ISSN 1936-5330

THE FEDERAL RESERVE BANK of KANSAS CITY
RESEARCH WORKING PAPERS

The Settlement of the United States, 1800 to 2000: The Long Transition Towards Gibrat's Law

Klaus Desmet and Jordan Rappaport

February 2013; Revised September 2013
RWP 13-02

RESEARCH WORKING PAPERS

The Settlement of the United States, 1800 to 2000: The Long Transition Towards Gibrat's Law[*]

Klaus Desmet
Universidad Carlos III

Jordan Rappaport
Federal Reserve Bank of Kansas City

September 2013

Abstract

Gibrat's law, the orthogonality of growth to initial levels, is considered a stylized fact of local population growth. But throughout U.S. history, local population growth has significantly deviated from orthogonality. In earlier periods smaller counties strongly converged whereas larger counties moderately diverged. Over time, due to changes in the age composition of locations and net congestion, convergence dissipated and divergence weakened. Gibrat's law gradually emerged without fully attaining it. A simple one-sector model, with entry of new locations, a growth friction, and decreasing net congestion closely matches these and many other observed relationships. Our findings suggest that orthogonal growth is a consequence of reaching a steady state population distribution, rather than an explanation of that distribution.

1 Introduction

Gibrat's law, the orthogonality of growth and initial levels, has long been considered a stylized fact of local population growth (Glaeser, Sheinkman and Shleifer 1995; Eaton and Eckstein, 1997; Ioannides and Overman, 2003). Orthogonal growth implies that the asymptotic distribution of local population will be distributed log normally across all locations and approximately according to Zipf's law across the largest locations (Gabaix, 1999; Eeckhout, 2004). This is indeed what is observed across cities in numerous countries at numerous points in time (Rosen and Resnick, 1980; Krugman, 1996a). More recently the robustness of Gibrat's law has been questioned in the context of small geographic units and historical growth rates. Holmes and Lee (2008) divide the U.S. into a grid of six-by-six mile squares and find an inverted-U relation between size and growth; Beeson and DeJong (2002) and Michaels et al. (2012) document a positive relation between size and growth—divergence—for intermediate-sized locations during the nineteenth century; and Dittmar (2011), using

[*] Michael Connolly 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 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. An extended appendix is available from www.kansascityfed.org/publications/research/rwp/index.cfm.

historical data on European cities, shows that Gibrat's law only emerged in the modern period, after 1500.

Building on this previous research, we empirically document and quantitatively match the changing relationship between the level of population and its growth rate over the nineteenth and twentieth centuries. Using U.S. counties and metro areas as our units of observation, we find that population growth was strongly negatively correlated with initial population across small locations throughout the nineteenth and early twentieth centuries. The implied convergence in population levels began to dissipate in 1900 and was completely gone by 1940. Population growth was also moderately positively correlated with initial population across intermediate and large locations during the late nineteenth and early twentieth centuries. The implied divergence in population levels gradually weakened during the late twentieth century but never completely. The evolution of the U.S. system of locations thus decisively differed from Gibrat's law throughout most of U.S. history.

This absence of orthogonal growth calls into question whether the presently-observed log normal population size distribution can be the asymptotic result of random growth. Reinforcing this skepticism is the fact that already in 1790 the U.S. population distribution across counties was almost exactly log normal. More consistent with our findings is that the current and late eighteenth century log normal U.S. population distributions reflect a log normal distribution of underlying total factor productivity and hence *steady state* location population (Krugman, 1996b; Rappaport and Sachs, 2003).[1] In our interpretation, random growth does not describe the population dynamics per se, but the evolution of total factor productivity (Davis and Weinstein, 2002).[2] Observed deviations from orthogonal population growth may therefore very well occur, reflecting transitions of local population to their steady state levels. But once steady state is reached, both growth will be orthogonal and the population distribution will be log normal.

We hypothesize that the observed convergence of small locations reflects the continual "entry" of new counties into the U.S. system of locations and their subsequent rapid transitions to their steady-state population. Over the 200 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 represents either a drop in the congestion from land or an increase in the importance of agglomeration economies or both. Michaels et. al (2012) argue for the former by emphasizing the structural transformation out of agriculture. Hansen and Prescott

[1] So long as underlying total factor productivity reflects the combination of a sufficient number of "fundamental" inputs (e.g., coastal proximity, natural resource abundance, weather), the implied distribution of total factor productivity will be log normal even if the distribution of each fundamental input is not (Lee and Li, 2013).

[2] In the absence of mobility frictions, the latter would imply the former (see, e.g., Eeckhout, 2004).

(2002) also document a drop in land shares. Alternatively, Desmet and Rossi-Hansberg (2009) argue for the latter by showing how the introduction of general purpose technologies at several points during the twentieth century increased the returns to agglomeration.

Our empirical findings document five salient features that lend support to these hypothesis. First, locations of similar "age" exhibit similar growth patterns independent of calendar year. Growth of young locations is *always* characterized by convergence. Growth of old locations is *never* characterized by 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 as a direct result of the waning of location entry 20 years earlier. Fourth, the divergence portion of local growth also dissipates over time. Fifth, positive local population growth is highly persistent across twenty-year periods.

Informed by these five salient features of U.S. local growth described above along with our two hypotheses, we develop a simple one-sector general equilibrium model of the evolution of a system of locations. An initial stochastic productivity draw and subsequent i.i.d. shocks uniquely determine each location's steady-state population. 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 the upward transition to each location's steady state and so cause population growth from low levels to be characterized by convergence. Overlapping this extended period of entry, the congestion arising from the fixed supply of land in each location gradually diminishes. This occurs because of either a decrease in land's share of factor income or an increase in the effect of agglomeration on productivity or both. The decrease in net congestion costs causes population levels to become more sensitive to underlying differences in productivity and so introduces a force towards divergence. Once entry is complete and the degree of net congestion stabilizes, the assumed orthogonality of productivity growth causes population growth to be orthogonal as well.

A numerical implementation of this setup with just a handful of "free" parameters tightly matches a large number of empirical relationships including the five we just described. Most model parameters are pinned down to match a specific empirical moment. For example, modeled aggregate population matches aggregate U.S. population, and modeled location entry matches newly enumerated locations in each decennial census. Only four model parameters retain any freedom to match more than fifty additional 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, and the non-linear persistence of population growth over nine adjacent pairs of twenty-year periods. Many of these distributions are also matched for "young" and "old" location subsamples. Additionally, the model matches population growth trajectories by different entry cohorts over multiple periods.

Such a close matching of such a rich set of empirical characteristics with very few degrees of freedom suggests that our simple model captures much of the essence that drove the local development of the United States. Naturally the model abstracts from many first-order events that have contributed to shaping the country's economic geography. Examples include steamships, railroads, the Civil War, the various homestead acts, electrification, the automobile, and highway building. Although all of these played pivotal roles, our results indicate that many of the important features of U.S. local growth over the last two centuries can be understood without them.

Our work builds on the literature on U.S. local development, but differs in several ways. Both Beeson and DeJong (2002) and Michaels et al. (2012) focus on why the relation between size and growth may be upward-sloping. Beeson and DeJong (2002) document this empirically, whereas Michaels et al. (2012) emphasize the role of the structural transformation in driving divergence. Instead, our aim is to analyze the slowly changing growth dynamics of the entire distribution over the last two centuries. This explains our focus on successive twenty-year periods from 1800 to 2000.

Also related is an incipient literature that analyzes the importance of the age of locations. Giesen and Südekum (2012), for example, explore the city size distribution, and find that older cities are on average larger. Their focus is not on describing how growth changes as cities age, but rather on how today's city size distribution depends on a city's age. Similarly, González-Val, Sánchez-Vidal and Viladecans-Marsal (2012) analyze U.S. city growth over the 20th century, and empirically document that new cities grow faster. Our findings are consistent with this, but we look at all locations over two centuries, and complement the empirical analysis with a model that is able to quantitatively match the main features of U.S. local growth since 1800.

The rest of the paper is organized as follows: Section 2 discusses the data. Section 3 reports the main empirical findings. Section 4 proposes a simple theory of the settlement of the United States. Section 5 calibrates the model. Section 6 presents numerical results. Section 7 concludes.

2 Data

Our dataset is built using data for county and county-equivalents from the 1790 through 2000 decennial censuses (Haines, 2005). The nineteenth century was a period during which U.S. settled land area grew rapidly. Correspondingly, the number of counties enumerated in each decennial census soared from just under 300 in 1790 to almost 2900 in 1900 (Table 1 column 1). Over the first few decades of the twentieth century, the number of enumerated counties further increased to more than 3100.[3]

	(1)	(2)	(3)	(4)	(5)	(6)
	counties enumerated in decennial census		**geographically-consistent counties**		**hybrid: geographically-consistent + metros**	
Year	number	change	number	change	number	change
1790	293		233		232	
1800	417	124	310	77	310	78
1820	762	345	545	235	544	234
1840	1,286	524	870	325	865	321
1860	2,086	800	1,706	836	1,692	827
1880	2,582	496	2,396	690	2,369	677
1900	2,841	259	2,696	300	2,655	286
1920	3,082	241	3,014	318	2,950	295
1940	3,116	34	3,062	48	2,982	32
1960	3,121	5	3,064	2	2,853	-129
1980	3,126	5	3,067	3	2,631	-222

Table 1: Alternative Measures of Locations and Location Entry. Geographically-consistent counties combine enumerated counties so as to keep borders approximately constant over a given twenty-year interval. The empirical baseline metro build additionally combines geographically-consistent counties into metro areas based on initial conditions.

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). This allows us to "re-combine" two or more counties that split off from each other during the interval over which growth is measured. 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. Additional details on the CLT are included in Appendix A.

The resulting geographic adjustments require that we construct a separate data set 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 data sets, the required joins reduce the number of observations by about one third (Table 1

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

column 3 versus column 1). For the twentieth century data sets, the reduction is relatively modest. A limitation of the CLT is that it formally tracks county changes back only to 1840. For the 1800 and 1820 data sets, we use the CLT's 1840 borders with a handful of modifications. Calculated growth rates that start in either of these years may fail to account for splits that cede significant land and population to a newly-formed county. Hence measured growth rates may be biased downward. Results for outcomes prior to 1840 should therefore be interpreted with extra caution.

When and where metro areas exist, we argue that they correspond better to unified labor markets than do their constituent counties. In particular, they are the smallest geography that encompasses were residents both live and work. Therefore, we use the hybrid of metro areas and remaining geographically-consistent counties as our baseline set of observations. For 1960 and 1980, we use respective delineations promulgated by the Office of Management and Budget (OMB) in 1963 and 1983 with some minor modifications. 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). Additional details are included in Appendix A.

3 Empirical Results

This section documents the changing 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 the ten twenty-year periods, 1800-1820 through 1980-2000. The first are non-linear kernels, which show a first-derivative continuous approximation of growth versus initial population level. These regressions allow for sharp visual comparisons of the relationship between growth and initial population as it develops over time. The second are comparable continuous, piecewise-linear spline regressions. These allow for more quantitative measures of the dependence of growth on initial population including the ability to formally test for orthogonality. The kernel regressions are of the form

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

where $L_{i,t}$ is the log population for location i in year t. The left hand side is just the average rate of population growth over twenty years. The estimation uses an Epanechnikov kernel (Desmet and Fafchamps, 2006). The continuous piecewise-linear spline regressions are of the form

$$(L_{i,t+20} - L_{i,t})/20 = \vec{\beta}_t(1 + \vec{L}_{i,t}) + e_{i,t}$$

where the vector $\vec{L}_{i,t}$ is 1 by k where k is the number of spline 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. The baseline regressions take the hybrid of metropolitan areas and remaining geographically-consistent counties as the geographic unit of observation. The robustness and age-based regressions use only geographically-consistent counties due to feasibility constraints. For geographic reasons, the residuals from regressions on county and metro observations are unlikely to be independent of each other. Using a generalization of the Huber-White algorithm, reported standard errors are constructed to be robust to spatial correlation between county pairs located within 200 km of each other (Conley 1999; Rappaport 2007).

3.1 All Locations

As a starting point, we analyze the relationship between size and growth for all locations for each of the twenty-year periods spanning 1800 to 2000. Both the kernel and spline regressions strongly reject the null hypothesis of orthogonal growth. The kernel regressions emphasize this visually (Figure 1).[4] Displayed growth rates are normalized by subtracting out the aggregate growth rate of all locations that were already active at the start of each twenty-year period.[5] This normalized growth rate is the steady-state expected growth rate of *all* counties/metros in the absence of further location entry. Henceforth it will be referred to simply as "aggregate growth".

Contrary to Gibrat's law, the fitted growth rates are clearly not horizontal. This violation of orthogonality holds for each of the ten 20-year periods and over most of the log population range. Several other features stand out. First, until 1940 the correlation between growth and size for counties/metros with low population is strongly negative. Low-population locations were thus converging towards intermediate-sized ones. Second, beginning in 1880 the correlation between growth and size for high-population locations is moderately positive. High-population counties were thus diverging away from lower-population ones. Third, the negative correlation of growth with size diminishes over time; beginning in 1940 it is no longer present. Fourth, the positive correlation of growth with size also diminishes over time though never completely. Population growth thus transitions towards Gibrat's law over time but never actually attains it.

The continuous piecewise-linear spline regressions give very similar results (Table 2). For example, the coefficients from the 1880-to-1900 regression are negative and statistically differ from zero for the five log population spline segments with lower bounds of 4, 6, 7, 8, and 9. Within the 4-to-6 bin, a 1 log point increase in initial population is associated with a decrease in average annual growth of 3.4 percentage points. Within the 10-to-11

[4] All fitted kernel regressions are trimmed to exclude from display the smallest 1 percent of observations plus the two largest observations.

[5] For example, a displayed 1800-to-1820 growth rate of zero implies actual 1800-to-1820 growth of 1.6 percent, which is the twenty-year aggregate growth rate for the 310 counties that were active in 1800. Aggregate growth of the entire United States from 1800 to 1820 was 3.0 percent. This higher growth rate compared to the aggregate growth of existing locations reflects the combined population of the 234 counties that entered the U.S. system between 1801 and 1820.

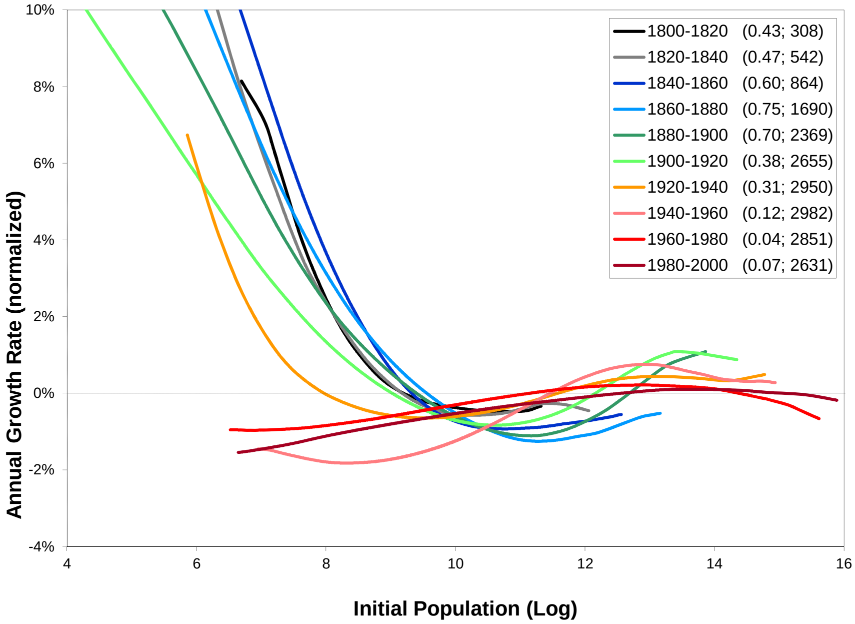


Figure 1: Population Growth vs Initial Population, 1800-2000. Figure shows fitted growth rates from kernel regressions of county/metro population growth on initial log population. Fitted growth rates are normalized by subtracting the aggregate growth rate of locations active at the start of each 20-year period. The first number in parentheses is R-squared statistic; the second is the number of observations. In this and all fitted kernels below, the smallest 1 percent of locations and the 2 largest ones have been trimmed from the display.

bin, growth is essentially uncorrelated with size. And within the 11-to-12 bin, a positive coefficient statistically differs from 0 at the 0.05 level. Its magnitude implies that a 1 log point increase in population within this range is associated with a 0.9 percentage point increase in growth.

Although Gibrat's can be rejected over most of the distribution, this is not true for the largest locations. The coefficient on increases in log population within the uppermost bin does not statistically differ from zero for all but one of the periods. This is consistent with growth being orthogonal to size for large metropolitan areas — as suggested by the literature — but not for small counties. The exception to this orthogonality is the 1960-to-1980 period. It admits a modestly negative, statistically significant coefficient for the highest bin. This indication of a net increase in congestion at very high population levels likely reflects the intense suburbanization taking place over this period.[6]

A possible critique of the present analysis is that the low-population counties that exhibit convergence for the 1800 through 1900 time periods (those with population less than 20,000, corresponding to approximately log 10) are of low importance to aggregate U.S. outcomes.

[6] The estimated convergence among very large counties and metros between 1960 to 1980 partly reflects suburbanization to the extent that eventual suburban counties were not designated as part of a metro area in 1960.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
log(pop) bin:	**1800-1820**	**1820-1840**	**1840-1860**	**1860-1880**	**1880-1900**	**1900-1920**	**1920-1940**	**1940-1960**	**1960-1980**	**1980-2000**
min to lowest lb	**-0.031** **(0.003)**	**-0.047** **(0.007)**	**-0.053** **(0.007)**	**-0.047** **(0.005)**	-0.007 (0.024)	**-0.023** **(0.008)**	**-0.039** **(0.006)**	0.004 (0.004)	0.000 (0.004)	0.001 (0.002)
lpop 04to06					**-0.034** **(0.008)**					
lpop 05to07				**-0.044** **(0.006)**						
lpop 06to07					**-0.036** **(0.012)**					
lpop 07to08				**-0.032** **(0.007)**	**-0.026** **(0.007)**	**-0.025** **(0.007)**				
lpop 08to09		**-0.020** **(0.005)**	**-0.039** **(0.006)**	**-0.025** **(0.005)**	**-0.022** **(0.003)**	**-0.015** **(0.004)**	-0.002 (0.002)	**-0.007** **(0.002)**	0.002 (0.003)	**0.005** **(0.002)**
lpop 09to10	0.002 (0.004)	**-0.008** **(0.004)**	-0.005 (0.003)	**-0.010** **(0.002)**	**-0.010** **(0.002)**	**-0.009** **(0.002)**	0.001 (0.001)	**0.006** **(0.001)**	**0.003** **(0.001)**	**0.003** **(0.001)**
lpop 10to11	-0.001 (0.007)	**0.010** **(0.004)**	0.000 (0.003)	0.001 (0.002)	0.003 (0.002)	**0.010** **(0.002)**	**0.006** **(0.001)**	**0.011** **(0.001)**	**0.004** **(0.001)**	0.001 (0.001)
lpop 11to12				0.002 (0.008)	**0.009** **(0.004)**	0.006 (0.005)	**0.006** **(0.002)**	**0.013** **(0.003)**	0.002 (0.002)	**0.004** **(0.002)**
lpop 12to13							-0.005 (0.003)	**-0.007** **(0.003)**	0.000 (0.002)	0.000 (0.002)
lpop 13to14									0.002 (0.004)	0.001 (0.003)
highest ub to max	0.002 (0.005)	-0.006 (0.007)	0.002 (0.005)	0.001 (0.005)	0.004 (0.003)	0.002 (0.003)	0.003 (0.002)	0.000 (0.002)	**-0.007** **(0.002)**	-0.002 (0.002)
Bins	4	5	5	8	9	7	7	7	8	8
N	304	538	860	1,685	2,354	2,648	2,943	2,979	2,849	2,630
R^2	0.401	0.478	0.615	0.759	0.654	0.376	0.306	0.128	0.043	0.074

Table 2: Population Growth vs Initial Population, 1800-2000. Table shows results from regressing county/metro population growth on a piecewise-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. Coefficients in the top-most row apply to each regression's lowest population bin, which has an upper bound equal to the lower bound of the next highest reported bin. Similarly, the last coefficient row has a lower bound equal to the upper bound of the next-lowest reported bin. Bins are constructed with the requirement that those with a lower bound of 7 or less have a minimum of 40 observations; those that have a lower bound of 8 through 10 have a minimum of 20 observations; and those that have a lower bound of 11 or higher have a minimum of 10 observations.

While possibly true today, this critique was definitely not true during the nineteenth century. In 1800 nearly half of the U.S. population lived in counties with log population below 10 and more than a quarter did so in 1900. Even in 1920, the start of the final twenty-year period characterized by convergence in the lower part of the distribution, more than a fifth of the U.S. population still resided in these small counties.

Another concern is that some counties are split apart exactly because growth is expected to be high. If widespread, this causal relationship from fast expected growth to the splitting up — and thus to smaller size — significantly changes the interpretation of the negative correlation between growth and population. To minimize this reverse causal channel, we recombine into single observations any county split-offs that occurred in the 40 years prior to each growth period. The resulting fitted curves are very similar to the base-specification ones.[7]

As a further robustness check, we rerun all regressions using geographically-consistent counties as the unit of observation rather than the hybrid of metros/geographically-consistent counties. The resulting kernels look very similar to those shown above. The only difference is the presence of some convergence among the largest counties. This difference makes sense since much of the migration out of the largest counties is capturing suburbanization. In the metro/county hybrid, a large portion of these suburbanization flows are within metro areas and so do not directly affect metro growth rates.

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 subsequent growth among relatively small locations arises from the transition of the U.S. system of locations towards a long run steady state. Specifically, 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 individual steady states. As entry waned over time, such transitional convergence died out. Those counties that remained small through more recent time periods are those with low productivity and hence low steady-state population.

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 arising from the fixed supply of land and an increase in agglomeration forces. One possibility is a 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

[7] See Appendix, Figure C.1.

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.

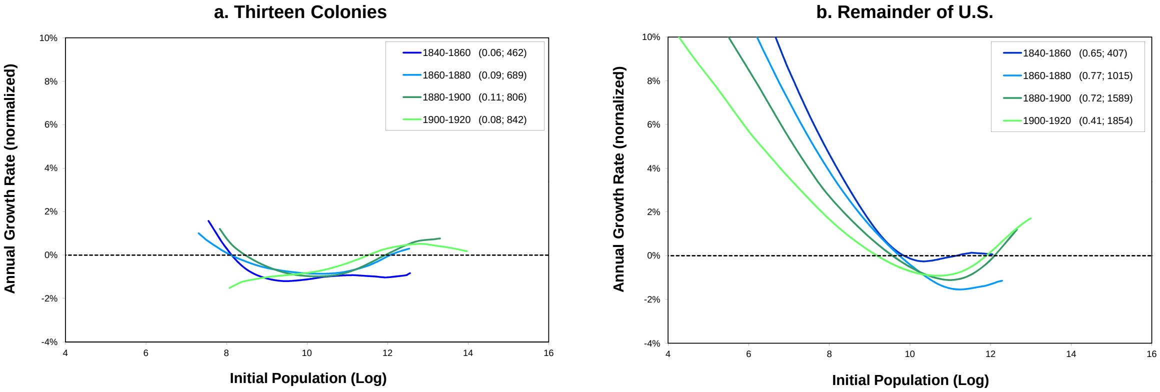


Figure 2: Growth versus Population, 1840 to 1940: Thirteen Colonies versus Remainder of U.S. Figure shows fitted kernel regressions of population growth on initial log population for geographically-consistent counties in states descended from the original thirteen U.S. colonies versus those in the remaining states. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

As a first way of distinguishing "old" from "young", we compare the correlation between growth and size among the counties that make up the initially-settled (by Europeans) thirteen colonies with counties in remaining states (Figure 2). We choose 1840 as the earliest starting year in which to consider counties in the thirteen colonies as primarily old. We choose 1900 as the last starting year in which it is possible to consider the remaining counties as primarily young. Over the twenty-year periods starting in 1840, 1860, and 1880, the counties making up the thirteen colonies did experience some small convergence (Panel A).[8] This likely reflects the presence of some younger counties in remoter areas such as those in Tennessee and Kentucky (which were respectively considered to be within Virginia

[8] For all regressions using subsets of locations, the normalization of growth rates subtracts the aggregate growth rate of *all* geographic-consistent counties in the initial year, not just those in the relevant age-specific subset.

and North Carolina prior to their statehood). Regardless, the estimated convergence dies out by a log population of 9. Over the 1900-to-1920 period, the thirteen colonies counties experienced no convergence. At higher populations—beginning at log population between 8 and 11—they experienced moderate divergence over the periods beginning in 1860, 1880, and 1900 as predicted.

Population growth among the remaining U.S. counties is sharply different. It exhibits strong convergence for each of the twenty-year time periods (Panel B). Expected growth steeply decreases as log population rises from about 7 to about 10.5.[9] Importantly, this convergence size range significantly overlaps with the size range over which the thirteen colonies counties are characterized by divergence. In other words, over an intermediate range of population, growth of the thirteen colonies' counties is characterized by divergence but growth of the remaining counties is characterized by convergence.

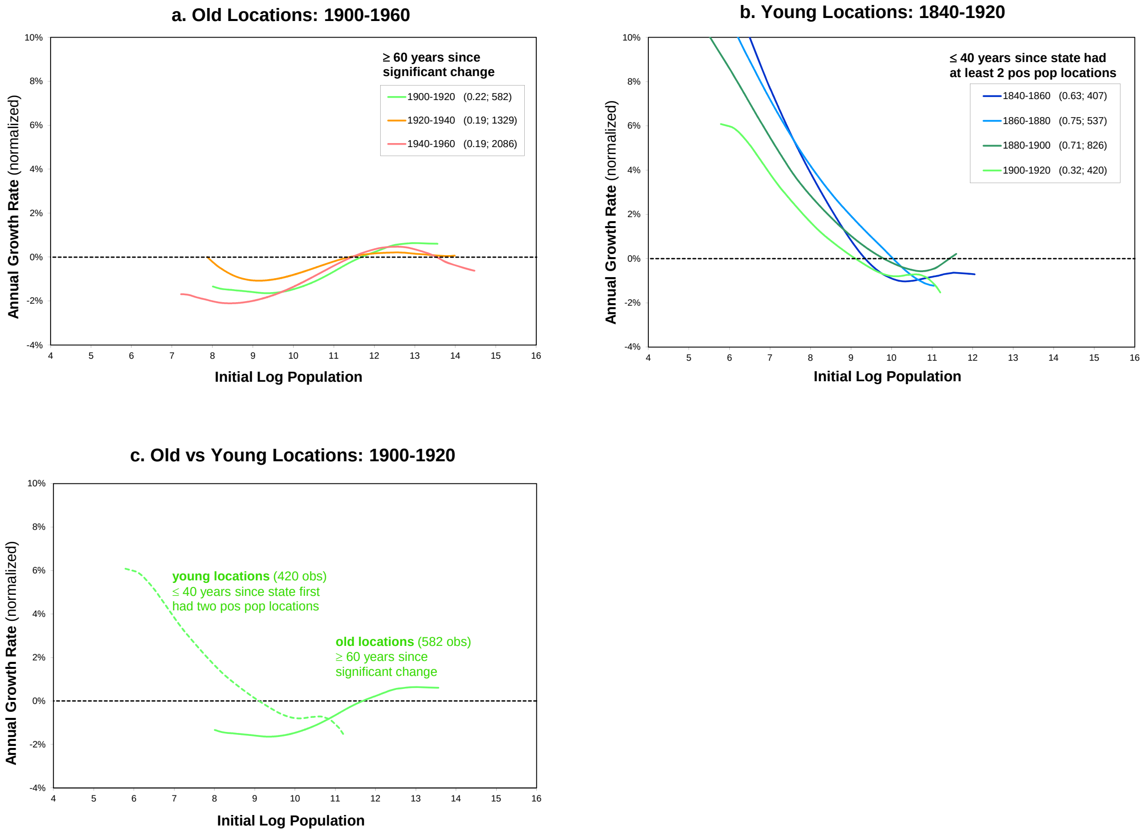


Figure 3: Population Growth vs Initial Population by Location Age. Figure shows fitted kernel regressions of population growth on initial log population for sub-samples of counties based on the number of years since they entered the system of U.S. locations. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

[9] At higher sizes there is some divergence, but the counties with such high population tend to be those which were the earliest settled among the non-thirteen colonies. By 1880 they really should be considered old. Examples of such locations include Cook County (Chicago) and St. Louis, both of which entered the U.S. system by 1810 according to an entry criterion described below and in the Appendix A.

As a second way of distinguishing age, we use the following algorithm: a geographically-consistent county is considered to be young if no more than 40 years have passed since the state or territory in which it is located first had two or more counties with positive population enumerated in a decennial census. A geographically-consistent county is considered to be old if it meets two criteria. First, at least 60 years must have elapsed since each of the enumerated counties being combined to form the geographically-consistent one experienced its final significant geographic change, typically a split or join of land area as documented in Forstall (1996). Second, at least 60 years must have passed since the CLT required any combining of enumerated counties to construct a geographic-consistent one. In other words, all old geographically-consistent counties must have been made up of a single enumerated county for at least 60 years. Typically these two criteria yield equivalent results. But for the minority of counties where they differ, we err towards falsely excluding it from the old group. Examples of applying these criteria are included in Appendix A.[10]

The dynamics for this old-young split are very similar to those for the split based on location in the thirteen colonies. For the twenty-year periods beginning in 1900, 1920, and 1940, counties classified as old show only slight convergence at low population levels (Figure 3 Panel A). They show moderate-to-strong divergence at intermediate populations. In sharp contrast, counties classified as young exhibit strong convergence in all twenty-year periods starting in 1840 through 1920 (Panel B).[11] This difference between old and young is not just driven by the older counties being on average larger. Even in the same time period, similarly sized old and young counties exhibit considerably different growth dynamics (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 date. Additionally, growth should be high initially and then decline with elapsed time. To determine the entry date of each geographically-consistent county, we use essentially the same algorithm that determines county-age, except with a 10-year cutoff rather than a 40-year one.[12] The relatively small number of counties that meet this tight entry criterion—552 over the near 200 year time span—makes clear that many actual (but unobserved) entries are be-

[10] A benefit of this methodology compared to the thirteen colonies division is that the number of locations classified as old increases over time. Conversely, the number classified as young decreases beginning in the late nineteenth century. An offsetting cost is that the algorithm fails to classify more than a handful of counties as old prior to 1900. But common sense suggests that many of these should indeed be classified as old.

[11] For counties classified as young in 1920 that had at least moderate population in that year, growth over the subsequent twenty years was approximately orthogonal. This reflects that the *proportion* of these young counties that had entered in the preceding twenty years (1900-1920) compared to those who had entered in the previous twenty years (1880-1900) was relatively low. Hence a much higher proportion of counties classified as young were near their steady state in 1920 compared to earlier starting years.

[12] The explicit criterion for "entry" is that a historically-consistent location be in a U.S. state or territory that first included two or more enumerated locations with positive population in the current or previous census. An illustration is included in Appendix A.

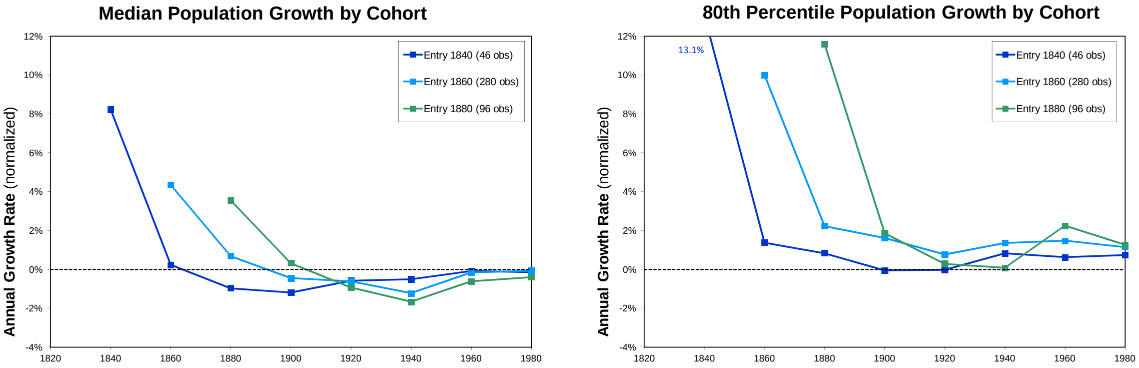


Figure 4: Population Growth by Entry Cohort. Figure shows growth trajectories for counties that entered the U.S. system of locations in the twenty years prior to each of 1840, 1860, and 1880. The displayed growth rates are the median and the 80th percentile for each cohort over each 20-year period.

ing 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.)

Both implications for growth trajectories are strongly borne out empirically. Growth trajectories of counties that became active in the ten years prior to each of 1840, 1860, and 1880 are characterized by nearly identical, high initial growth that dissipates over time (Figure 4). For each cohort, median initial growth is at least 4 percentage 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.[13] The below-aggregate growth during the later periods following entry reflects that very recently entered cohorts "steal" population from older cohorts, 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 individual steady state.

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.[14] Additionally, the young population *level* distribution should be skewed to the left and the young *growth* distribution should be skewed to

[13] The especially fast initial median population growth of the 1840 cohort may reflect that the constituent counties had higher productivity on average than that of counties that entered later.

[14] The algorithm we use classifies very few geographically-consistent counties as young prior to 1900. Hence for prior years, we focus on the contrast between young and non-young counties rather than between young and old counties.

the right. All of these are matched empirically.[15]

3.3 Population Growth Persistence

The extended transition before growth became orthogonal suggests that counties may face substantial growth frictions. If so, county growth rates should exhibit persistence. This is indeed the case. Using a four-way spline, for county/metros that experienced positive growth during an initial twenty-year period, subsequent growth is positively correlated with initial period growth.[16] Nearly all coefficients corresponding to the lagged positive growth segments in our spline estimation are positive and statistically different from zero at 0.01 level. The persistence of positive population growth weakens as growth increases. Negative initial-period growth, in contrast, is positively correlated with subsequent growth for some periods but negatively correlated with it for other periods. When the negative correlation holds, growth "bounces back" from negative rates.[17]

Persistence accounts for a significant share of the variation in growth rates. Regression R-squared values range from 0.14 to 0.31 for the initial periods through 1940. For the initial period beginning in 1960, R-squared spikes to 0.59, which suggests that some sort of shock or structural break occurred circa 1960. The acceleration of suburbanization along with the pickup in migration to the Sunbelt are the most likely candidates.[18]

4 Model

Informed by our empirical findings, we develop a simple one-sector general equilibrium model of a system of locations transitioning towards a long-run steady state. Over time there is exogenous entry of new locations into the system. The productivity of each location is drawn from a log normal probability distribution and then evolves stochastically. Agents are freely mobile, but positive local population growth leads to a friction that dampens a location's productivity.[19] Idiosyncratic shocks and agglomeration economies also affect

[15] See Appendix, Figure C.2. Consistent with Cuberes (2011), the population growth distribution across all counties, regardless of age, is highly skewed to the right. But in contrast to Cuberes, the right tail of the consolidated growth distribution is made up of young, low-population counties rather than high-population ones.

[16] See Appendix, Figure C.4.

[17] This overall pattern of persistence is similar to the pattern reported and modeled in Rappaport (2004). In contrast, Glaeser and Gyourko (2005) establish that municipal decline from 1970 to 2000 was more persistent when initial period growth was negative, and argue this captures the slow decline of durable housing. Of course, our results may change when using finer geographies than historically-consistent counties.

[18] The substitution of metro areas for counties where and when they can be identified significantly dampens the effect of suburbanization on measured persistence. However, the persistent, above-average growth of counties that abut a metro area—and which may eventually be classified to be part of that metro area—allow suburbanization to increase measured persistence.

[19] As we will later argue, in terms of its effect on the distribution of population, this is equivalent to a model with mobility frictions in which locations with high labor demand growth do not attract enough workers from other locations.

local productivity. Local population growth rates become characterized by Gibrat's law as the system of locations approaches its steady state.

4.1 Locations, Endowments and Preferences

The economy consists of N potential locations indexed by i. At time t, a number $N_t \leq N$ of locations are active. The timing of each potential location's activation is exogenous, and the order of entry is random. Denote the activation period of location i by t_i. Once active, a location remains active forever after. Each location 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,

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

That is, in each period new people enter the system. Every period each individual supplies one unit of labor where she lives and works, so 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 an agent's receipt of land income does not depend on where she lives. Agents are freely mobile across locations. Preferences are standard. An agent that enters the system in period s maximizes the presented discounted value of utility, $\sum_{t=s}^{\infty} \beta^{t-s} u(c_t)$, where c_t denotes consumption in period t, and $u(c_t)$ is the period-utility which we assume to be strictly increasing and continuously differentiable.

4.2 Production

Production is perfectly competitive and firms produce an identical non-storable good. Aggregate production in each location i is a Cobb-Douglas combination of labor and land multiplied by location- and time-specific total factor productivity (TFP), $Z_{i,t}$,

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

where α_t is the factor share of income accruing to land. Labor and land are each paid their marginal product. In particular, location wages are given by

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

TFP of location i at time t, $Z_{i,t}$, is the product of four components: a time-zero productivity draw, Z_i^0; the product of idiosyncratic shocks $\prod_{s=1}^{t} Z_{i,s}^{\text{idio}}$; a discount term arising from growth frictions, $Z^{\text{fric}}(\Delta L_{i,t}/L_{i,t-1})$; and an agglomeration factor, $Z_t^{\text{irs}}(L_{i,t})$:

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

More specifically, at time $t = 0$, each location i draws beginning productivity, Z_i^0, from a log normal distribution, $\log Z_i^0 \sim \mathcal{N}(0, \sigma_{Z^0})$. Each period, beginning at $t = 1$, each location, both active and inactive, draws an idiosyncratic shock, $Z_{i,t}^{\text{idio}}$, from a second log-normal distribution $\log Z_{i,t}^{\text{idio}} \sim \mathcal{N}(0, \sigma^{\text{idio}})$.[20] The agglomeration factor assumes that TFP endogenously increases as population increases with elasticity $\varepsilon_t \geq 0$,

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

In the numerical section, a key driving force comes from the assumed gradual decrease in the difference between the congestion coming from land, α_t, and the agglomeration parameter, ε_t. This decrease in net congestion, $\hat{\alpha}_t \equiv \alpha - \varepsilon_t$, can equivalently come from a decrease in the first component or an increase in the second component.

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 external characterization greatly simplifies the forward-looking nature of migration decisions. In the numerical section 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 $\xi_1 \geq 0$ and $\xi_2 > 0$. This specification implies that $G \in [0, 1]$, and that its first derivative is weakly positive, $G'(\cdot) \geq 0$.

4.3 Equilibrium

The decision to migrate is in principle dynamic and requires agents to compare the future time paths of utility flows across all locations. In the present setup, however, where an agent lives in the current period has no effect on her utility in future periods. Because a location's growth frictions affect all its residents equally, there are no wage differences within locations. In addition, land rents do not depend on an individual's place of residence. Together with free mobility, this implies that in each period wages equalize across locations.[21] Therefore, the forward-looking migration decision simplifies to a static decision. Since agents are homogeneous and the good the economy produces is non-storable, agents consume all their income every period. Hence the dynamic equilibrium for the model economy collapses to a sequence of static equilibria that are linked through the laws of motion

[20] The presence of idiosyncratic shocks is not strictly necessary. As we will show in the numerical part, the idiosyncratic shocks, together with the declining share of land, help account for the increasing dispersion of the employment distribution over time. If we eliminate the idiosyncratic shocks we could still account for the same increase in dispersion by assuming a slightly higher, though still plausible, decline in net congestion.

[21] This may seem counterfactual to the evidence of nominal and real wages being substantially higher in the West than in the rest of the country during the 19th century (Easterlin, 1960; Rosenbloom, 1990; Mitchener and McLean, 1999). Instead of being modeled as external, we could model growth frictions as being borne only by migrants. Such a friction would allow wages to diverge over intermediate time frames (Rappaport, 2004). This would not substantially change the population dynamics though.

for the population and the number of locations.[22] We are now ready to define the economy's *dynamic equilibrium.*

Definition of Dynamic Equilibrium *A dynamic equilibrium is a sequence of a number of locations $\{N_t\}$ and a sequence 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 as long as agglomeration forces are not too strong. This is stated in the following proposition.

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 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. ■

Abstracting from idiosyncratic shocks, once all the transition dynamics have played out, the economy reaches a *global steady state*, which we now define.

Definition of Global Steady State *In the absence of idiosyncratic shocks, the economy reaches a global steady state when all potential locations have become active, $N_t = N$, and each location's population growth equals the aggregate rate, $\forall_i \Delta L_{i,t}/L_{i,t-1} = \Delta L_t/L_{t-1}$.*

Proposition 2. *In the absence of idiosyncratic shocks and as long as all parameters are constant and $\alpha > \varepsilon$, the economy converges to a global steady state in which the population of all locations grow at the aggregate rate.*

Proof. 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

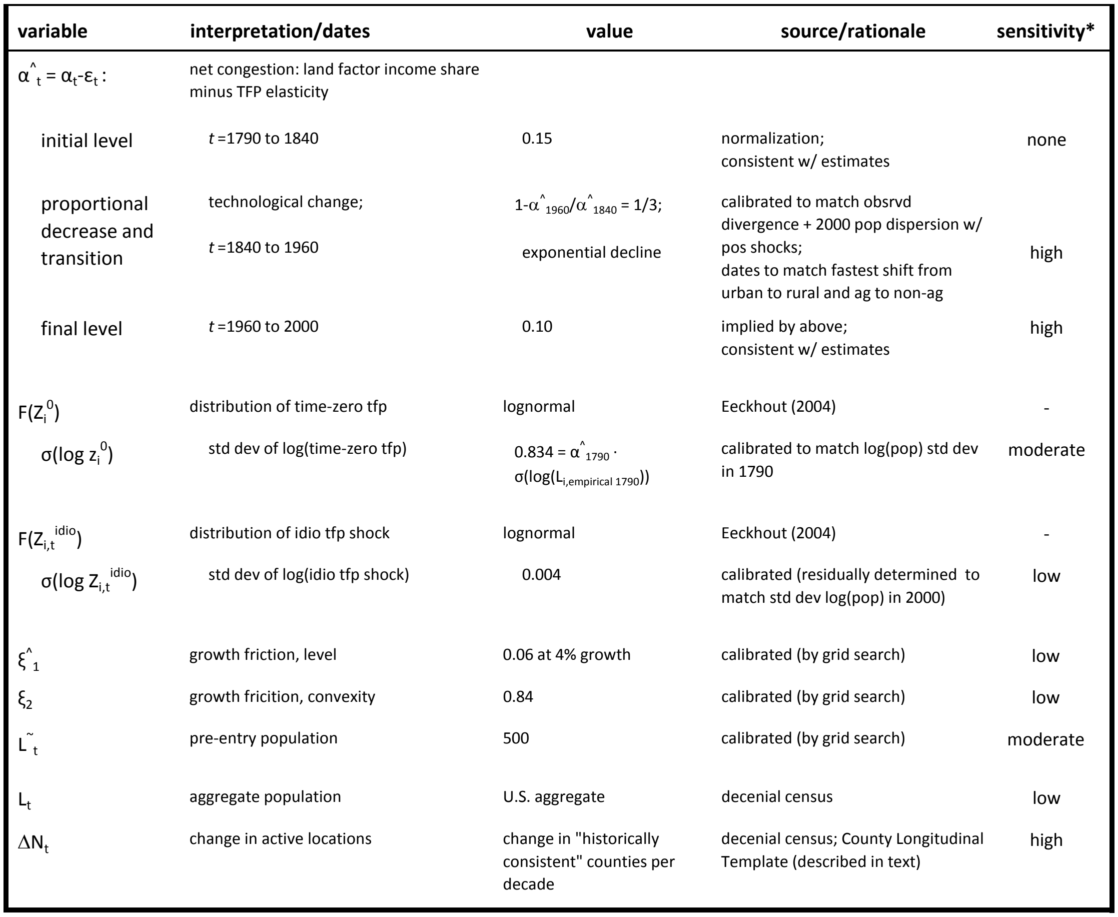
[22] There is the possibility of saving through the land market. This is the only dynamic decision, but since it does not affect the allocation of people across locations nor the equilibrium wage, we can abstract from the land market without loss of generality.

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. ■

Thus, in a deterministic setting, the economy converges to a degenerate form of Gibrat's law in which population growth rates are identical across locations. In a setting with orthogonal shocks to locations' productivity, the assumed asymmetry of the population friction may induce a slight positive correlation between population levels and growth, implying Gibrat's law may not strictly hold.

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. The parameters fall into two groupings: those that determine steady-state population and those that affect each location's transition to its steady-state population.



variable	interpretation/dates	value	source/rationale	sensitivity*
$\hat{\alpha}_t = \alpha_t - \varepsilon_t$:	net congestion: land factor income share minus TFP elasticity			
initial level	t =1790 to 1840	0.15	normalization; consistent w/ estimates	none
proportional decrease and transition	technological change; t =1840 to 1960	$1-\hat{\alpha}_{1960}/\hat{\alpha}_{1840} = 1/3$; exponential decline	calibrated to match obsrvd divergence + 2000 pop dispersion w/ pos shocks; dates to match fastest shift from urban to rural and ag to non-ag	high
final level	t =1960 to 2000	0.10	implied by above; consistent w/ estimates	high
$F(Z_i^0)$	distribution of time-zero tfp	lognormal	Eeckhout (2004)	-
$\sigma(\log z_i^0)$	std dev of log(time-zero tfp)	$0.834 = \hat{\alpha}_{1790} \cdot \sigma(\log(L_{i,\text{empirical } 1790}))$	calibrated to match log(pop) std dev in 1790	moderate
$F(Z_{i,t}^{idio})$	distribution of idio tfp shock	lognormal	Eeckhout (2004)	-
$\sigma(\log Z_{i,t}^{idio})$	std dev of log(idio tfp shock)	0.004	calibrated (residually determined to match std dev log(pop) in 2000)	low
$\hat{\xi}_1$	growth friction, level	0.06 at 4% growth	calibrated (by grid search)	low
ξ_2	growth fricition, convexity	0.84	calibrated (by grid search)	low
$\tilde{L}_t$	pre-entry population	500	calibrated (by grid search)	moderate
L_t	aggregate population	U.S. aggregate	decenial census	low
ΔN_t	change in active locations	change in "historically consistent" counties per decade	decenial census; County Longitudinal Template (described in text)	high

Table 3: Model Calibration. The parameter $\hat{\xi}_1$ is a transformation of ξ_1. It gives the friction penalty to TFP when a location is growing at a 4 percent annual rate. *Sensitivity results are with respect to moderate changes to each model parameter.

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

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

where $\hat{Z}_{i,t}(\cdot) \equiv Z^0_i \cdot \prod_{s=1}^{t} Z^{\text{idio}}_{i,s} \cdot Z^{\text{fric}}(\cdot)$ 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 B).

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 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 B). 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, $\sigma(Z^{\text{idio}}_i)$, 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.

A final set of three parameters jointly govern the friction that slows positive population growth. The level friction parameter, $\hat{\xi}_1$, specifies the penalty to location productivity when population growth is at a benchmark rate of 4 percent (the choice of the benchmark growth rate is solely for intuition and is without loss of generality). The convexity friction parameter, ξ_2, specifies the curvature with which the productivity discount increases as growth increases. Lastly, newly-entering locations can join the system with a population of $\tilde{L}$ or less without incurring any frictional discount to their TFP. Thus $\tilde{L}$ 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, $\hat{\xi}_1$, is set to 0.06; the friction convexity, ξ_2, is set to 0.84 (moderately concave); and the pre-entry population proxy, $\tilde{L}$, is set to 500 ($\log \tilde{L} = 6.2$). These level and convexity parameters imply a degree of labor mobility consistent with empirical estimates in Gallin (2004) and Rappaport (2004). The sensitivity of results to these choices is discussed in an online appendix.[23]

The two remaining variables to be calibrated are the total population and the number of active locations in each year. The former is assigned to match the population of the continental United States. The latter is assigned to match the number of geographically-consistent locations in each decennial year (Table 1 columns 3-4). Together these two assumptions allow for a clean comparison of simulated population levels with the geogaphically-consistent empirical results.[24]

6 Numerical Results

The calibrated model approximately matches the population dynamics of U.S. counties and metro areas over each of the ten twenty-year periods from 1800 to 2000. This approximate match holds for population convergence and divergence, for the growth trajectories over time of newly entering locations, for the distribution of population levels, for the distribution of growth rates, for the persistence of growth rates, and for the differential growth patterns of young and old locations.

6.1 Simulated Transitional Growth

To measure the match of simulated to empirical convergence and divergence, simulated growth is regressed on the same initial population spline used in the empirical analysis. For each of the twenty-year intervals, this is repeated for each of 400 stochastic seeds. The mean values of the resulting 400 sets of coefficients are the basis for fitted simulated curves, which can then be compared to the fitted empirical curves.

The overall pattern of simulated growth closely matches that of empirical growth throughout the 200 years from 1800 to 2000 (Figure 5 versus Figure 1). During the early nineteenth century simulated growth is characterized by significant convergence from low population levels as newly entering locations transition to their "individual steady states". The latter is simply the population level consistent with a location growing at the aggregate rate.

[23] See www.eco.uc3m.es/~desmet/ or www.kansascityfed.org/speechbio/rappaport.cfm.

[24] For results on population growth, in the empirical section we argued that it is more sensible to work with the hybrid county-metro data build (Table 3 columns 5-6). The main reason is that any congestion and agglomeration forces are likely to be determined at the metro level). Since in the numerical exercises we do not combine suburban counties with core cities, there will be a difference between modeled and empirical growth.

Equivalently, it is the population level that allows a location to match the system-wide shared wage when it experiences the friction associated with growth at the aggregate rate. At higher population levels, growth is approximately orthogonal to size reflecting that large locations are typically close to their individual steady state.

Beginning very slightly in 1860 and then strengthening, population growth is characterized by moderate divergence at intermediate and high population levels.[25] Population convergence ends in about 1940, reflecting the dwindling number of entering locations. And the continuing divergent population growth attenuates following the end of the drop in net congestion in 1960. By the 1980-2000 period, simulated divergence is mostly absent.

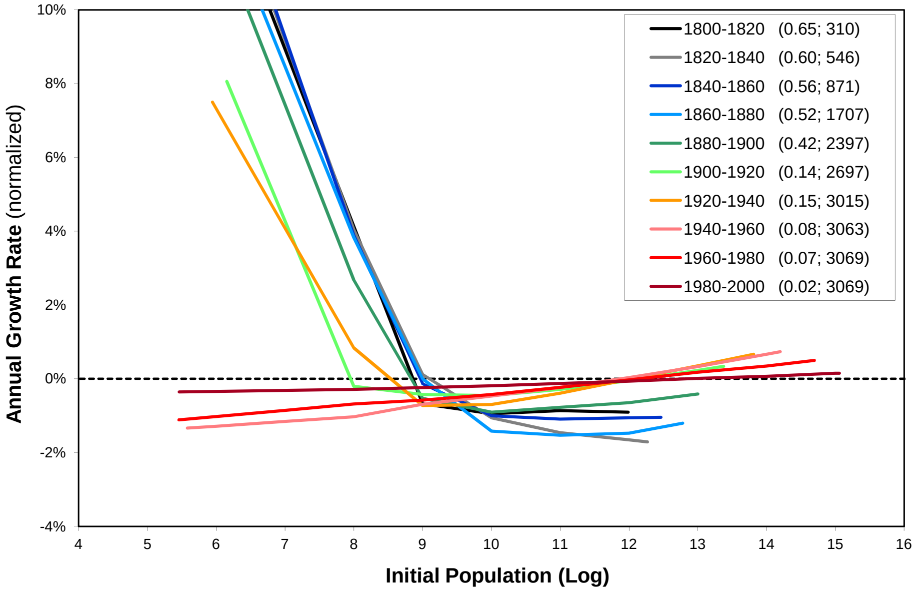


Figure 5: Simulated Population Growth vs Initial Population, 1800-2000.

The close match of simulated to empirical convergence and divergence is especially evident in a period-by-period comparison of fitted growth rates (Figure 6). For the twenty-year periods starting in 1840 and 1860, the match is almost exact (top left panel). For the periods starting in 1880 and 1900, simulated convergence at low levels is modestly steeper than is empirical convergence; at higher population levels, simulated divergence is slightly more gradual than is empirical divergence (top right panel). For the subsequent two periods, the key result is that simulated convergence continues from 1920-1940 but is absent from 1940-1960. This exactly matches the timing of the end of empirical convergence (bottom left panel). Simulations using alternative scenarios make clear that this collapse of convergence

[25] The beginning of simulated divergence lags the start of the underlying technology change by twenty years. This reflects that the simulated system remains far from its global steady state in 1840. Specifically many high-productivity locations are still transitioning up and other, intermediate-productivity, locations are drifting down. Decreasing $\hat{\alpha}$ when the system is at a steady state immediately causes divergence.

arises from the dwindling of location entry during the 1930s rather than from the sharp declines in entry that occurred during earlier decades. Between 1960 and 1980, modest simulated divergence almost exactly matches empirical divergence except at high populations, where empirical growth is approximately orthogonal (bottom right panel, bright red lines). Finally, simulated growth from 1980 to 2000 is approximately orthogonal, reflecting that the simulated system has nearly attained its steady state. In contrast, empirical divergence modestly strengthens in the 1980-2000 period. This suggests that the distance between the actual population distribution and the steady state population distribution may actually have widened circa 1980. We will discuss this possibility in a later subsection.

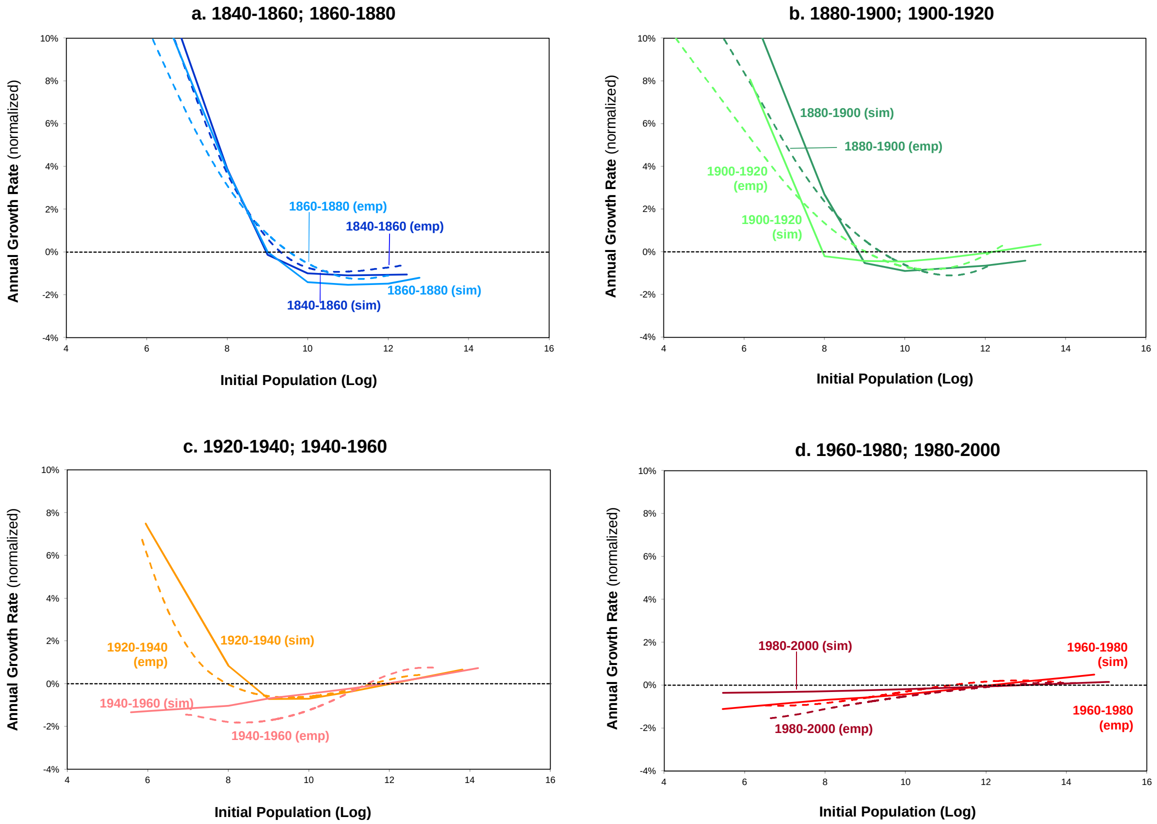


Figure 6: Population Growth Fitted on Initial Population: Simulated Versus Empirical. Population growth rates are normalized by subtracting the aggregate growth rate of locations existing locations at the start of each twenty-year period.

The ability of population levels to account for population growth differs between the simulated and empirical regressions. The simulated R-squared value gradually declines, from 0.65 for the twenty-year period beginning in 1800 to 0.42 for the one beginning in 1880, and then jumps down to 0.14 in 1900. Thereafter it declines gradually to just 0.02 in 1980. This overall downward trend reflects that the simulated system is steadily transitioning to a long-run steady state, at which normalized growth is exclusively driven by the idiosyncratic shocks. The empirical regressions, in contrast, have an R-squared value that rises from 0.43 in 1800 to 0.70 in 1880. It then jumps down to 0.31 in 1900, after which it declines steadily to 0.07 in 1980. The initially lower empirical R-squared values likely reflect the multitude of forces driving population movements in the newly-born United States. The identical timing in 1900 of the downward jump in R-squared values reflects, at least in part, the significant decline in location entry during the 1890s. The higher empirical R-squared values for the 1860 through the 1980 periods reflects, at least in part, that simulated transitions are moderately faster than are empirical ones. As a result, a larger portion of simulated locations during these periods will be at their individual steady state and so their growth will be approximately orthogonal to size.

6.2 Simulated Transitional Growth and Location Age

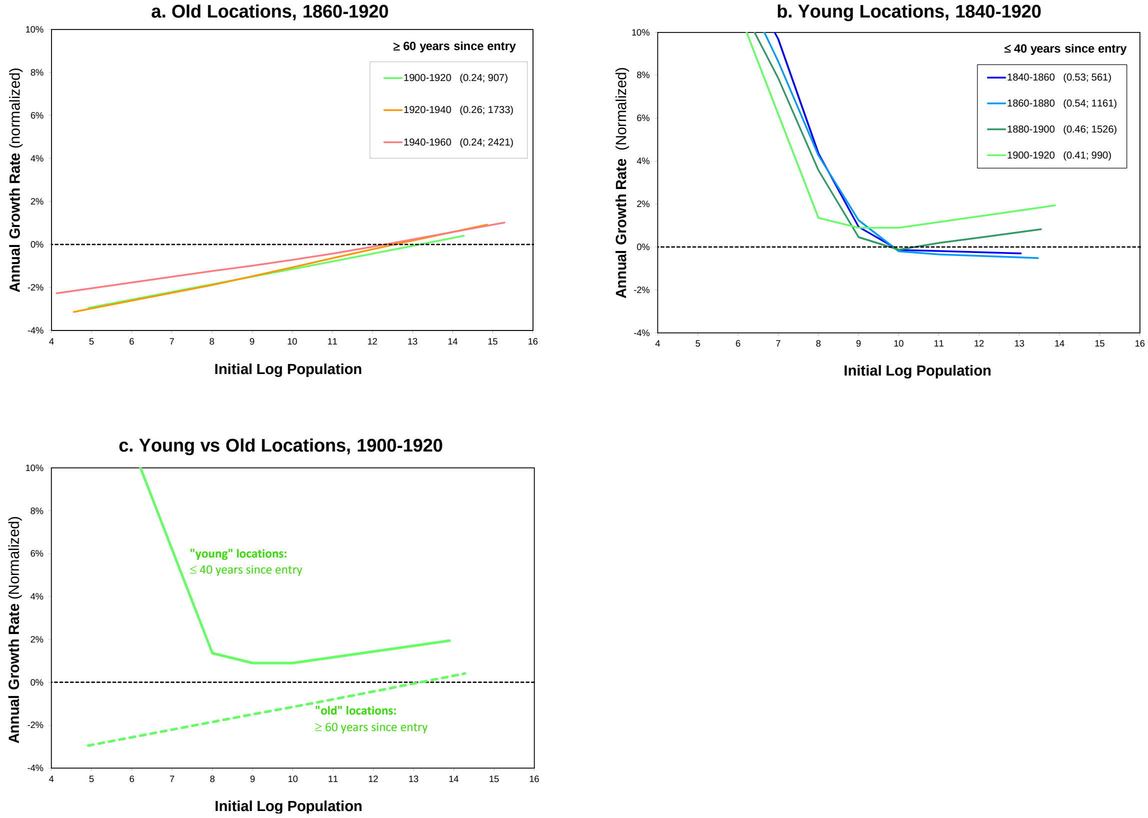


Figure 7: Simulated Population Growth vs Initial Population by Location Age. Figure shows fitted spline regressions of population growth on initial log population for sub-samples of locations based on the number of years since they entered the modeled system of locations. Fitted growth rates are normalized by subtracting the aggregate growth rate of *all* locations active at the start of each period.

Simulated convergence, as is hypothesized for empirical convergence, is driven solely by the entry and subsequent upward transition of locations to their individual steady state.

Similar to its empirical counterpart above, Figure 7 divides locations into "young" (≤ 40 years since entry) and "old" (≥ 60 years since entry). As is readily seen, growth fitted on log population is characterized by convergence only for the young locations (Panel B). In contrast, growth fitted on log population for old locations is characterized by divergence over the entire rang of log population (Panel A). Because of this age distinction, convergence and divergence can coexist in the same time period (Panel C).

Simulated growth trajectories match empirical ones moderately closely (Figure 8 versus Figure 4). The simulated trajectory shapes are essentially identical to those of the empirical trajectories, but the growth rates themselves are lower. This is most obvious for the trajectories of the 1840 cohort. For that cohort, growth rates from 1840 to 1860, both for the median and at the 80th percentile, fall short of their empirical counterparts by almost 4 percentage points. For the remaining trajectory periods and cohorts, simulated growth falls short of empirical growth by at least 1 percentage point.

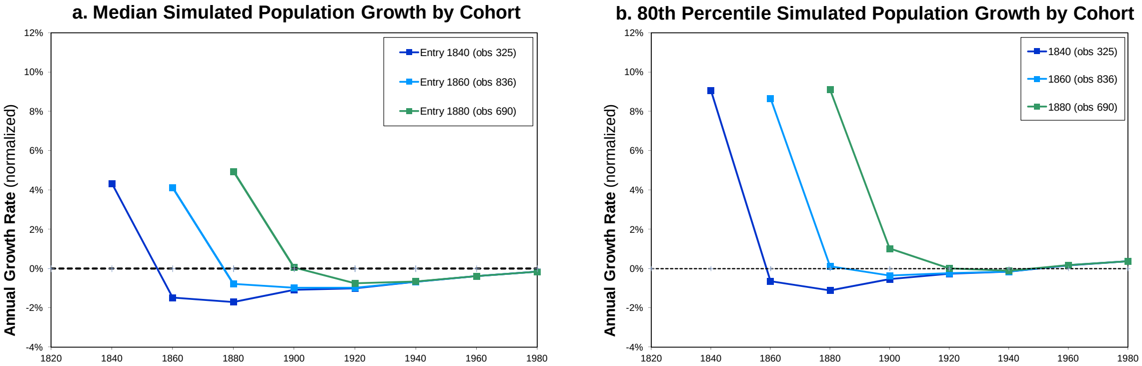


Figure 8: Simulated Population Growth by Entry Cohort. Figure shows growth trajectories for counties that became active in the twenty years prior to each of the listed entry years. The displayed growth rates are the median and the 80th percentile for each cohort over each 20-year period.

Notwithstanding slightly lower growth rates, the simulated transition durations are moderately shorter than the empirical durations. As with the empirical trajectories, duration will be measured by the time required for a given simulated cohort's trajectory to match the trajectory of the previous cohort. For example, the median growth rate of the 1860 cohort matches that of the 1840 cohort after 40 years. Similarly, the 1880 median trajectories, and the 1860 and 1880 80th percentile trajectories each have a duration of 40 years. These transition times match the corresponding empirical ones for the 1880 cohort. But for the 1860 cohort, both median and 80th percentile simulated durations are 20 years shorter than the corresponding empirical durations. The simulated trajectories exhibit slower growth and have shorter durations, compared to the empirical ones. This suggests that at least some growth frictions may depend on absolute changes in population rather than relative ones.[26]

[26] As with the empirical trajectories, the below aggregate growth portions of the simulated trajectories

Together, the fitted convergence/divergence by age relationships and the growth trajectories strongly suggest that simulated younger locations tend to be smaller and grow quicker than simulated non-young locations. This is indeed the case. The distribution of simulated young location populations by age in a representative year, 1860, is significantly shifted to the left from the distribution of non-young location populations. This approximately matches the empirical distributions by age except that the empirical young distribution has a significantly fatter lower tail. The distribution of simulated young location growth rates is significantly skewed to the right, which contrasts with a relatively symmetric simulated non-young location growth distribution. This pattern again matches the comparable empirical growth distributions.[27]

Simulated persistence qualitatively matches empirical persistence in several important ways.[28] First, expected second-period growth depends positively on positive initial-period growth. Second, the simulated persistence spline is concave over positive initial-period growth rates, as is also the case with the empirical spline. Third, initial-period simulated growth accounts for a large share of the variation of subsequent-period growth. The fitted simulated persistence spline also differs from the empirical one in one important way. The coefficient on the negative initial-period growth segment is always positive. This largely reflects the gradual population decline of low-productivity locations due to the decreasing importance of land. In contrast, the empirical coefficient on negative initial growth is positive for some initial periods and negative for others.

6.3 Simulated Growth and Level Distributions

The distribution of simulated growth rates across all locations approximately matches the changing distribution of empirical growth throughout the nineteenth and twentieth centuries (Figure 9).

During the nineteenth century, both the simulated and empirical growth distributions are significantly skewed to the right (top panels). Compared to the empirical distributions, the simulated ones are denser at low growth rates but less dense at intermediate ones. These differences balance each other out in the sense that the simulated and empirical growth means are approximately equal.

During the twentieth century, the dispersion of both the simulated and empirical growth rates is much narrower (bottom panels). The right tails are mostly absent, reflecting that relatively few locations remain far below their steady-state populations. As described above, the simulated system has approximately reached its long-run steady state by 1980. The empirical system, however, has probably not. Consistent with the continued transition of

reflect the disproportionate absorption of aggregate population growth by entering locations.

[27] Figures that show the population level and population growth distributions for young vs non-young in several years are included in Appendix, Figure C.3.

[28] See Appendix, Figure C.4.

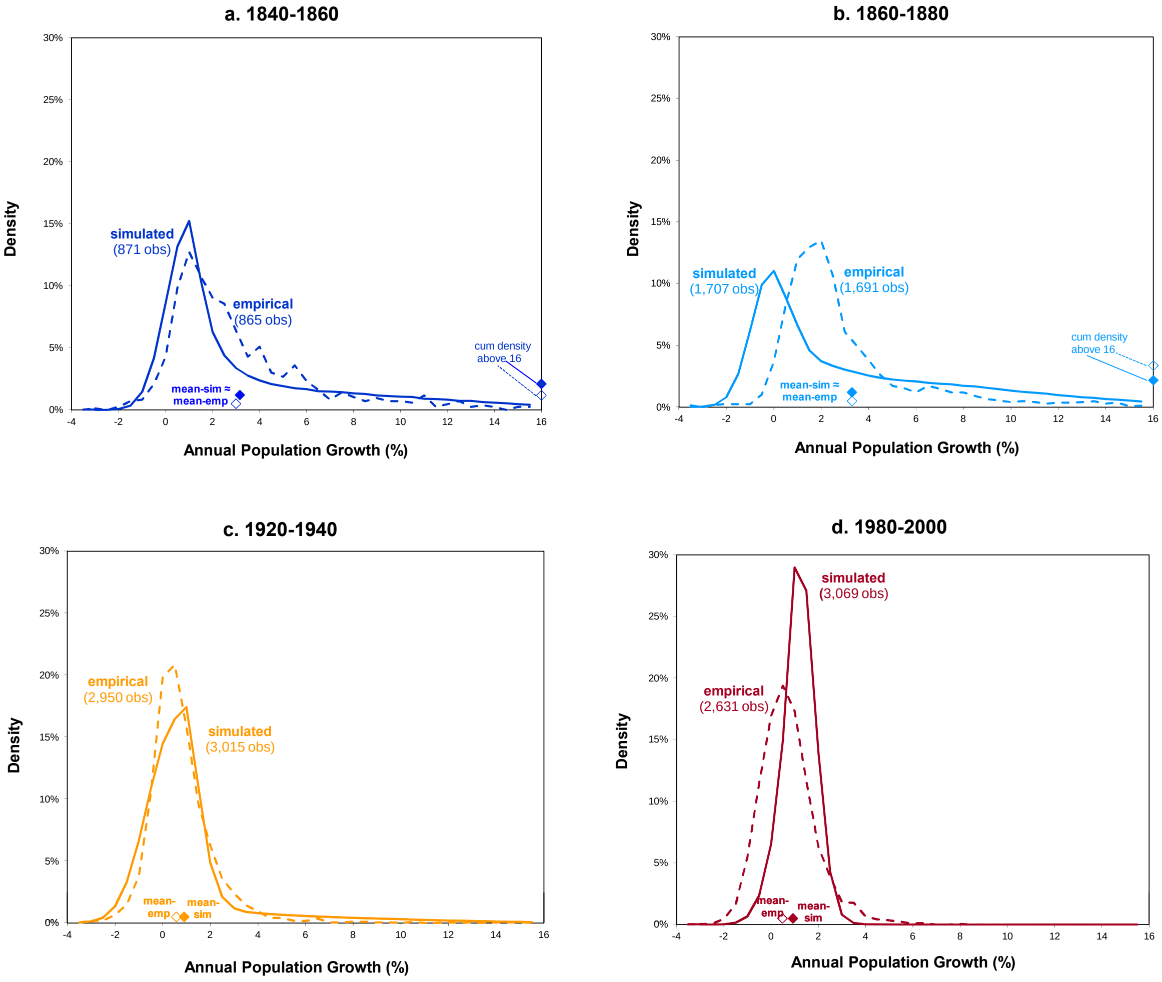


Figure 9: Population Growth Distributions, Simulated vs Empirical. Growth rates are not normalized.

the empirical system, its growth distribution is shifted to the left relative to the simulated distribution and is also more dispersed. As will be discussed in one of the counterfactual exercises, these differences would follow if net congestion resumed decreasing during the 1980s and 1990s.

The simulated distributions of log population levels do an even better job of matching their empirical counterparts (Figure 10). In 1790 (the year in which the simulation actually starts), the simulated level distribution almost exactly overlays the empirical one (Panel A). As the simulated level distribution is log normal by construction, the empirical level distribution must be log normal as well. The 2000 simulated level distribution, which too is log normal by construction, approximately matches the empirical level distribution (Panel D). However, the empirical distribution is moderately shifted to the left relative to the simulated one. The empirical log normality of the population distribution in 1790 is a key piece of evidence suggesting that the current log normal population distribution is unlikely

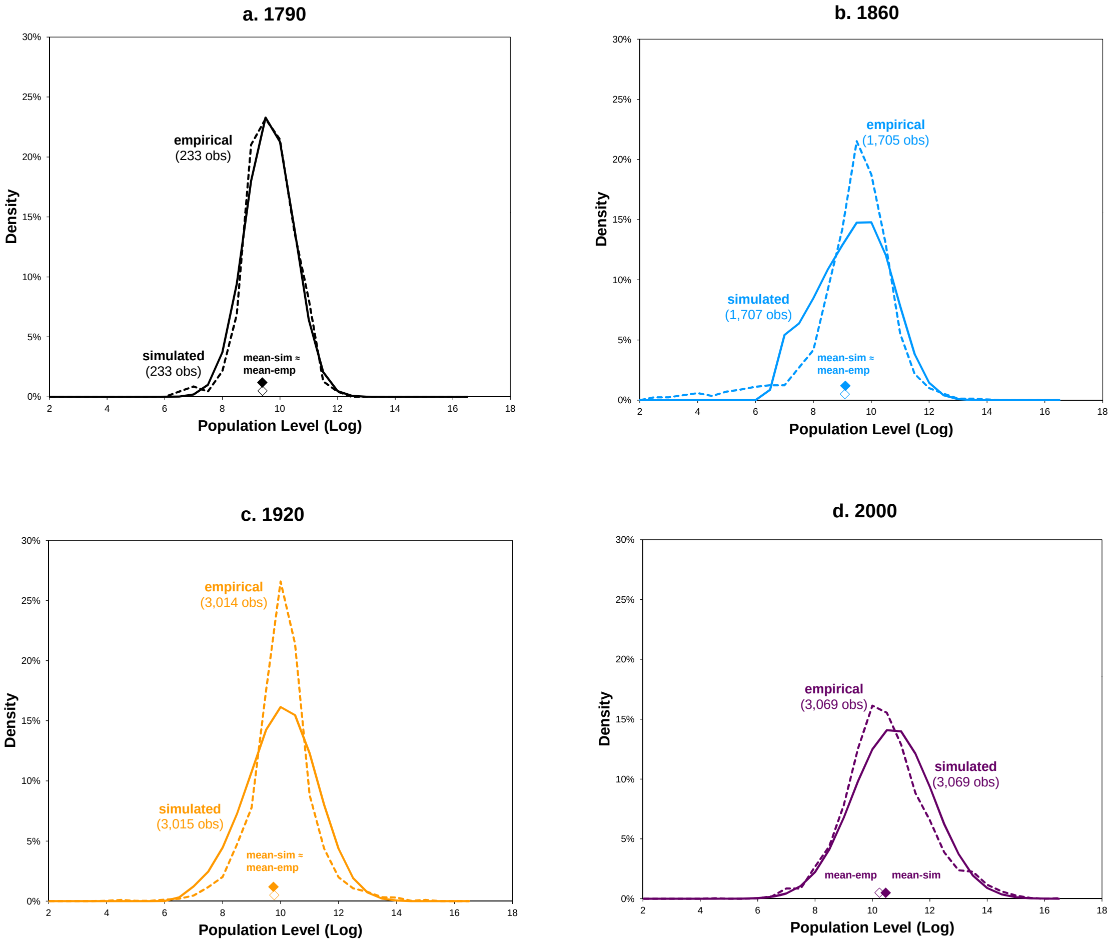


Figure 10: Population Level Distributions, Simulated vs Empirical. Note that the empirical distributions here are based on all U.S. counties rather than the hybrid of counties and metro areas.

to be the asymptotic result of orthogonal growth.

During the interim between 1800 and 2000, the simulated population level distributions tend to be moderately more dispersed than the empirical ones. One reason is that the calibrated entry population level ($\tilde{L} = 500$ or equivalently, $\log \tilde{L} = 6.2$) may be somewhat low. For example, the high density of the 1860 simulated population distribution for log population between 6 and 8 (panel B) diminishes when $\log \tilde{L}$ is assumed to be larger. A possible additional reason for the greater simulated dispersion is linked to the assumed gradual decrease in net congestion. By construction, the decrease in simulated net congestion captured by the decline of $\hat{\alpha}_t$ increases the simulated *steady-state* population dispersion while maintaining its log normal form. By 1920, the increase in the simulated dispersion exceeds the increase in the empirical dispersion (Panel C). This might be because the calibrated friction level is too low. A too-low calibrated friction would also help account for the comparatively fast simulated transitions following entry.[29]

6.4 Alternative Scenarios

The model's success in matching such a wide range of empirical relationships throughout 200 years of varying U.S. population and geographic development suggests that it can provide insights into alternative, counterfactual, and hypothetical scenarios. What would U.S. development have looked like if frictions were much smaller or larger? How would an early end to the U.S. westward expansion have affected population dynamics? What would recent and future growth look like if the decrease in net congestion resumed?

Growth Friction. Conventional wisdom holds that the U.S. has generally been characterized by relatively high labor mobility and so implicitly low growth frictions. Within the framework of the model, in the complete absence of frictions, all locations would immediately jump to their local steady state upon entry. Hence there would be no observed convergence. With a positive, but very low growth friction (an assumed loss of productivity of 1 percent rather than 6 percent at 4 percent annual growth), convergence would end in 1900, forty years earlier than under the baseline (Figure 11, Panel A). The divergence of population would also end earlier: in 1960 rather than 1980. More specifically, population growth would revert to near orthogonality immediately following the stabilization of net congestion in 1960 rather than needing twenty more years to transition.

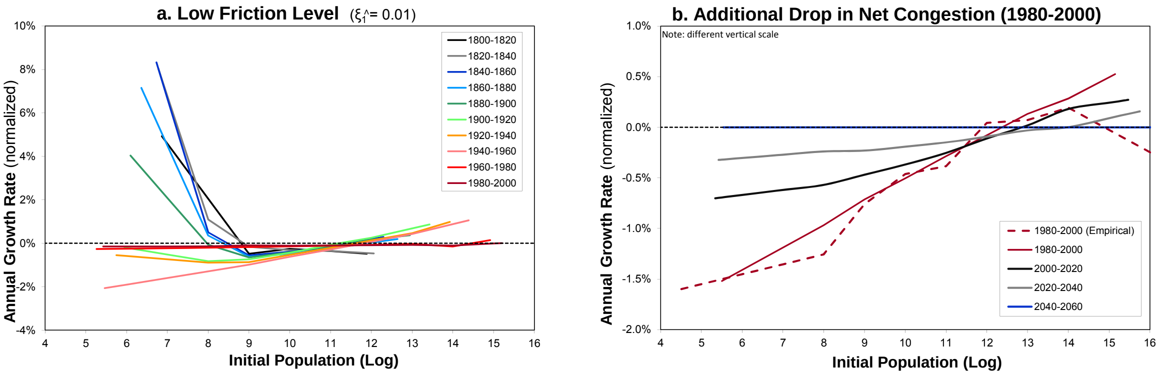


Figure 11: Alternative Scenarios. Panel A: growth friction equal to 1 percent of TFP rather than the baseline 6 percent at 4 percent annual growth ($\hat{\xi}_1 = 0.01$). Panel B: Net congestion decreases from $\hat{\alpha} = 0.100$ to $\hat{\alpha} = 0.095$ from 1980 to 2000.

[29] An additional difference between the simulated and empirical distributions is the fat lower tail in the empirical one. This can be seen in the 1860 empirical distribution; it also characterizes the 1880 one (not shown). In other words, there were significantly fewer very small simulated locations in these years than there were empirical ones.

Early End to Location Entry. The discussion so far has emphasized the key role of location entry in driving aggregate dynamics. In particular, the end of entry causes convergence to end shortly thereafter. Consider the alternative scenario under which entry matches the baseline until 1860 and is zero thereafter. In that case, the 1860-1880 period would be the last during which convergence is observed. Locations that had become active immediately before 1860 (all of which are still small in 1860) would continue to exhibit convergent growth during these twenty years. But thereafter, any residual convergence would be completely masked by the divergence associated with the ongoing gradual decrease in net congestion. Essentially, the switch from convergent growth to divergent growth among smaller locations would occur 60 years earlier than under the baseline.

Decrease in Net Congestion from 1980 to 2000. The implicit share of information and communications technologies in aggregate U.S. output began to increase rapidly during the mid 1970s. Desmet and Rossi-Hansberg (2009) argue that the introduction of such a general purpose technology increases the incentive to agglomerate. Hence net congestion may have resumed decreasing during the last decades of the 20th century. We model this as a gradual decline of $\hat{\alpha}$ from the baseline value of 0.100 to 0.095 between 1980 and 2000. For perspective, this 5 percent decrease over twenty years corresponds to a more moderate rate of decline than the baseline one-third decrease over 120 years. Compared to the baseline, divergence from 1980 to 2000 becomes stronger (Figure 11, Panel B). Divergence diminishes but remains present over the subsequent forty years. Beginning in 2040, the system resumes orthogonal growth. The increase in divergence under this alternative scenario during the period 1980-2000 is consistent with the increase in divergence observed in the data (Figure 1) which was absent in the numerical simulations (Figure 5). This supports the hypothesis of stronger agglomeration economies coinciding with the introduction of information technology.

7 Conclusions

This paper has studied the long run development of the U.S. system of locations between 1800 and 2000. From the beginning, population growth across young, smaller locations was characterized by convergence. Then, as the entry of new locations dwindled during the early twentieth century, convergence first dissipated and then disappeared. Beginning in the mid nineteenth century, population growth across older, larger locations was characterized by divergence. This moderated during the second half of the twentieth century but never fully died out. In consequence, the orthogonality of local population growth to initial population levels is decisively rejected throughout almost all of U.S. history. Given that, it would seem impossible for orthogonal population growth to be the cause of the recently-observed log normal population distribution. An explanation more consistent with observed dynamics is that orthogonal growth is a consequence of a system of locations being in its steady state, whatever its distribution.

A simple one-sector model of a system of locations closely matches observed dynamics. For an extended period, new locations continually enter the system with low initial population. Transitions up to their steady-state population are slowed by an external friction on population growth. Overlapping this extended period of entry, the net congestion arising from the fixed supply of land net of any agglomeration economies begins to gradually lessen. As a result, the steady-state population distribution begins to diverge. But such divergence is only observed across larger locations. Across smaller ones, divergence is masked by even stronger convergence until location entry dies out.

With only a handful of parameters, the model is able to match a wide range of empirical relationships that have characterized U.S. local growth over the last 200 years. 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, this modeling framework may help us understand other economies where the geographic distribution of population is still rapidly changing, such as China and Brazil. Of particular interest is how the relationship between population growth and size is likely to evolve over the next decades. Similarly, how will different public policies, such as the restriction of mobility through the Chinese *Hukou* system, affect it. More generally, the model can be used to help understand and project the effect on local growth dynamics of different shocks to technology, productivity and net congestion.

References

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

[2] Caselli, F. and Coleman, W.J., 2001. "The U.S. Structural Transformation and Regional Convergence: A Reinterpretation," *Journal of Political Economy*, 109, 584-616.

[3] Ciccone, A., 2002. "Agglomeration Effects in Europe." *European Economic Review*, 46, 213–227.

[4] Conley, T., 1999. "GMM rstimation with cross sectional dependence," *Journal of Econometrics*, 92, 1-45.

[5] Cuberes, D., 2011. "Sequential City Growth: Empirical Evidence," *Journal of Urban Economics*, 69, 229-239.

[6] Davis, D.R. and Weinstein, D.E, 2002. "Bones, Bombs, and Break Points: The Geography of Economic Activity," *American Economic Review*, 92, 1269-1289.

[7] Davis, M.A. and Heathcote, J., 2007. "The Price and Quantity of Residential Land in the United States," *Journal of Monetary Economics*, 54, 2595-2620.

[8] Desmet, K. and Fafchamps, M., 2006. "Employment concentration across U.S. counties," *Regional Science and Urban Economics*, 36, 482-509.

[9] Desmet, K. and Rossi-Hansberg, E., 2009. "Spatial growth and industry age," *Journal of Economic Theory*, 144, 2477-2502.

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

[11] Easterlin, R.A., 1960. "Interregional Difference in Per Capita Income, Population, and Total Income, 1840-1950," in: *Trends in the American Economy in the Nineteenth Century, Studies in Income and Wealth Vol. 24*, ed. W. Parker, 73-140, Princeton, NJ: Princeton University Press.

[12] Eaton, J. and Eckstein, Z., 1997. "Cities and growth: Theory and evidence from France and Japan," *Regional Science and Urban Economics*, 27,] 443-474.

[13] Eeckhout, J., 2004. "Gibrat's Law for (All) Cities," *American Economic Review*, 94, 1429-1451.

[14] 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

[15] Gabaix, X., 1999. "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics*, 114, 739-767.

[16] Gallin, J. H., 2004. "Net Migration and State Labor Market Dynamics." *Journal of Labor Economics*, 22, 1-23.

[17] Gardner, T., 1999. "Metropolitan Classification for Census Years before World War II", *Historical Methods*, 32, 3, 139–150.

[18] Gibrat, R., 1931. *Les inégalités économiques: applications aux inégalités de richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel*, Paris: Librairie du Recueil Sirey.

[19] Giesen, K. and Südekum, J., 2012. "The Size Distribution across All Cities: A Unifying Approach," unpublished manuscript.

[20] Glaeser, E.L. and Gyourko, J. 2005. "Urban Decline and Durable Housing," *Journal of Political Economy*, 113, 345-375.

[21] Glaeser, E.L., Scheinkman, J.A., and Shleifer, A., 1995. "Economic Growth in a Cross-Section of Cities," *Journal of Monetary Economics*, 36, 117-143.

[22] Gónzalez-Val, R., Sánchez-Vidal, M., and Viladecans-Marsal, E., 2012. "Sequential City Growth in the U.S.: Does Age Matter?", unpublished manuscript.

[23] 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].

[24] Hansen, G. and Prescott, E.C., 2002. “Malthus to Solow,” *American Economic Review*, 92, 1205-17.

[25] Holmes, T.J. and Lee, S., 2010. “Cities as Six-by-Six-Mile Squares: Zipf’s Law?,” NBER Chapters, in: *Agglomeration Economics*, 105-131, National Bureau of Economic Research.

[26] Horan, P.M. and 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.

[27] Ioannides, Y.M. and Overman, H.G., 2003. “Zipf’s Law for Cities: An Empirical Examination,” *Regional Science and Urban Economics*, 33, 127-137.

[28] Jackson, J.R., Johnson, R.C., and Kaserman, D.L., 1984. “The Measurement of Land Prices and the Elasticity of Substitution in Housing Production.” *Journal of Urban Economics* 16, 1–12.

[29] Jorgenson, D.W., Ho, M.S., and Stiroh, K.J., 2005. “Growth of U.S. Industries and Investments in Information Technology and Higher Education.” in *Measuring Capital in the New Economy*, eds. Corrado, C., Haltiwanger, J., and Sichel, D., Chicago IL: University of Chicago Press.

[30] Krugman, P., 1996a. *The Self-Organizing Economy*, Cambridge, MA: Blackwell.

[31] Krugman, P., 1996b. “Confronting the Mystery of Urban Hierarchy,” *Journal of the Japanese and International Economies*, 10, 399-418.

[32] Lee, S. and Li, Q., 2013. “Uneven Landscapes and City Size Distributions,” *Journal of Urban Economics*, forthcoming.

[33] Michaels, G., Rauch, F. and Redding, S., 2012. “Urbanization and Structural Transformation,” *Quarterly Journal of Economics*, 127, 535-586.

[34] Mitchener, K.J. and McLean, I.W., 1999. “U.S. Regional Growth and Convergence, 1880-1980,” *Journal of Economic History*, 59, 1016-1042.

[35] Mundlak, Y., 2001. “Production and Supply,” in *Handbook of Agricultural Economics*, eds. Gardner, B. and Rausser, G., Elsevier Science B.V.

[36] Rappaport, J., Sachs, J.D., 2003. “The United States As a Coastal Nation,” *Journal of Economic Growth*, 8, 5-46.

[37] Rappaport, J., 2004. “Why Are Population Flows So Persistent?” *Journal of Urban Economics*, 56, 554-580.

[38] Rappaport, J., 2007. “Moving to Nice Weather,” *Regional Science and Urban Economics* 37, 375-398.

[39] Rappaport, J., 2008. “Consumption Amenities and City Population Density,” *Regional Science and Urban Economics* 38, 533-552.

[40] Rosen, K. and Resnick, M., 1980. “The Size Distribution of Cities: An Examination of the Pareto Law and Primacy,” *Journal of Urban Economics*, 8, 165186.

[41] 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.

[42] Solow, R.M., 1973. “Congestion Cost and the Use of Land for Streets,” *Bell Journal of Economimcs and Management Science*, 4, 602-618.

[43] Thorndale, W. and Dollarhide, W., 1987. *Map Guide to the Federal Censuses, 1790–1920.* Baltimore: Genealogical Publishing Company.

[44] 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.

[45] U.S. Bureau of the Census, 1975. *Historical Statistics of the United States Colonial Times to 1970 Bicentennial Edition*, Washington D.C.

Appendix A: Data Construction Details

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 into two or more of a single, previously-settled county. 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 293 counties enumerated in the 1800 census into 233 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 data for the 417 counties enumerated in the 1820 decennial census, these are also additively collapsed into 233 counties.

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.

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. All of the counties must have no more than one-third of employed persons working in the agricultural sector, at least 10,000 nonagricultural workers, and at least 10 percent 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 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.

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 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.

Determining age and entry year

The algorithmic approach to 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 criterion one 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 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 the near 200 year 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: 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 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 percent (Jackson, Johnson, and Kaserman, 1984; Thorsnes, 1997). Also suppose that the housing service share of household consumption was 15 percent, 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 points. The additional 0.03 percentage point contribution required to attain a 0.15 aggregate land share of factor income requires that the remaining 52 percent 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 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 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 percent 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 percent rate of structure depreciation suggested by Davis and Heathcote and a 4 percent required real rate-of-return) implies a 39 percent 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 points. 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, Ho, and Stiroh, 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 percentage point. Using a one-third consumption expenditure share for manufacturing residually implies a 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 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 (1974) 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.

Appendix C: Additional Figures

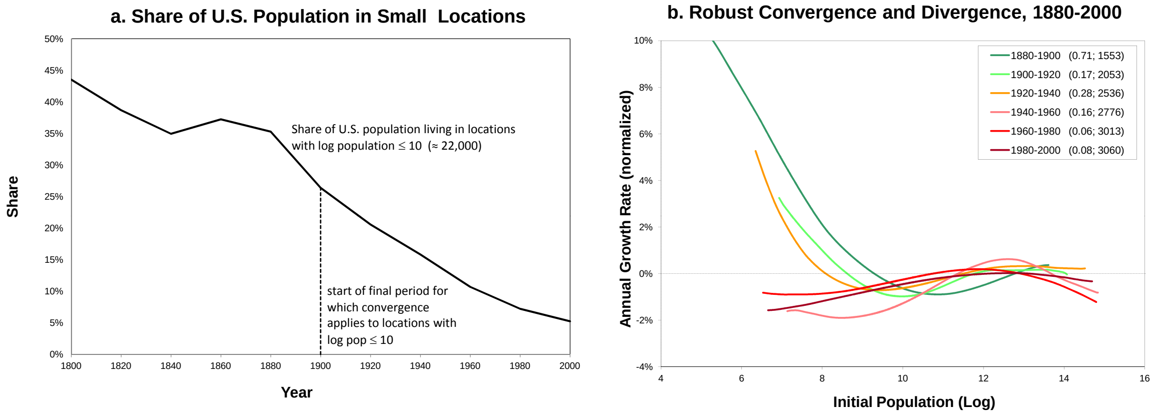


Figure C.1: Empirical Relevance and Robustness of Convergence

Panel A shows share of U.S. population living in county/metros with log population ≤ 10 (population $\lesssim 22,000$). Panel B shows fitted spline regressions of county growth on initial log population using a non-metro build of the data. County observations are combined to match the composition of their predecessor locations 40 years earlier. Fitted growth rates are normalized by subtracting the aggregate growth rate of all locations active at the start of each period.

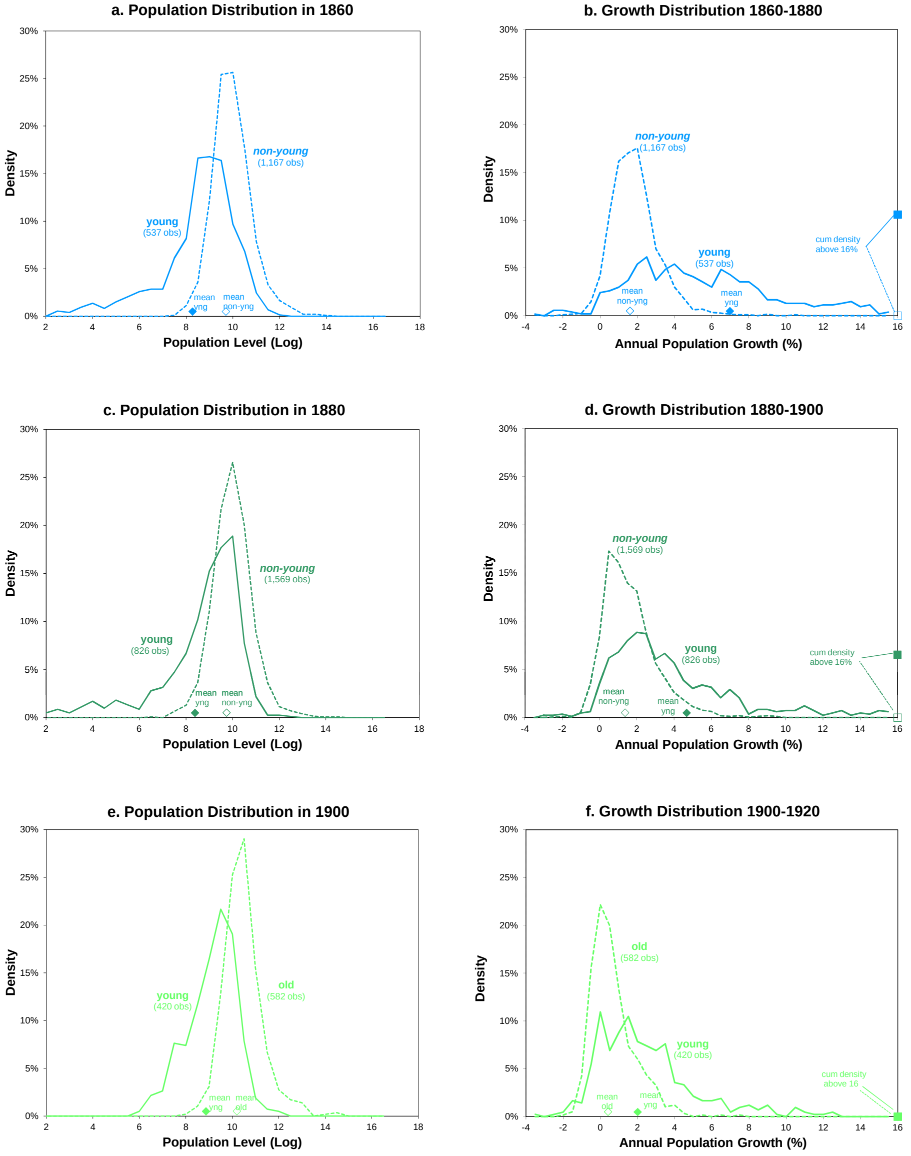


Figure C.2: *Empirical* Population Level and Growth Distributions by Age in 1860, 1880 and 1900

Figure shows the distribution of population across locations (Panel A) population growth across locations, in each case split 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 the main text. Note that for all years, the density of young locations by population is shifted to the left compared to the density of non-young/old locations by population (panels A, C, and E). For all three twenty-year periods, the density of young locations by growth rate is shifted to the right compared to the density of non-young/old locations by growth rate (panels B, D, and F).

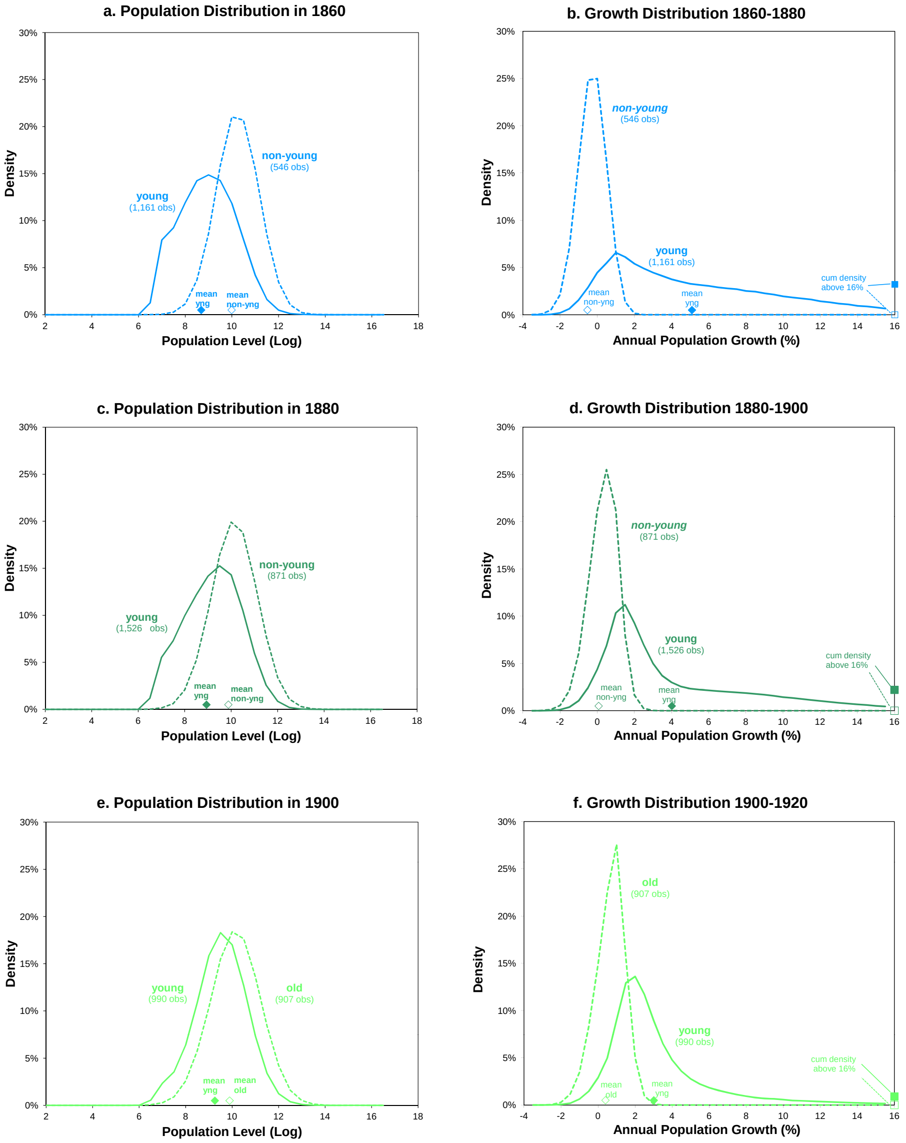


Figure C.3: *Simulated* Population Level and Growth Distributions by Age in 1860, 1880 and 1900

The simulated distributions of location population and population growth by age in 1860, 1880, and 1900 approximately match their empirical counterparts. The largest difference is that the simulated distributions are moderately more dispersed than are the empirical ones.

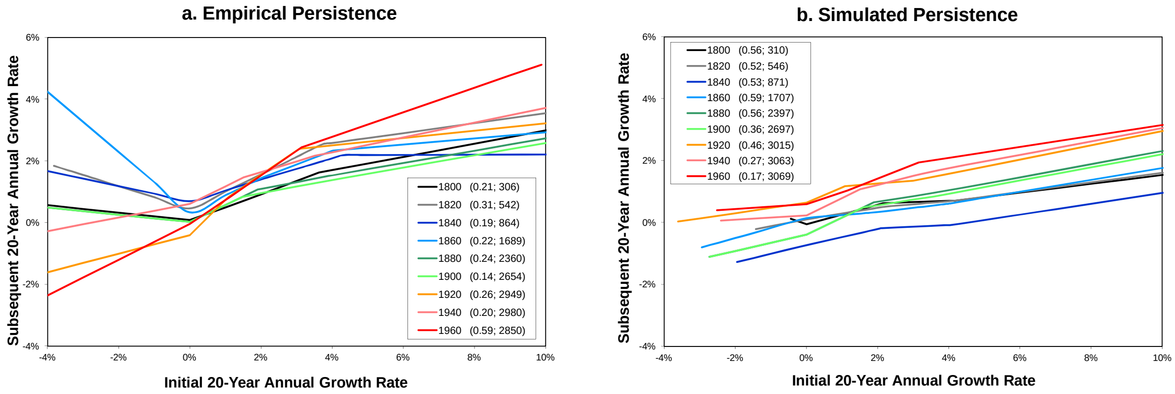


Figure C.4: Empirical and Simulated Growth Persistence

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 initial twenty-year period. Numbers in parenthesis are R-squared values.