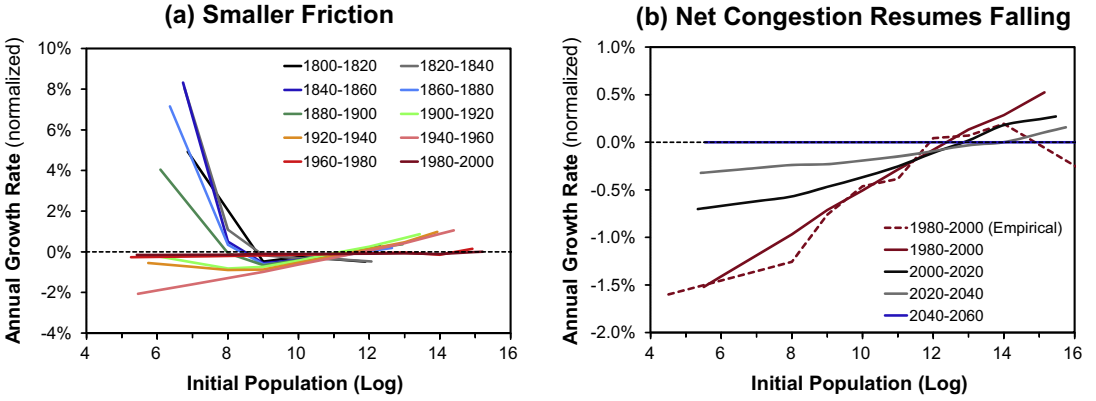

Fig. 13. Alternative scenarios. Panel A: growth friction equal to 1% of TFP rather than the baseline 6% at 4% annual growth. Panel B: Net congestion decreases from $\hat{\alpha} = 0.100$ in 1980 to $\hat{\alpha} = 0.095$ in 2000.

In this case, locations immediately jump to their relative population along a balanced growth path and so there is no observed convergence. Divergence ends in 1960, concurrent with the stabilization of net congestion and twenty years earlier than under the baseline. With a slight positive growth friction (a loss of productivity at 4% annual growth of 1% rather than 6%), net convergence ends in 1900, forty years earlier than under the baseline; divergence ends in 1960, twenty years earlier than under the baseline (Fig. 13, Panel A). With a friction significantly higher than under the baseline (a loss of productivity at 4% annual growth of 20%), net convergence ends in 1960 and divergence ends in 2000, both of which are twenty years later than under the baseline.

6.4.3. Decrease in net congestion from 1980 to 2000

Desmet and Rossi-Hansberg (2009) argue that the rapidly expanding output share of information and communications technology beginning in the mid-1970s was akin to the introduction of a general purpose technology that increased the return to agglomeration. We model this as resumption of the decline in net congestion from 1980 to 2000. Specifically, we assume that $\hat{\alpha}$ gradually declines from 0.100 to 0.095. For perspective, this 5% decrease over twenty years corresponds to a more moderate rate of decline than the baseline one-third decrease over 120 years. The implied increase in divergence from 1980 to 2000 closely matches empirical growth (Fig. 13 Panel B, dashed line and solid line for

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

1980-2000). Divergence then diminishes over the subsequent forty years with orthogonal growth resuming in 2040.

7. Conclusions

This paper has studied the long-run development of the U.S. system of locations. Throughout the nineteenth and early twentieth centuries, population growth among young, smaller locations was characterized by strong convergence as newly entered locations rapidly transitioned up to their long-run relative population. As the entry of new locations dwindled during the early twentieth century, this convergence gave way to moderate sustained divergence across small locations. Beginning in the late nineteenth century, growth across older, intermediate and larger locations became characterized by moderate divergence. This divergence soon ended among the larger locations but persisted through 2000 for the intermediate locations. In consequence, the orthogonality of local population growth to initial population levels is decisively rejected throughout all of U.S. history. For this reason, orthogonal growth cannot be the proximate cause of the distribution of population across U.S. locations, whether Zipf's, log normal, or anything else.

We develop a simple one-sector model of a system of locations that closely matches observed dynamics. For an extended period, new locations enter the system with low initial population. Transitions up to their long-run relative population levels are slowed by an external friction on population growth. As a result, growth among smaller locations is characterized by strong convergence. Overlapping this extended period of entry, the congestion from the fixed supply of land net of agglomeration gradually lessens. As a result, population growth among intermediate and larger locations is characterized by divergence. Soon after entry ends, small locations diverge as well.

With only a handful of parameters, the model matches a wide range of evolving empirical relationships. This success suggests that the model captures much of the essence that has driven the evolution of the U.S. system of locations. If so, the model may help us to better understand the rapidly changing geographic distribution of population in countries such as China and Brazil. More generally, it may help us project how future shocks to technology, productivity and net congestion will effect the geographic distribution of economic activity.

Appendix A. Data construction details

A.1. Constructing geographically-consistent counties

The County Longitudinal Template (CLT) (Horan and Hargis, 1995; as corrected and amended by Beeson and DeJong; as corrected and amended by Rappaport) provides a simple way of joining data for county observations delineated in different census years. For example, the 1800 decennial census enumerated population and other aggregate outcomes for 417 counties (see Table 1). The 1820 census did so for 762 counties. Much of the increase over the intervening twenty years came from the settlement of land areas previously unoccupied by Europeans. But some of the increase also came from the splitting of previously-settled counties into two or more successor ones. The CLT "re-combines" the resulting successor counties in order to measure population growth rates for fixed geographic land areas.

For example, suppose enumerated county A in 1800 splits into enumerated counties B and C in 1820. In this case, the CLT additively collapses all 1820 data for counties B and C into a single observation, to which it assigns a vintage 1800 identification code. The CLT attaches this same vintage 1800 identification code to the 1800 data for county A. The 1800 and 1820 data can then be merged together using the vintage 1800 identification code. More specifically, the CLT collapses the 417 counties enumerated in the 1800 census into 310 counties. Doing so is the minimum joins necessary to create geographically-consistent borders to match the 1800 observations with observations from future years. To match the 1800 data for the 762 counties enumerated in the 1820 decennial census, these are also additively collapsed into 310 counties.

The CLT also merges back independent cities—county equivalents located primarily in Virginia—into the counties that surround them. Because independent city borders are endogenously drawn around densely-settled cities, combining them with the surrounding county removes a potentially serious bias.

A drawback of the CLT is that it requires combining counties whenever a change has occurred between the start date and the year 2000 even if the end date is prior to the change. For example, suppose that counties A and B merge together in 1821 to form county C. Matching the 1800 data for county A with the 1820 data for geographically-identical county A is not possible. Instead, only the additively collapsed data for counties A and B can be matched over these years.

A considerably more detailed documentation of the evolution of U.S. country borders from 1629 to 2000 is available from the Newberry Library. It allows for the careful study of the aggregate dynamics of one or a handful of counties over time, but it is less suited for studying the dynamics of all settled counties over time.

A.2. Combining counties into year-specific metro areas

The criteria for being considered a metropolitan area changes moderately over time. In all cases, metro areas are constructed as the combination of whole counties. For 1800 through 1940, we rely on Gardner (1999), who applies Census Bureau criteria from 1950 retroactively using the IPUMS data from each of the relevant decennial censuses. The Census Bureau criteria, in turn, consider a range of characteristics. Specifically, a metropolitan designation requires a contiguous group of counties that includes at least one urban center of at least 50,000 inhabitants. Each of the counties must have no more than one-third of employed persons working in the agricultural sector. Each must also have at least 10,000 nonagricultural workers or at least 10% as many nonagricultural workers as total workers in the primary county of the metropolitan area. If the number of nonagricultural workers is below one of these thresholds, the county can nevertheless be considered a metropolitan area if at least half of its population resides in a thickly settled area (at least 150 persons per square mile) contiguous to the central city.

For 1960 we use the "standard metropolitan statistical areas" delineations released by the Office of Management and Budget (OMB) in 1963 based on 1960 census data. We additionally use the New England County Metropolitan Area delineations released in 1975. For 1980, we use the 1983 OMB metro definitions with the exception that we retain the 1963 delineation of the New York City metro area. The 1983 Consolidated Metropolitan Area delineation for the New York City metro is far too large geographically to be considered a single labor market. But the Primary Metropolitan Statistical Area component that includes New York City is far too small. For New England, we use the 1983 New England County Metropolitan Area classification.

A.3. Creating time-lagged geographical borders

An important concern is that changes in counties' borders may be endogenous. For example, a predecessor enumerated county may legally fragment into two or more successor ones when its growth is expected to be fast. The successor counties will tend to

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

be small due to the fragmentation and to grow quickly if expectations were correct. To address this concern, we construct county-equivalents using the required geographically-consistent joins from 40 years prior to the start of the period over which growth is measured. In other words, we create longer-term geographically-consistent counties with borders from 40 years earlier. Our main results do not change with this robustness check. The validity of this strategy requires that any endogeneity of a county's geographic composition must dissipate within 40 years. As shown in the main text, growth transitions following entry typically die out within 40 years. This suggests that using geographic compositions lagged by 40 years to establish robustness should be an effective strategy.

A.4. Determining age and entry year

The algorithm for determining a county's age goes as follows: a geographically-consistent county is considered to be young if no more than 40 years have passed since its state or territory first had two or more enumerated counties with positive population. For example, the 1860 decennial census enumerated only one county in the geographic area that was to eventually become the state of Colorado. Ten years later in 1870, the decennial census enumerated 26 counties with positive population in the newly-formed Territory of Colorado. Thus we consider all geographically-consistent Colorado counties in the starting years of 1880 and 1900 to be young.

Counties not considered to be young are considered to be old if they meet two criteria. First, at least 60 years must have elapsed since each of the enumerated counties being combined to form a geographically-consistent county experienced its final significant geographic change, typically a split or join of land area other than a minor border adjustment (Forstall, 1996). For example, San Bernardino County, California, experienced its last significant geographic change in 1893, when a portion of it was ceded to form Riverside County. Hence this first criterion considers San Bernardino County to be old starting in 1953. Similarly, Peoria County, Illinois, experienced its last significant geographic change in 1831 when it ceded land to form Warren and Mercer counties. Hence we consider Peoria County to be old beginning in 1891. Second, at least 60 years must have passed since the CLT required any combining of raw counties to construct a historically-consistent one. In other words, all geographically-consistent old counties must have been made up of a single enumerated county for at least 60 years. Typically this second criterion yields equivalent results as the first. But for the minority of counties where the two criteria differ, we err towards falsely excluding it from the old group.

In order to construct growth trajectories by entry cohort, we need to identify when a county actually enters the U.S. system of locations. In doing so, we especially want to exclude any newly enumerated counties that are simply legal offshoots within long-occupied land area. For this reason, the change in the number of enumerated or geographically-consistent counties poorly corresponds to entry. Instead, we identify entries as all active counties in a state or territory that for the first time had at least two raw counties with positive population in the previous twenty years. Continuing an example from above, all counties in Colorado that were active in 1880 are considered to be entrants, because the first time the census enumerated more than two Colorado counties with positive population was 1870. Note that this criterion is similar to the designation of a location as young, with the exception that it is based on a shorter time horizon, twenty years instead of forty years. The relatively small number of counties that meet the entry criterion—552 over a two-century time span—makes clear that a large number of actual entries are being missed. As a result, a sufficient number of entrants to construct growth trajectories is available only for 1840, 1860, and 1880. (We do not to construct growth trajectories for the 1820 entry cohort because of the lower accuracy of the CLT for that year.)

Appendix B. Proof of Proposition 2

Proposition 2. *Assume that $\alpha > \varepsilon$, all potential locations are active, and all parameters are constant. Then in the absence of idiosyncratic shocks or in the absence of growth frictions, the economy converges to a balanced growth path characterized by population in all locations growing at the same expected rate.*

Proof. We first look at the case without any idiosyncratic shocks (but with growth frictions). For any period t, denote by $\{t,0\}$ the beginning of the period (before anyone moves) and by $\{t,1\}$ the end of the period (after people move). Take any two locations, i and j at time $\{t,0\}$ for which $w_{i,\{t,0\}} > w_{j,\{t,0\}}$. In other words, $Z_i^0 L_{i,\{t,0\}}^{-(\alpha-\varepsilon)} > Z_j^0 L_{j,\{t,0\}}^{-(\alpha-\varepsilon)}$. Free mobility implies that people will move until $w_{i,\{t,1\}} = w_{j,\{t,1\}}$. With $\alpha > \varepsilon$, such wage equalization requires that $\Delta L_{i,t}/L_{i,t-1} > \Delta L_{j,t}/L_{j,t-1}$. The larger proportional change in L_i implies that in $\{t+1,0\}$ the relative difference in wages will have decreased, $w_{i,\{t+1,0\}}/w_{j,\{t+1,0\}} < w_{i,\{t,0\}}/w_{j,\{t,0\}}$, which in turn implies that $\Delta L_{i,t+1}/L_{i,t} - \Delta L_{j,t+1}/L_{j,t} < \Delta L_{i,t}/L_{i,t-1} - \Delta L_{j,t}/L_{j,t-1}$, so that growth rates across locations converge over time, and Gibrat's law is reached.

We now look at the case without growth frictions (but with idiosyncratic shocks). At the end of any period t, wages are equal across i and j, which in the absence of frictions implies that $Z_i^0 \prod_{s=1}^{t} Z_{i,s}^{\text{idio}} L_{i,t}^{-(\alpha-\varepsilon)} = Z_j^0 \prod_{s=1}^{t} Z_{j,s}^{\text{idio}} L_{j,t}^{-(\alpha-\varepsilon)}$. At time t, the expected difference between location i and location j in their growth rates between t and $t+1$ is then $E_t((\log L_{i,t+1} - \log L_{i,t}) - (\log L_{j,t+1} - \log L_{j,t})) = (\alpha-\varepsilon) E_t(\log Z_{i,t+1}^{\text{idio}} - \log Z_{j,t+1}^{\text{idio}})$. Since $E_t(\log Z_{i,t+1}^{\text{idio}})$ is the same across locations, the expected difference in growth rates will be zero, so that Gibrat's law holds. □

Appendix C. Parameterization of net congestion

As described in the main text, the model depends on the net congestion arising from the land share of factor income partly offset by any increase in productivity arising from agglomeration ($\hat{\alpha}_t \equiv \alpha_t - \varepsilon_t$). A key calibration choice is the *proportional* decrease in this net congestion parameter over the mid-nineteenth and early-twentieth centuries. The respective start and end dates in 1840 and 1960 are pinned down by the rapid acceleration and eventual deceleration in rural to urban migration (U.S. Bureau of the Census, 1975: Series A 57–72). The calibrated 0.15 starting level of net congestion is without loss of generality. The calibrated one-third proportional decrease is based primarily on matching observed convergence and divergence over successive twenty-year intervals. A moderately smaller proportional decrease results in too little divergence. A moderately larger proportional decrease results in too much divergence.

An important constraint on the calibrated proportional decrease in net congestion is the need to match the *increase* in the standard deviation of the population distribution between 1790 and 2000. Absent any stochastic shocks to productivity, the one-third decrease in net congestion achieves almost three-quarters of the required increase. Correspondingly, only slightly more than a quarter of the observed increase in population dispersion depends on the stochastic shocks to location productivity. A two-fifths proportional decrease in net congestion is the largest possible change (combined with no stochastic shocks to productivity) that does not cause the implied increase in population dispersion to exceed

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

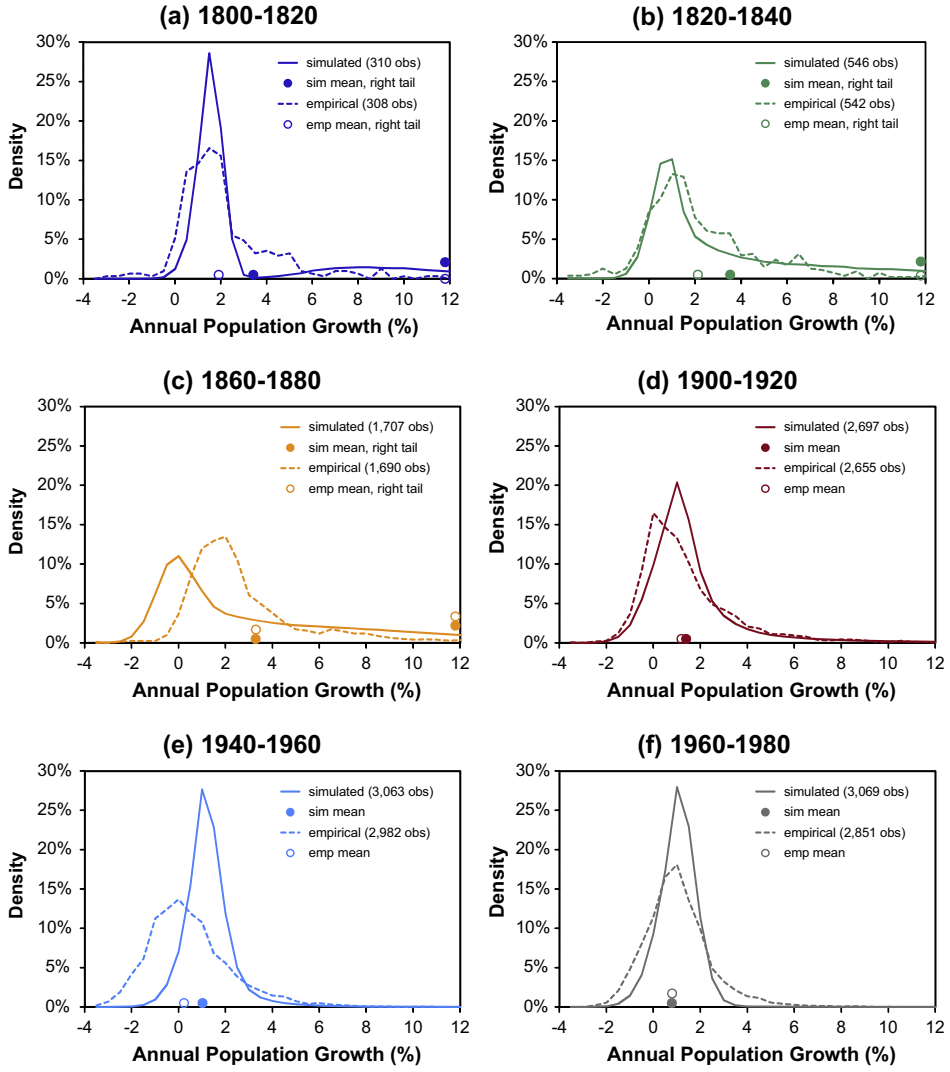

Fig. D.1. Population growth distributions, simulated versus empirical, additional 20-year intervals. These are the twenty-year periods not included in the main text, Fig. 11. Growth rates are not normalized. The right tail is calculated as the cumulative density of locations with growth above 12%.

what is observed. But with respect to the distribution of observed growth rates—for all locations during the late twentieth century and for old locations only in earlier years—a higher-variance stochastic shock would achieve a better match.

To see the plausibility of the calibrated one-third proportional decrease in net congestion, consider it as arising solely from a decrease in the land factor income parameter, α_t. More specifically, consider a decrease in the land share from 0.15 in 1840 to a value of 0.10 in 1960 (along with an implicit constant zero value for the agglomeration parameter). The United States was obviously a very agrarian economy in the early nineteenth century. Based on data from the mid-twentieth century, Mundlak (2001) reports estimates of the land share of agricultural goods factor income that range from 0.20 to 0.36. Over the mid-nineteenth and early-twentieth centuries there was significant capital-biased technological progress in agricultural production (e.g., the steel plow, the grain elevator, chemical fertilizers, barbed wire, and powered tractors). Suppose that in 1840, prior to these innovations, the land factor income share in agriculture was in the upper third of Mundlak's range, say 0.30. Assuming that agriculture's share of household consumption expenditure was about one third in 1840 implies a 0.10 percentage point additive contribution to land's share of aggregate factor income in that year.

Next, suppose that land's share of factor income from housing services was about 0.15 in 1840, which is considerably below modern-day estimates. One reason why it might be lower than today is the much higher cost of the structure input in 1840 relative to the cost of the unimproved land input in that year (along with sufficient complementarity between the structure and land inputs). Consistent with a relatively low historical land share, the estimated present-day land share of housing factor income in land-abundant metro areas is estimated to be between 17% and 27% (Jackson et al., 1984; Thorsnes, 1997). Also suppose that the housing service share of household consumption was 15%, which is slightly below its share circa 2000. The resulting contribution from housing services to land's aggregate factor income share is 0.02 percentage point. The additional 0.03 percentage point contribution required to attain a 0.15 aggregate land share of factor income requires that the remaining 52% of household consumption expenditures have a land factor income share of 0.05. Consistent with this, Caselli and Coleman (2001) calibrate the land share of manufacturing consumption in 1880 to be 0.06.

To the extent that there was some increasing returns to production, a 0.15 calibrated value for net congestion in that year would require a higher land share α_t to allow for the offset. A land share

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

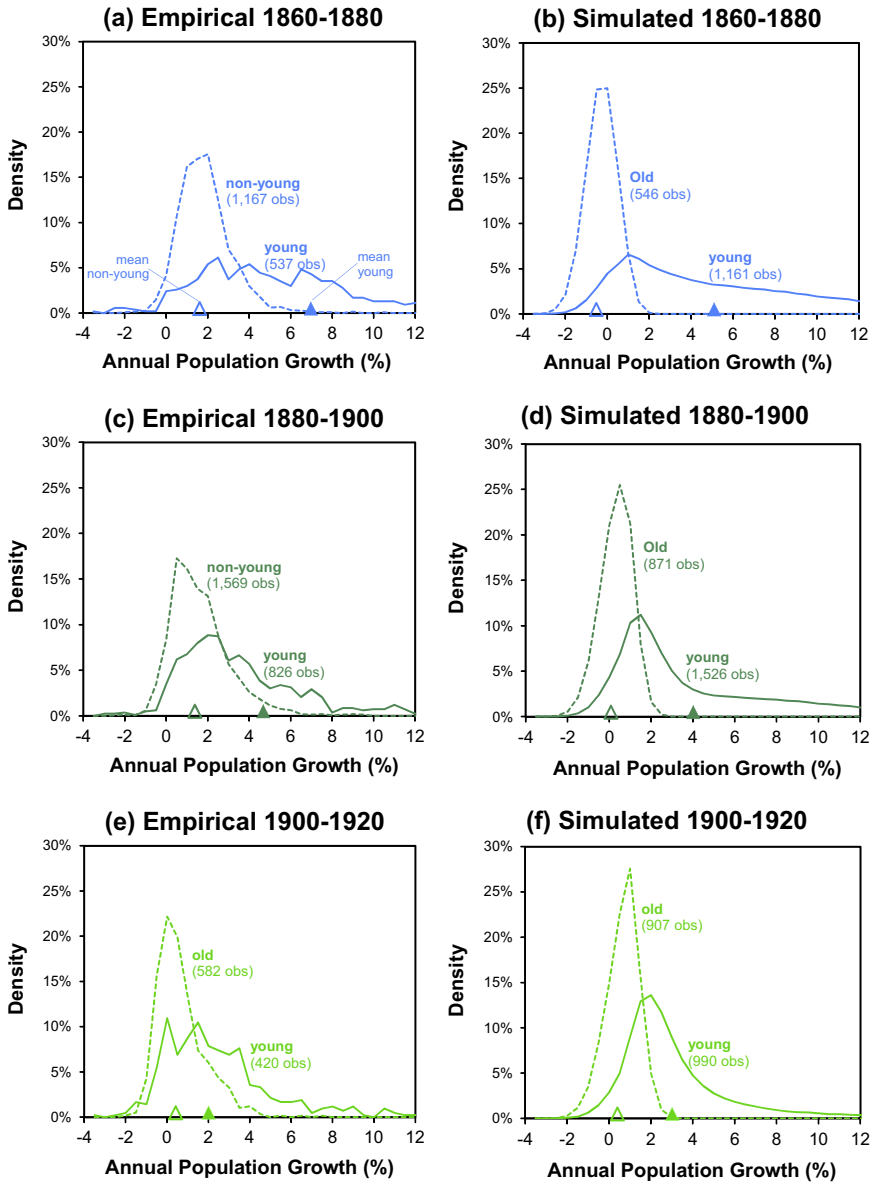

Fig. D.2. Simulated and empirical population growth distributions by age, 1860–1920. Figure shows the empirical and simulated distribution of population growth across locations by age groups for the 20-year periods beginning in 1860, 1880, and 1900. For the first two of these periods, the age split is between "young" and remaining locations. For the period beginning in 1900, the split is between "young" and "old" locations. Definitions of these age categories are included in Appendix A.

above 0.15 in 1840 is easily plausible. An assumed land share of agricultural factor income of 0.42 in that year (moderately above Mudlak's reported upper-bound estimate of 0.36 for the mid-twentieth century) boosts the agricultural contribution to the aggregate land share of factor income to 14 percentage points. The land share of housing-service factor income and the housing-service share of consumption expenditures can also be justified as being somewhat higher in 1840 than posited immediately above. Any of these three possibilities would leave room for even a generous positive value for ε.

The plausibility of a 0.10 aggregate land factor income share in 1960 rests on land's contribution to the production of housing services. Between 1975 and 2004, land accounted for an average of 47% of the sales value of the aggregate U.S. housing stock (Davis and Heathcote, 2007). Adjusting for the fact that structures depreciate but land does not (using the 1.6% rate of structure depreciation suggested by Davis and Heathcote and a 4% required real rate-of-return) implies a 39% land factor share. Based on U.S. consumption data, assume that the housing services share of aggregate consumption expenditure in 2000 was 0.18. This is slightly above housing's 0.15 share in the U.S. NIPA accounts for 2000 but well below its 0.30 weight in the U.S. Consumer Price Index in that year. The implied contribution to the aggregate land share of factor income is 0.07 percentage point. Agriculture and manufacturing each contribute very little to land's aggregate share in 2000. For agriculture the reason is its very small share of consumption expenditures, estimated to be about 0.014 (Caselli and Coleman, 2001). For manufacturing, the reason is a very small land share of factor income in that sector, estimated to be about 0.016 (Jorgenson et al., 2005; Rappaport, 2008; Ciccone, 2002). As a result, the combined contribution to the aggregate land factor income share from agriculture and manufacturing is only about 0.01% point. Using a one-third consumption expenditure share for manufacturing residually implies a 0.47 consumption expenditure share on non-residential services. To match a 0.10 aggregate land share target in 2000 requires that non-residential services

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

must have an average 0.047 land share. This is definitely high for many services such as retail, restaurants, entertainment, transportation, and utilities. But a broader interpretation of consumption to include non-market goods such as streets, highways, airports, and parks easily justifies the 0.10 aggregate parameterization. For example, Solow (1973) argues that streets occupy about one quarter of the land area of residential structures within a metro area. If so, this would represent an approximate 0.02 percentage point contribution to land's share of aggregate factor income.

There are a number of reasons to think that the decrease in net congestion may have been bigger than one third. First, as argued above, there is considerable scope to justify a calibrated early-nineteenth-century land share significantly above 0.15. But there is much less scope to argue for a late twentieth century land share above 0.10. Second, there are many reasons to believe that the agglomerative offset to net congestion has increased over time. For example, decreased transport costs, accelerating technological change, and increased specialization all suggest that the elasticity of productivity with respect to population (or some alternative measure of size) is likely to have been higher in the late twentieth century than it was in the early nineteenth century. Third, technological progress over time has almost surely decreased net congestion at any given population level, broadly interpreted. For example, mass transit systems in general and subways in particular were key developments that allowed for both denser and less geographically compact living. Similarly, the automobile and commuter highways considerably increased the feasible geographic distance between where someone lived and where they worked.

If the decrease in net congestion was indeed above one third, the dispersion of population levels across counties should be greater than is observed. Large federal and state government subsidies and transfers to rural counties may be one reason this greater dispersion has not yet occurred.

Appendix D. Additional tables and figures

Figs. D.1–D.3.

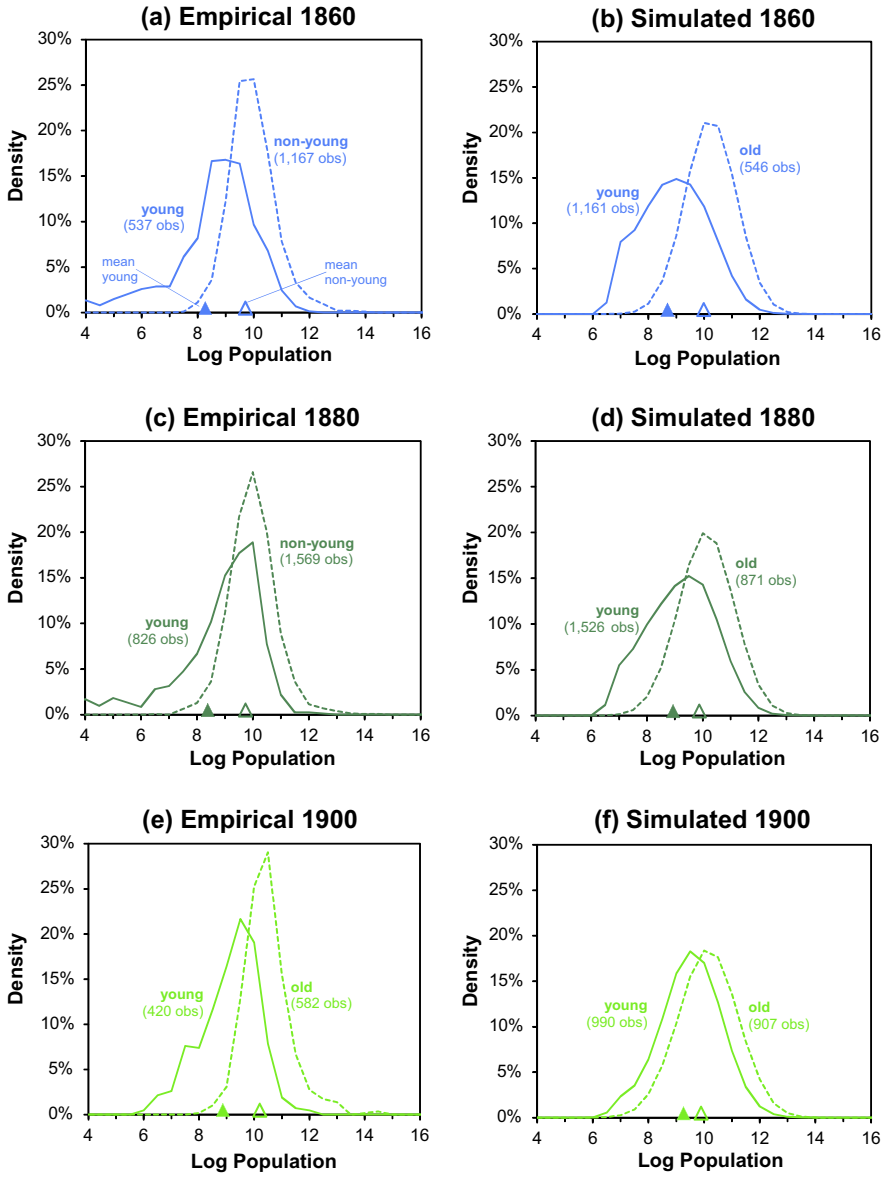

Fig. D.3. Empirical and simulated population level distributions by age, 1860–1900. Figure shows the empirical and simulated distribution of population across locations by age groups for 1860, 1880, and 1900. For the first two of these years, the age split is between "young" and remaining locations. For 1900, the split is between "young" and "old" locations. Definitions of these age categories are included in Appendix A.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004

References

Barro, R.J., Sala-i-Martin, X., 1991. Convergence across states and regions. Brookings Papers on Economic Activity 1, 107–182.

Barro, R.J., Mankiw, N.G., Sala-i-Martin, X., 1995. Capital mobility in neoclassical models of growth. American Economic Review 85, 103–115.

Beeson, P.E., DeJong, D.N., 2002. Divergence. Contributions to Macroeconomics, 2(1), Article 6 (B.E. Press).

Caselli, F., Coleman, W.J., 2001. The U.S. structural transformation and regional convergence: a reinterpretation. Journal of Political Economy 109, 584–616.

Ciccone, A., 2002. Agglomeration effects in Europe. European Economic Review 46, 213–227.

Conley, T., 1999. GMM estimation with cross sectional dependence. Journal of Econometrics 92, 1–45.

Davis, M.A., Heathcote, J., 2007. The price and quantity of residential land in the United States. Journal of Monetary Economics 54, 2595–2620.

Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and break points: the geography of economic activity. American Economic Review 92, 1269–1289.

Desmet, K., Fafchamps, M., 2006. Employment concentration across U.S. counties. Regional Science and Urban Economics 36, 482–509.

Desmet, K., Rossi-Hansberg, E., 2009. Spatial growth and industry age. Journal of Economic Theory 144, 2477–2502.

Dinkelman, T., Schulhofer-Wohl, S., 2013. Migration, Congestion Externalities, and the Evaluation of Spatial Investments, Working Paper 700. Federal Reserve Bank of Minneapolis.

Dittmar, J., 2011. Cities, Markets and Growth: The Emergence of Zipf's Law, unpublished manuscript.

Easterlin, R.A., 1960. Interregional difference in per capita income, population, and total income, 1840–1950. In: Parker, W. (Ed.), Trends in the American Economy in the Nineteenth Century, Studies in Income and Wealth, vol. 24. Princeton University Press, Princeton, NJ, pp. 73–140.

Eaton, J., Eckstein, Z., 1997. Cities and growth: theory and evidence from France and Japan. Regional Science and Urban Economics 27, 443–474.

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

Forstall, R., 1996. Population of States and Counties of the United States: 1790–1990. U.S. Bureau of the Census <http://www.census.gov/population/www/censusdata/pop1790-1990.html>.

Gabaix, X., 1999. Zipf's law for cities: an explanation. Quarterly Journal of Economics 114, 739–767.

Gallin, J.H., 2004. Net migration and state labor market dynamics. Journal of Labor Economics 22, 1–23.

Gardner, T., 1999. Metropolitan classification for census years before world war II. Historical Methods 32 (3), 139–150.

Gaspar, J., Glaeser, E.L., 1998. Information technology and the future of cities. Journal of Urban Economics 43, 136–156.

Gennaioli, N., La Porta, R., Lopez de Silanes, F., Shleifer, S., 2014. Growth in regions. Journal of Economic Growth 19, 259–309.

Giesen, K., Suedekum, J., 2014. City age and city size. European Economic Review 71, 192–208.

Glaeser, E.L., Gyourko, J., 2005. Urban decline and durable housing. Journal of Political Economy 113, 345–375.

Glaeser, E.L., Scheinkman, J.A., Shleifer, A., 1995. Economic growth in a cross-section of cities. Journal of Monetary Economics 36, 117–143.

Haines, M.R., 2005. Historical, Demographic, Economic, and Social Data: The United States 1790–2002 [Computer File]. ICPSR Study 2896. Inter-university Consortium for Political and Social Research [distributor].

Hansen, G., Prescott, E.C., 2002. Malthus to Solow. American Economic Review 92, 1205–1217.

Holmes, T.J., Lee, S., 2010. Cities as six-by-six-mile squares: Zipf's law? NBER chapters. In: Agglomeration Economics. National Bureau of Economic Research, pp. 105–131.

Horan, P.M., Hargis, P.G., 1995. County Longitudinal Template, 1840–1990. [computer file]. ICPSR Study 6576. Inter-university Consortium for Political and Social Research [distributor]. Corrected and amended by Patricia E. Beeson and David N. DeJong, Department of Economics, University of Pittsburgh, 2001. Corrected and amended by Jordan Rappaport, Federal Reserve Bank of Kansas City, 2010.

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

Jackson, J.R., Johnson, R.C., Kaserman, D.L., 1984. The measurement of land prices and the elasticity of substitution in housing production. Journal of Urban Economics 16, 1–12.

Jorgenson, D.W., Ho, M.S., Stiroh, K.J., 2005. Growth of U.S. industries and investments in information technology and higher education. In: Corrado, C., Haltiwanger, J., Sichel, D. (Eds.), Measuring Capital in the New Economy. University of Chicago Press, Chicago, IL.

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

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

Michaels, G., Rauch, F., Redding, S., 2012. Urbanization and structural transformation. Quarterly Journal of Economics 127, 535–586.

Mitchener, K.J., McLean, I.W., 1999. U.S. regional growth and convergence, 1880–1980. Journal of Economic History 59, 1016–1042.

Mundlak, Y., 2001. Production and supply. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier Science, B.V.

Rappaport, J., 2004. Why are population flows so persistent? Journal of Urban Economics 56, 554–580.

Rappaport, J., 2005. How does labor mobility affect income convergence? Journal of Economic Dynamics and Control 29, 567–581.

Rappaport, J., 2007. Moving to nice weather. Regional Science and Urban Economics 37, 375–398.

Rappaport, J., 2008. Consumption amenities and city population density. Regional Science and Urban Economics 38, 533–552.

Rappaport, J., Sachs, J.D., 2003. The United States as a coastal nation. Journal of Economic Growth 8, 5–46.

Rosenbloom, J.L., 1990. One market or many? Labor market integration in the late nineteenth-century United States. Journal of Economic History 50, 85–107.

Sánchez-Vidal, M., González-Val, R., Viladecans-Marsal, E., 2014. Sequential city growth in the U.S.: does age matter? Regional Science and Urban Economics 44, 29–37.

Solow, R.M., 1973. Congestion cost and the use of land for streets. Bell Journal of Economics and Management Science 4, 602–618.

Thorndale, W., Dollarhide, W., 1987. Map Guide to the Federal Censuses, 1790–1920. Genealogical Publishing Company, Baltimore.

Thorsnes, P., 1997. Consistent estimates of the elasticity of substitution between land and non-land inputs in the production of housing. Journal of Urban Economics 42, 98–108.

U.S. Bureau of the Census, 1975. Historical Statistics of the United States Colonial Times to 1970 Bicentennial Edition, Washington DC.

Please cite this article in press as: Desmet, K., Rappaport, J. The settlement of the United States, 1800–2000: The long transition towards Gibrat's law. Journal of Urban Economics (2015), http://dx.doi.org/10.1016/j.jue.2015.03.004