

# Hinterlands, City Formation and Growth: Evidence from the U.S. Westward Expansion

DÁVID KRISZTIÁN NAGY

CREi, Universitat Pompeu Fabra and BSE

*First version received January 2018; Editorial decision October 2022; Accepted January 2023 (Eds.)*

I study how geography shaped city formation and aggregate development in the U.S. prior to the Civil War. To guide my analysis, I present a conjecture that cities' farm hinterlands fostered both city development and aggregate growth: the *hinterland hypothesis*. The hinterland hypothesis has rich implications on how various elements of U.S. geography—railroads, changes in U.S. political borders, and international trade—affected city formation and U.S. growth. To evaluate the hinterland hypothesis and its implications, I assemble a novel historical dataset on population, trading routes, and agricultural productivity at a high spatial resolution. I combine the data with a dynamic quantitative model of economic geography that yields the hinterland hypothesis as a prediction. I find evidence for the hinterland hypothesis by showing that the model can replicate the key patterns of U.S. urbanization and city formation. Finally, I conduct a series of counterfactuals in the model to quantify the effect of geography on cities and growth, guided by the implications of the hinterland hypothesis. Results indicate that railroads were responsible for 1.8% of urban population in 1860 and for 25% of real GDP growth between 1830 and 1860. The effect of international trade was similar in magnitude, while the effect of political border changes was small during the period.

**Key words:** Quantitative economic geography, Economic growth

**JEL codes:** O41, R13, R41

## 1. INTRODUCTION

In the very end of the 18th century, the U.S. was still a rural, pre-industrial country. The average income of a U.S. person was below average income in England and Wales (Lindert and Williamson, 2013). In 1790, the country had only five cities with more than 10,000 inhabitants, and its total urban population was around 100,000. By the middle of the 19th century, on the other hand, the U.S. was nearly the richest country on Earth (Weiss, 1992). It had almost 100 cities, and its total urban population exceeded 1,800,000 in 1860. What can explain this astonishing episode of take-off between the American Revolution and the Civil War? Scholars have offered the country's high-quality institutions as a prominent explanation (North, 2003). Another influential view, dating back to Turner (1894), is that the abundance of land on the country's Western frontier played a key role: "The existence of an area of free land, its continuous

recession, and the advance of American settlement westward, explain American development” (Turner, 1894).

How could the abundance of land in the West stimulate U.S. city formation and economic take-off? In his famous book on the rise of Chicago, William Cronon opines that “urban-rural commerce was the motor of frontier change [...] ‘the discovery, cultivation, and capitalization’ of land meant bringing it into the marketplace and attaching it to the metropolis” (Cronon, 1991, p. 54). “Western cities [...] grew in tandem with the countryside and played crucial roles in encouraging settlement from a very early time. City and country formed a single commercial system, a single process of rural settlement and metropolitan economic growth. To speak of one without the other made little sense” (Cronon, 1991, p. 47).

Motivated by Cronon’s observations, in this paper, I formulate a conjecture on how the abundance of land in the West fostered pre-Civil War U.S. city formation and growth: Cronon’s *hinterland hypothesis*. According to this hypothesis, the abundance of land attracted farmers to the West. Farmers brought their products to nearby trading places, helping these places develop into cities. The prosperity of these cities, in turn, contributed to aggregate U.S. growth.

Besides rationalizing the westward expansion of U.S. economic activity and the rapid development of cities, the hinterland hypothesis has rich implications on how different elements of U.S. geography shaped growth. The first such implication has to do with the effect of railroads. Railroads, which started to be built in the early 1830s, increased the sizes of cities’ potential farm hinterlands by decreasing the costs of inland trade. The hinterland hypothesis thus suggests that railroads fostered city development and aggregate growth. A second implication of the hinterland hypothesis is that the shift of U.S. political borders to the West, which increased the country’s land area by a factor of three, might have played a similar role as railroads. A third implication is that international trade, which favoured Eastern regions that had good access to trade with foreign markets, slowed down the westward expansion. At the same time, it might have fostered city development and growth in the East.

In this paper, I conduct the first quantitative evaluation of the hinterland hypothesis and its implications on how geography shaped pre-Civil War U.S. growth. To this end, I first assemble a novel historical dataset that encompasses county, city, and town population data from U.S. censuses between 1790 and 1860, data on the geographic location of waterways and railroads, changing U.S. political borders and agricultural productivity, all at a high spatial resolution. The data allow me to characterize the patterns of urbanization and city formation prior to the Civil War. I start by documenting the first-order feature of the data, consistent with the hinterland hypothesis: between 1790 and 1860, U.S. economic activity shifted substantially toward previously unoccupied land in the West. At the same time, the share of people living in cities increased by a factor of five. Also consistent with the hinterland hypothesis, I find that the vast majority of newly forming cities emerged at locations with good access to trade, such as near navigable rivers. Using the data, I document further facts about the location patterns and the growth of cities.

To rationalize and quantify the importance of hinterlands in pre-Civil War U.S. development and city formation, I combine my high-resolution historical data with a dynamic quantitative model of economic geography.<sup>1</sup> Studying the interaction between cities and their farm hinterlands calls for a model with two sectors: a farm and a non-farm sector. Locations endogenously specialize in farm or non-farm activities in the model. The farm sector uses land and labour to produce a homogeneous good under constant returns to scale, and sells the good to consumers

1. A rapidly growing literature uses quantitative models to study how geography shapes local and aggregate economic outcomes. I discuss this literature in my literature review below.

and to the non-farm sector. Firms in the non-farm sector combine the farm good with labour under economies of scale. They also hire workers to innovate, which leads to endogenous growth in their productivity. Due to scale economies and shipping costs, a large proportion of non-farm activities, and hence innovation, concentrate at a few locations with large population. In other words, cities form and develop endogenously.

I show that the model yields the hinterland hypothesis as a prediction by proposing a model-based measure of a location's *effective hinterland size* and proving that a location with larger effective hinterland size attracts more farmers to its surroundings, is more likely to become a city, and innovates more. This is due to the fact that a large farm input market implies higher returns to both labour use and innovation in the local non-farm sector. This interconnection between the farm and non-farm sectors helps locations in the Midwest, with access to a large hinterland especially after the placement of railroads, grow fast and attract a substantial fraction of population. As a result, the model is able to qualitatively replicate the first-order features of the data: the westward expansion and the increasing share of people living in cities. I test the model's fit to these and various other untargeted moments, and find that the model fits the data not only in a qualitative but also in a quantitative sense. For instance, the correlation between locations' population levels in the model and in the data is around 0.6 throughout most of the period of investigation. This is despite the fact that the model's calibration does not target any moment of the evolution of the spatial distribution of population.

To further highlight the importance of hinterlands in explaining the first-order patterns in the data, I contrast the fit of my model with the model of [Desmet et al. \(2018\)](#), henceforth DNR. DNR is also a dynamic quantitative geography model in which local innovation activity induces aggregate growth and an endogenous evolution of the spatial distribution of population. Crucially however, DNR features one sector. Thus, it does not allow farm hinterlands to foster growth in the non-farm sector, hence city growth and aggregate development. I take the DNR model to the data and use it to predict how the spatial distribution of population evolved in the pre-Civil War U.S. I find that the DNR model is unable to replicate the substantial westward shift in U.S. population and the emergence of cities outside the Northeast. These results can thus be viewed as further evidence for the quantitative relevance of the hinterland hypothesis.

Armed with quantitative evidence that the model is able to rationalize the data, I conduct a series of counterfactuals to measure how geography shaped U.S. growth, guided by the implications of the hinterland hypothesis. In the first counterfactual, I measure the effect of railroads on city formation and growth. I find that the absence of railroads would have led to a 1.8% lower urban population in 1860, as well as a 25% drop in U.S. real GDP growth between 1830 and 1860. Railroads thus played an important role in aggregate U.S. growth and city development, in line with the implication of the hinterland hypothesis that railroads fostered city growth through increasing the sizes of farm hinterlands. At the same time, I find that the effect of political border changes was positive but small. This is consistent with the fact that most of newly occupied land was hard to access prior to railroad construction in the far West, which only took place after the Civil War. Finally, in line with the third implication of the hinterland hypothesis, I find that international trade slowed down the westward expansion but fostered city development in the Northeast (a region that enjoyed a better geographic position to trade with foreign markets), leading to faster growth in the aggregate.

I test the robustness of my quantitative findings to changes in the values of model parameters and modelling assumptions. One particularly important concern is that the placement of railroads is endogenous. I carefully design my calibration strategy to alleviate this concern when it comes to the identification of parameters and the results of the counterfactual with no railroads. Nevertheless, the endogeneity of railroads is a concern for the results of the other two counterfactuals, in which alternative railroad networks could emerge. To address the concern, I develop

an extension of the model in which railroads are endogenously placed at locations with high demand for investment in transport infrastructure. I find that railroad locations predicted by my endogenous railroad placement framework feature a good fit to the actual location of railroads. I also find that the quantitative results of the counterfactuals are not substantially affected by the endogeneity of railroad placement.

This paper is related to four strands of the literature. One contribution of the paper is that it investigates urbanization and the interaction between city development and aggregate growth in a spatially heterogeneous environment, which makes it suitable for quantitative analysis. Previous studies of city formation and urban growth either assume a set of identical sites, as [Helsley and Strange \(1994\)](#), [Henderson and Venables \(2009\)](#), or [Rossi-Hansberg and Wright \(2007\)](#), or two types of locations, as [Black and Henderson \(1999\)](#) or [Bertinelli and Black \(2004\)](#). Another related strand of the literature, including [Ngai and Pissarides \(2007\)](#), [Baumol \(1967\)](#), [Buera and Kaboski \(2012a, 2012b\)](#), and [Caselli and Coleman \(2001\)](#), focuses on structural change, which has been recognized as a key source of urbanization. The models in these papers are also highly stylized in their spatial structure or are completely aspatial. On the other hand, the flexible geographic structure in my model allows me to quantitatively evaluate the hinterland hypothesis and its implications on how different elements of geography affected city development and aggregate growth.<sup>2</sup>

The fact that my paper models the economy over a rich geography relates it to the rapidly growing literature applying quantitative models to examine the spatial distribution of economic activity, such as [Allen and Arkolakis \(2014\)](#), [Caliendo \*et al.\* \(2018\)](#), [Fajgelbaum and Redding \(2022\)](#), [Monte \*et al.\* \(2018\)](#), and [Redding \(2016\)](#). Similar to my model, these models can accommodate a large number of locations that are heterogeneous in their factor endowments, productivity, and shipping costs. However, given that these models are static, they cannot be used to measure how geography shapes growth. My main contribution to this literature is developing a dynamic quantitative framework that allows me to study not only the spatial distribution of economic activity but also its evolution. Besides opening the door to measuring the effect of geography on growth, incorporating the dynamics also allows me to test the model on how well it can predict the evolution of the spatial distribution of population.

The paper is also related to the set of papers that model the evolution of productivity and growth across space, such as [Desmet and Rossi-Hansberg \(2014\)](#), [Desmet \*et al.\* \(2018\)](#), [Eckert and Peters \(2018\)](#), and [Michaels \*et al.\* \(2012\)](#).<sup>3</sup> Relative to this literature, I emphasize the key role of cities in determining the relationship between geography and growth. Geography, and the size of hinterlands in particular, shapes the formation and growth of cities through affecting the non-farm sector's incentives to innovate. The spatial distribution of cities, in turn, drives the spatial distribution of innovation, hence aggregate growth in the economy. Within this literature, my work is closest to [Desmet \*et al.\* \(2018\)](#) ("DNR"). DNR is also a spatial growth model in which aggregate growth is induced by firms' incentives to innovate locally. Crucially however, DNR features a single sector. In this paper, I show that a two-sector framework in which access

2. See [Herrendorf \*et al.\* \(2014\)](#) for a review of the literature on structural change. It is important to note that although my paper introduces a geographic dimension into a model with structural change, tractability requires it to abstract from certain forces that are typically present in aspatial models of structural change. Most importantly, the model features homothetic preferences and a unitary elasticity of substitution between the farm and non-farm sectors in consumption, even though non-homotheticity and consumption-side complementarity have been pointed out as potentially important sources of structural change in the literature ([Herrendorf \*et al.\*, 2014](#)).

3. [Caliendo \*et al.\* \(2019\)](#) also develop a dynamic quantitative model of economic geography. In their model, the dynamics are driven by households making dynamic location choices subject to mobility frictions. Unlike in [Caliendo \*et al.\* \(2019\)](#), my model features frictionless labour mobility but an endogenous dynamic evolution of productivity.

to a farm hinterland drives innovation in the non-farm sector is essential to replicate the evolution of the spatial distribution of population and the patterns of city formation seen in the data.

Finally, my paper is related to the literature quantifying the economic impact of U.S. railroads. In his seminal work, [Fogel \(1964\)](#) aims at carefully accounting for the various local and aggregate effects of 19th-century U.S. railroads. [Donaldson and Hornbeck \(2016\)](#) revisit Fogel's analysis, using a static quantitative trade model. They focus on the impact of railroads on real output through railroads allowing for cheaper transportation of agricultural products.<sup>4</sup> Relative to [Donaldson and Hornbeck \(2016\)](#), I broaden the scope of the investigation by also accounting for the effect of railroads on the formation and development of cities. Using a model-based decomposition, I show that this new effect is responsible for at least 34% of the total effect of railroads on output.<sup>5</sup> The result that railroads foster economic growth through city formation is consistent with empirical estimates by [Atack et al. \(2010\)](#), who document increasing urbanization in counties along railroad lines in the pre-Civil War Midwest.

The structure of this paper is as follows. Section 2 uses pre-Civil War U.S. data to document the key patterns of urbanization and city formation during the period. Section 3 outlines the theoretical model, while Section 4 describes the steps of taking the model to the data. Section 5 evaluates the model's fit to the data and compares it to the fit of the DNR model. It also presents the results of the three counterfactuals aimed at quantifying the role of geography in U.S. growth and city formation. Section 6 tests the robustness of the results. Section 7 concludes.

## 2. STYLIZED FACTS: CITY FORMATION IN THE PRE-CIVIL WAR U.S.

This section presents a series of stylized facts that characterize the process of urbanization and city formation in the pre-Civil War U.S. Consistent with the hinterland hypothesis, the U.S. saw a substantial shift in its population and economic activity toward the West (fact 1). At the same time, the share of people living in cities increased by a factor of five, and most of this urbanization can be attributed to new cities forming as opposed to pre-existing cities growing (fact 2). The location of these newly forming cities seems to be driven by better trading opportunities, such as access to navigable waterways or railroads (fact 3). Furthermore, there were substantial geographic differences in the patterns of city formation: new cities appeared relatively far from each other, but less so in the Northeast (fact 4). Finally, the growth of cities was approximately orthogonal to city size (Gibrat's Law) but with slight convergence; at the same time, small towns grew slower than cities above 10,000 inhabitants (fact 5).

I use three datasets to document these facts. First, I use census data on county, city, and town populations that are available for every decade starting from 1790. The census reports the population of any settlement above 2500 inhabitants. I classify the settlements above 10,000 as *cities*, and those between 2500 and 10,000 as *towns*.<sup>6</sup> Second, I use data on the location of land, rivers, canals, lakes, and oceans at a high spatial resolution, which come from the ESRI Map of U.S. Major Waters. Third, railroad maps coming from the website *oldrailhistory.com* show

4. Other recent papers studying the economic impact of transport infrastructure in quantitative spatial frameworks include [Fajgelbaum and Schaal \(2020\)](#), [Allen and Arkolakis \(2021\)](#), [Herrendorf et al. \(2012\)](#), [Santamaría \(2022\)](#), [Swisher \(2014\)](#), and [Trew \(2020\)](#). Also see the review of this literature in [Redding and Turner \(2015\)](#).

5. This finding is in line with [David \(1969\)](#), who argues that scale economies might lead to railroads having indirect effects not captured in Fogel's analysis. In his response, [Fogel \(1979\)](#) points to the fact that scale economies are weak in the farm sector, but does not consider scale economies in non-farm activities.

6. This distinction is motivated by the finding that settlements above 10,000 inhabitants exhibited significantly different growth patterns from those below 10,000 (fact 5), in line with the findings of [Desmet and Rappaport \(2017\)](#).

the rail network in 1835, 1840, 1845, 1850, and 1860. [Supplementary Material, Appendix B](#) provides a detailed description of these datasets.

My period of investigation starts with the first population census (1790) and ends right before the U.S. Civil War (1860). The choice of 1860 as the end of the period is motivated by the well-known fact that the Civil War had a large and long-lasting economic impact on the U.S. economy. Besides one million people dead, the war caused dramatic losses in both transport infrastructure and city productivity, especially in the South ([Foote, 1974](#)). It also ended the institution of slavery, affecting agricultural production in the South. These events took the U.S. economy on a different development path, characterized by divergence between the North and the South, which lasted for a century ([Kim and Margo, 2004](#)). Since modelling the Civil War is outside the scope of my theoretical framework, I focus on the years 1790–1860 throughout the paper.

### *Fact 1: Westward expansion*

U.S. population increased by a factor of eight between 1790 and 1860. Total U.S. population was growing primarily because birth rates were above death rates; immigration did not play a major role at the time, as shown by the fact that only 13% of U.S. population was foreign-born in 1860. Also, the U.S. saw significant changes in its political borders during the period, mainly as a result of the Louisiana Purchase (1803), the annexation of Texas (1845), and the Mexican–American War (1847).<sup>7</sup> Population growth was, however, far from uniform across space. Population in the West increased much more rapidly, as evident from [Figure 1](#) which shows population density and the mean centre of U.S. population in 1790, 1820, 1840, and 1860. By the end of the period, the mean centre of population shifted about 400 miles to the West from Baltimore, Maryland to Columbus, Ohio.

This process led to rapid convergence in population sizes among three large regions of the U.S.: the Northeast, the South, and the Midwest.<sup>8</sup> Whereas the Midwest only had 1.3% of U.S. population in 1790, its share was almost equal to that of the Northeast and the South in 1860 ([Figure 15 in Supplementary Material, Appendix E](#)).<sup>9</sup>

### *Fact 2: rapid urbanization*

While U.S. population grew and moved to the West, the fraction of people living in cities above 10,000 inhabitants rose from 2.8% to 14.8% (blue line in the left panel of [Figure 2](#)). Part of this increase can be the result of population growth: as total population grows, the population of each settlement is likely to rise as well, hence more and more settlements reach the 10,000-inhabitant threshold. To show that my results are not driven by this mechanical effect, I set an alternative threshold that equals 10,000 in 1790, but then grows in proportion to total population. The red line in the left panel of [Figure 2](#) shows that the increase in urbanization was still three-fold if I use this alternative threshold.

7. Although modelling changes in political borders and total population is outside the scope of this paper, the model presented in [Section 3](#) allows for exogenous changes in both. In [Section 5.3](#), I conduct a counterfactual to study the effect of political border changes on cities and aggregate growth.

8. I follow [Caselli and Coleman \(2001\)](#) when defining the geographic boundaries of these regions. See [Supplementary Material, Appendix B](#) for the list of states belonging to each region.

9. Although political border changes increased the area of the Midwest and the South during the period, the increase in these regions' population was more than proportional to the increase in their area. Population density in the South tripled between 1790 and 1860, while it increased by more than a factor of 100 in the Midwest.



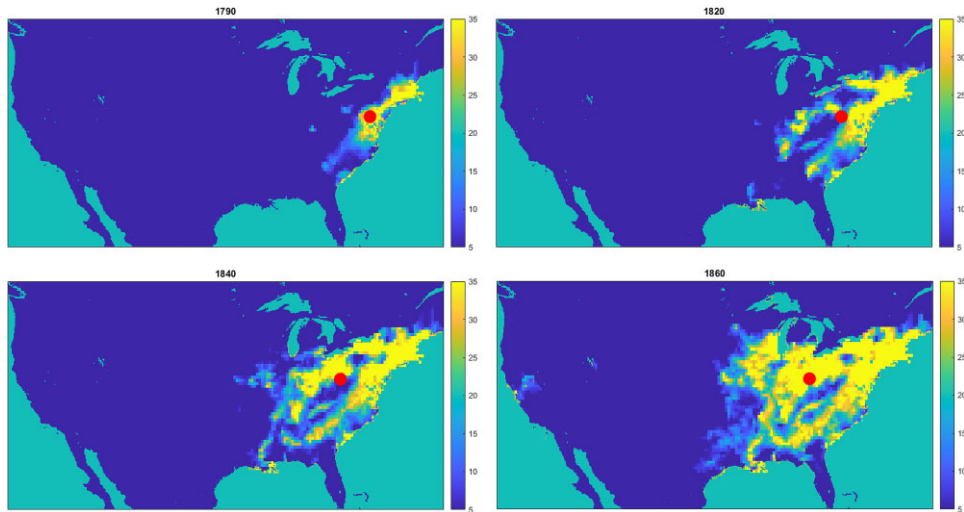


FIGURE 1

## Evolution of the U.S. population distribution

*Notes:* The four panels of the figure present population (number of inhabitants) per square mile in each 20 by 20 arc minute grid cell of the U.S. in 1790, 1820, 1840, and 1860, respectively. The red dot represents the mean centre of U.S. population. The scale is bottom-coded at five and top-coded at thirty-five for visibility. Source: U.S. census.

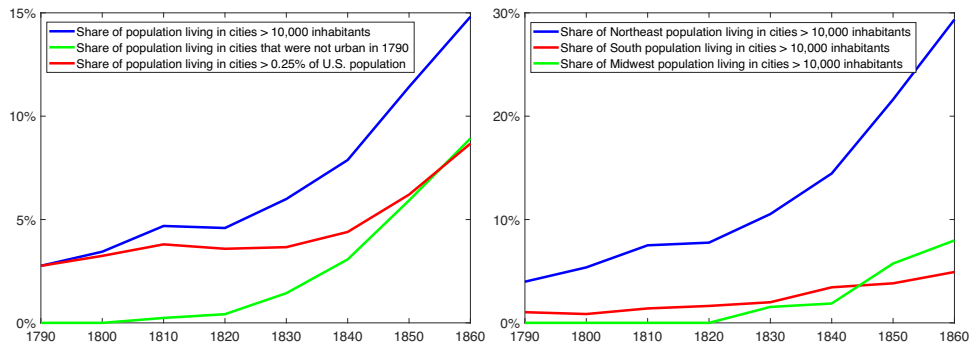


FIGURE 2

## Urbanization in the pre-Civil War U.S.

*Notes:* In the left graph, the blue line shows the fraction of U.S. population living in cities (settlements above 10,000 inhabitants) for each census year. The green line shows the fraction of U.S. population living in cities that were not classified as urban places by the census in 1790. The red line shows the fraction of U.S. population living in settlements whose population exceeds 0.25% of total U.S. population in the given year. In the right graph, the blue, red, and green lines show the fraction of population living in cities (settlements above 10,000 inhabitants) in the Northeast, South, and Midwest, respectively. Source: U.S. census.

The fact that the number of cities above 10,000 inhabitants rose from 5 (New York City, Philadelphia, Boston, Charleston, and Baltimore) to ninety-nine seems to indicate that most of the urbanization came from new cities forming, as opposed to pre-existing cities growing. Figure 2 provides additional evidence for this. Taking out cities that were already classified as “urban places” in 1790, the share of cities in total population increased to 8.9% by 1860 (green line in the left panel of Figure 2). This implies that 74% of urbanization was due to cities that did not yet exist in 1790.

Finally, the right panel of Figure 2 shows that the rapid urbanization was far from uniform across space. The share of people living in cities increased to almost 30% in the Northeast, while

TABLE 1  
*Fraction of cities forming at trading routes*

	Fraction of cells	Fraction of land	Fraction of cities
Water	39.1%	35.6%	98.6%
Confluence	23.0%	19.9%	85.5%
Railroad	24.3%	23.2%	95.7%

*Notes:* The first column of this table shows the number of 20 by 20 arc minute grid cells located at water, a confluence, or a railroad relative to the total number of cells in the U.S. The second column shows the fraction of U.S. land that belongs to cells located at water, a confluence, or a railroad. The last column shows the fraction of 1860 U.S. cities that formed in cells located at water, at a confluence, or at a railroad. A cell is located at water if the cell, or the cell next to it, has a navigable river, canal, lake, or the sea in 1860. It is located at a confluence if the cell, or a cell next to it, is surrounded by at least three cells with water in 1860. It is located at a railroad if the cell, or a cell next to it, was part of the railroad network in 1860. Source: U.S. census, ESRI Map of U.S. Major Waters, and *oldrailhistory.com*.

it remained relatively modest in the South. On the other hand, the end of the period saw rapid urbanization in the Midwest, even though the region had no cities at all until 1820.

### *Fact 3: cities formed at trading routes*

Where did new cities form? Figure 16 (Supplementary Material, Appendix E) shows the locations of the ninety-four cities coming into existence between 1790 and 1860. A striking regularity is that almost all new cities are close to places with good trading opportunities such as rivers (especially the Mississippi, the Ohio River, and the Hudson), the Great Lakes, or the sea. To confirm this, I discretize the territory of the U.S. into 20 by 20 arc minute grid cells<sup>10</sup> and classify each cell depending on whether it is located near a waterway (*i.e.* next to a cell that includes a navigable river or canal, a lake, or the sea), near a confluence of waterways, or near a railroad in 1860. As railroads and canals might be placed near newly forming cities, the emergence of cities at railroads and canals should not be interpreted as a causal relationship. Nevertheless, it indicates a correlation between city locations and a good access to trade.

Table 1 shows that a disproportionately high share of cities appear in cells with better access to trade. The fraction of cities forming in cells near water is more than 98%, an order of magnitude larger than the fraction of cells near water, or the fraction of land that belongs to these cells. Similarly, cities are more likely to appear near confluences than at other places. Clearly, locating at a confluence increases trading opportunities even more than locating only at water. Finally, cities are more likely to appear near railroads.

The formation of most cities at trading routes does not necessarily imply that trade attracted people to these locations. In particular, locations with good trading opportunities might have better natural amenities or higher agricultural productivity, and such advantages, not trade itself, may have attracted people to these locations. To control for natural amenities and agricultural productivity, I collect data from the FAO GAEZ database<sup>11</sup> and consider specifications of the form

$$newcity_i = \beta_0 + \beta_1 tradeopp_i + \beta_2 amen_i + \beta_3 prod_i + \epsilon_i, \quad (2.1)$$

10. This means that each cell is approximately 20 by 20 miles large.

11. The natural amenity variables are based on measures of temperature and precipitation, while the productivity variables are based on 1961–90 yields of the major crops grown in 19th-century U.S. Supplementary Material, Appendix B describes each of these variables in detail. Section 4.4 further discusses the assumption that 1961–90 yield data are used to proxy historical productivity levels.



TABLE 2  
*The role of trading routes, controlling for amenities and productivity*

	Dependent variable: newcity					
	(1)	(2)	(3)	(4)	(5)	(6)
Water	0.023** (0.003)			0.018** (0.002)		
Confluence		0.032** (0.004)			0.029** (0.004)	
Railroad			0.035** (0.004)			0.030** (0.004)
Productivity and amenities	No	No	No	Yes	Yes	Yes
Number of observations	7641	7641	7641	7641	7641	7641

*Notes:* This table presents the results of estimating equation (2.1). The unit of observation is a 20 by 20 arc minute grid cell in the 1860 U.S. The dependent variable, *newcity*, is a dummy variable equal to 1 if the cell contains a city formed after 1790 in 1860. The three independent variables, *water*, *confluence*, and *railroad*, are dummy variables equal to 1 if the cell is located at water, at a confluence, or at a railroad, respectively. “Prod and amenities” refer to productivity and amenity controls. A cell is located at water if the cell, or the cell next to it, has a navigable river, canal, lake, or the sea in 1860. It is located at a confluence if the cell, or a cell next to it, is surrounded by at least three cells with water in 1860. It is located at a railroad if the cell, or a cell next to it, was part of the railroad network in 1860. Heteroskedasticity-robust standard errors in parentheses. \*\*Significant at 1%. Source: U.S. census, ESRI Map of U.S. Major Waters, *oldrailhistory.com*, and FAO GAEZ.

where  $tradeopp_i$  is one if cell  $i$  was located at a specific trading opportunity which is water, confluence, or rail depending on the specification, and zero otherwise.  $newcity_i$  is a variable that equals one if a new city appeared in cell  $i$  between 1790 and 1860, and zero otherwise.  $amen_i$  and  $prod_i$  are amenity and productivity controls. I estimate specification (2.1) as a linear probability model.<sup>12</sup>

The estimation results suggest that even after controlling for amenities and productivity, cities are significantly more likely to appear near trading routes. Table 2 shows the estimated  $\beta_1$  coefficients for all three types of trading opportunities, both with and without controls. Locations at a waterway, a confluence, or a railroad receive a city with a 2–4% higher probability than other locations. The inclusion of amenity and productivity controls hardly affects the significance or the magnitude of these estimates.

Finally, it could have also happened that migration to the West itself led to the emergence of cities near trading routes. Occupying the West was a slow transition process, and this transition might have played out in a way that more settlers happened to come together at trading routes. For instance, river confluences could be hubs not only for trade but also for settlers, which could have led to a larger concentration of population at these locations. To address this concern, I model a process of slow population migration from the East in [Supplementary Material, Appendix C](#). In this process, the frequency at which people move to another location is negatively related to the cost of moving to that location. I consider two calibrations for moving costs: one in which the cost is a function of geographic distance, and another in which the cost is lower along water routes than inland. Then I simulate the process, calculate the fraction of population clusters forming at water, confluences, and railroads, and run equation (2.1) on the simulated data. In both calibrations, the process of slow migration is unable to replicate the disproportionate emergence of cities near trading routes. These results together motivate the choice

12. Probit and logit estimates are very similar. These results are available from the author upon request.

TABLE 3  
*Clustering of cities*

Average distance from cities in the region (miles)	
– Northeast	315.8
– Rest of the U.S.	1057.4
Average distance from the closest city in the region (miles)	
– Northeast	50.2
– Rest of the U.S.	142.7

*Notes:* The first two numbers in this table correspond to the unweighted average of pairwise distances between cities forming by 1860 in the Northeast and in the rest of the U.S., respectively. The last two numbers correspond to the unweighted average of a city's distance from the closest city forming by 1860 in the Northeast and in the rest of the U.S., respectively. Source: U.S. census.

of a theoretical framework in which incentives to trade, and not other forces, attract people to locations with good trading opportunities.

*Fact 4: cities clustered more in the Northeast*

Besides cities' locations relative to trading routes, cities' locations relative to each other might also be of interest. A quick glance at Figure 16 (Supplementary Material, Appendix E) suggests that cities tended to form relatively far from other cities, but this was less true in the Northeast. To check if this was indeed the case, I calculate two statistics measuring the clustering of cities in these two regions: cities' average distance from other cities in the region and cities' average distance from their closest neighbour.

Both statistics confirm the prediction that city location patterns exhibited more clustering in the Northeast than outside the Northeast, as can be seen from Table 3. Needless to say, these results could be driven by the fact that the Northeast was both denser and substantially smaller in its land area. To address this concern, I can split the rest of the country into smaller units and re-compute measures of clustering for these units. With the classification of locations into regions used earlier in this section (Northeast, South, and Midwest), for instance, I find that cities' average distance from their closest neighbour was 166.4 miles in the South and 136.7 in the Midwest. Both are substantially larger than the corresponding number in the Northeast (50.2), and very close to the number obtained for the entire area outside the Northeast (142.7). While each of the South and the Midwest is still somewhat larger than the Northeast, the fact that the results barely change relative to Table 3 suggests that the higher measured clustering in the Northeast is not a mechanical consequence of this region being relatively small.<sup>13</sup>

*Fact 5: city growth*

The final set of facts that I document is related to the growth of cities (that is, settlements above 10,000 inhabitants) and towns (that is, settlements between 2500 and 10,000 inhabitants). Figure 3 plots a binscatter of the log decennial growth rate of these settlements against their

13. One might also wonder whether the differences in location patterns are caused by history, not by economic forces. In particular, land in the West was allocated for different types of land use according to a rectangular system called the Public Land Survey System (Cazier, 1976). Schools, roads, and railroads were designed to be built at pre-selected locations within each rectangular survey unit, which led to an equidistant placement of these pieces of infrastructure. However, given that each rectangular unit was 6 by 6 miles large, this system was only likely to influence the location of cities *within* 20 by 20 mile cells, not *across* these cells.

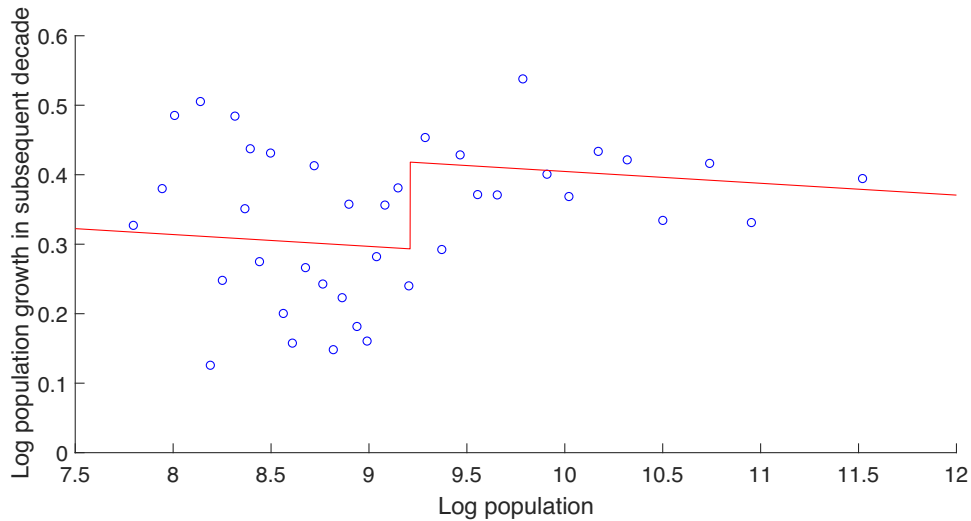


FIGURE 3  
Population growth versus size

*Notes:* This figure presents a binscatter of log decennial growth against log population among settlements above 2500 inhabitants in the U.S. in each of the census years (1790, 1800, 1810, 1820, 1830, 1840, or 1850). The horizontal axis corresponds to the log of the settlement's population in the given year. The vertical axis corresponds to the log of the settlement's population growth over the subsequent decade. Each circle represents average log population and average log growth within one of forty equally sized population bins. The red line is a regression line corresponding to a regression of log growth on log population and a dummy equal to one if population exceeds 10,000 inhabitants. Source: U.S. census.

log size in the beginning of the decade.<sup>14</sup> The red line is a regression line that corresponds to a regression of log growth on log size, controlling for a dummy variable that equals one if the settlement is a city:

$$\ln(\text{growth}_{st}) = \gamma_0 + \gamma_1 \ln(\text{size}_{st}) + \gamma_2 \text{city}_{st} + \eta_{st}.$$

Thus, the dip of the red line in the middle of the plot corresponds to the size threshold of 10,000 inhabitants (9.21 in logs) between cities and towns.

The difference between the average growth rate of cities and towns is substantial: it is about 1 percentage point per year. To confirm this discontinuity in growth rates between cities and towns, column (1) of Table 14 (Supplementary Material, Appendix E) presents the results of the above regression specification. The estimated coefficient on  $\text{city}_{st}$  is indeed significantly different from zero. Columns (2)–(4) consider thresholds different from 10,000, and find no significant break in the growth rate at these alternative thresholds. The finding that cities grow faster than towns is in line with the different growth patterns of small versus large counties during the same period documented by Desmet and Rappaport (2017). More generally, the fact that cities can act as focal points of growth has been pointed out by the urban growth literature, such as Black and Henderson (1999).

14. Cities and towns can grow through the annexation of surrounding areas. For instance, the Northern Liberties District became part of Philadelphia in 1854, increasing the population of the city by as much as 39% (according to 1850 census numbers). To abstract from this source of population growth, I consider cities and towns in the same 20 by 20 arc minute grid cell as a single city or town, depending on whether its total population is above or below 10,000. What even this strategy cannot account for is the annexation of surrounding villages below 2500 inhabitants.

Both cities and towns exhibit convergence in their size as larger settlements within each of the two groups grow on average slower than smaller ones, as can be seen in Figure 3. However, convergence is slow. The average growth rate of large cities does not differ strikingly from the average growth rate of small cities, and the same is true if we compare the growth rates of large and small towns. That is, the patterns of city and town growth are quite close to Gibrat's Law. Gibrat's Law (the independence of growth and size) is a well-known fact of city growth that has been documented for various countries and time periods (Eaton and Eckstein, 1997; Ioannides and Overman, 2003; Eeckhout, 2004). However, in the context of 19th-century U.S., Desmet and Rappaport (2017) show that population growth deviates from Gibrat's Law and exhibits convergence among small counties, in line with my findings.

### 3. A DYNAMIC QUANTITATIVE MODEL OF CITY FORMATION

Motivated by the facts shown in Section 2, this section presents the theoretical framework I use to study the forces shaping pre-Civil War U.S. city formation and aggregate growth. The economy is populated by a mass of consumers who choose in which sector to work and where to live. Locations differ in their endowment of land, their productivity, and their shipping costs to other locations. Sectors differ in their returns to scale and in their intensity of land use. This implies a higher concentration of non-farm production than that of farm production in equilibrium, and thus the existence of cities. Finally, non-farm productivity grows due to firms' innovation activity, and diffuses across space. This leads to the dynamic evolution of sectoral specialization patterns, and thus to the endogenous evolution of the geographic locations and sizes of cities.

#### 3.1. *Single-country setup*

**3.1.1. Geography, goods, and factor endowments.** A country—the U.S.—occupies a compact subset  $S$  of a two-dimensional surface. Individual locations (elements of  $S$ ) are indexed by  $r$ ,  $s$ , or  $u$ . There are two sectors in the economy: a sector producing a homogeneous *farm good* and a sector producing a continuum of *non-farm goods* indexed by  $i \in [0, n_t]$ . The mass of non-farm goods,  $n_t$ , is determined endogenously in equilibrium. Classifying economic activities into these two broad sectors follows the literature examining the pre-Civil War period (Towne and Rasmussen, 1960; Weiss, 1992). Note that the processing and the distribution of agricultural goods are non-farm activities under this classification.

Factor endowments are as follows. In period  $t \in \{1, 2, \dots\}$ , an exogenously given mass of  $\bar{L}_t$  consumers live in the U.S.<sup>15</sup> Each consumer owns one unit of labour which she supplies inelastically. Finally, the supply of land at location  $r$  is denoted by  $H_t(r)$ .  $H_t(\cdot)$  is an exogenous measurable function.

Besides choosing where to live and what quantity of farm and non-farm goods to consume, consumers choose in which sector they want to work in each time period.<sup>16</sup> Section “Farmers” describes the choices and constraints of those choosing to work in the farm sector (“farmers”), while section “Non-farm workers” describes the choices and constraints of those choosing to work in the non-farm sector.

15. Note that I allow U.S. population to change over time, which allows me to incorporate birth, death, and immigration. Modelling consumers' endogenous fertility and immigration decisions is outside the scope of this paper. See Desmet *et al.* (2018) for a spatial growth model with endogenous immigration.

16. Although it is possible to introduce frictions to consumers' mobility between sectors, I assume free mobility. This is motivated by the finding that there was no gap in real hourly income between the agricultural and non-agricultural sectors in the pre-Civil War U.S. (David, 2005).

**3.1.2. Farmers.** Farmers pick a *residential location*,<sup>17</sup> where they live and work, as well as a *trading place* where they sell their product and buy non-farm goods. For simplicity, I assume that consumption also happens at the trading place.<sup>18</sup>

Farmers order non-farm varieties according to constant elasticity of substitution (CES) preferences (Dixit and Stiglitz, 1977), while they have Cobb–Douglas preferences over the index of non-farm varieties and the farm good. Therefore, a farmer with residence at  $r$  and trading place at  $s$  obtains the following utility at time  $t$ :

$$U_t^F(r, s) = \zeta \left[ \int_0^{n_t} x_t^N(r, s, i)^{\frac{\epsilon-1}{\epsilon}} di \right]^{\frac{\epsilon}{\epsilon-1}} x_t^F(r, s)^{1-\nu},$$

where  $x_t^N(r, s, i)$  is the quantity of non-farm variety  $i$ , and  $x_t^F(r, s)$  is the quantity of the farm good consumed by the farmer.  $\zeta = \nu^{-\nu} (1 - \nu)^{-(1-\nu)}$  is a constant that simplifies the subsequent formulas algebraically.

Farmers choose their production and consumption levels, their residence, and their trading place to maximize utility, taking all prices as given. They face three constraints: the production function in the farm sector, the shipping technology, and a budget constraint.

The production function that a farmer with residence at  $r$  faces is Cobb–Douglas in labour and land:

$$q_t^F(r) = B_t(r) \ell_t^F(r)^\alpha h_t(r)^{1-\alpha},$$

where  $q_t^F(r)$  is the quantity of the good produced,  $B_t(r)$  is location-specific farm TFP,  $\ell_t^F(r)$  is the quantity of labour used (which equals one due to the inelastic supply of labour), and  $h_t(r)$  is the quantity of land used.

The farmer is subject to iceberg shipping costs when shipping her product to the trading place. If  $q_t^F(r)$  units are shipped from  $r$  to  $s$ , the quantity that *actually arrives* at  $s$  is

$$\tilde{q}_t^F(r, s) = \frac{q_t^F(r)}{\tau_t^F(r, s)},$$

where  $\tau_t^F(r, s) \geq 1$  is the shipping cost between  $r$  and  $s$ .  $\tau_t^F(\cdot, \cdot)$  is assumed to be an exogenous measurable function that satisfies the triangle inequality, that is,

$$\tau_t^F(r, u) \leq \tau_t^F(r, s) \tau_t^F(s, u)$$

for any  $r, s$  and  $u \in S$ . I normalize  $\tau_t^F(s, s)$  to one. Note that the model allows for changes in shipping costs over time. As we will see, shipping costs lead to the concentration of farm production around trading places in equilibrium. This agglomeration force is counterbalanced by the dispersion force that farmers need to use land, which is available in fixed supply at each location.

Finally, the farmer faces the budget constraint

$$\int_0^{n_t} p_t^N(s, i) x_t^N(r, s, i) di + p_t^F(s) x_t^F(r, s) \leq p_t^F(s) \tilde{q}_t^F(r, s) - R_t(r) h_t(r) + y_t(r),$$

17. I abstract from slavery. I discuss this modelling choice in the last paragraph of this section.

18. This assumption allows me to abstract from shipping costs incurred between the residential location and the trading place. [Supplementary Material, Appendix A.5](#) presents a version of the model in which consumption happens at the residential location and argues that the difference between the two models is small under the values of shipping costs chosen in the calibration.

where  $p_i^N(s, i)$  is the price of non-farm variety  $i$ ,  $p_i^F(s)$  is the price of the farm good at the trading place, and  $R_i(r)$  is the land rent that the farmer pays per unit of the land she uses at  $r$ .

I assume that land rents paid by farmers at a location are fully redistributed to them with equal shares. I denote this redistributed amount of local rents per farmer by  $y_i(r)$ , and include it on the right-hand side of the farmer's budget constraint. Note that I make two assumptions here: the first is that land rents are redistributed toward farmers rather than non-farm workers or firms, while the second is that they are redistributed locally rather than across the entire country (or within larger regions). The assumption of redistribution toward farmers rather than the non-farm sector is supported by historical estimates that land rents paid to non-farm land owners were tiny prior to the Civil War, never exceeding 2.5% of farm GDP (Towne and Rasmussen, 1960). The assumption of local redistribution follows quantitative economic geography models such as Redding and Sturm (2008) and Desmet *et al.* (2018), but is also supported by historical evidence. In particular, census data show that 74% of farms were owned by the farmers who operated them in the first census year in which such data are available (1880). This motivates using a framework in which farm income gets redistributed locally rather than globally.<sup>19</sup>

Land market clearing implies that land income per farmer equals land rents paid per farmer:

$$y_i(r) = R_i(r)h_i(r).$$

As a result, the farmer's nominal income that she can spend on consumption goods equals her production revenues,  $p_i^F(s)\tilde{q}_i^F(r, s)$ . Since the farmer's utility is Cobb–Douglas in non-farm and farm goods, the indirect utility of a farmer living at  $r$  and trading at  $s$  equals her nominal income divided by the Cobb–Douglas aggregate of non-farm and farm prices:

$$U_i^F(r, s) = \frac{p_i^F(s)\tilde{q}_i^F(r, s)}{P_i(s)^v p_i^F(s)^{1-v}} = \frac{p_i^F(s)\tau_i^F(r, s)^{-1} B_i(r) [H_i(r)/L_i^F(r)]^{1-\alpha}}{P_i(s)^v p_i^F(s)^{1-v}}, \quad (3.1)$$

where I have used the production function to substitute for  $\tilde{q}_i^F(r, s)$ ,  $L_i^F(r)$  is the farm population of  $r$  (*i.e.* the total number of farmers with residence at  $r$ ), and  $P_i(s)$  is the CES price index of non-farm goods at  $s$ , that is,

$$P_i(s) = \left[ \int_0^{n_i} p_i^N(s, i)^{1-\epsilon} di \right]^{\frac{1}{1-\epsilon}}.$$

I abstract from the institution of slavery in the model. This is important for tractability as a set of agents making location and sectoral choices for others would substantially complicate my setting. That said, it is not an innocuous choice. Economic historians have argued that both slavery and its abolition affected agricultural productivity in the U.S. South, even though it was unlikely to be the major root cause of economic divergence between the North and the South (Kim and Margo, 2004). In principle, such differences across locations could be incorporated in farm productivity  $B_i(r)$ . In practice, however, I lack data on actual farm productivity and match  $B_i(r)$  to the potential yield of location  $r$ 's main crop (Section 4.4). As these potential yields do not reflect man-made institutions, one would expect the model to underestimate farm productivity in the

19. Later in the 19th century, the fraction of farms owned by the farmers decreased but was still 65% in 1900. In light of this evidence, of course, assuming that farmers actually own (and, therefore, can buy and sell) the land at their location would be even more realistic. However, making this assumption would imply that farmers need to account for the whole future distribution of land rents when making their location decision. This would make it infeasible to solve the model with a large number of heterogeneous locations.



South during slavery. In fact, the model underestimates population and economic activity in the South compared to the data (Table 15 in [Supplementary Material, Appendix E](#)), though not by a large amount. This could be due to the fact that slavery's impact on farm productivity is missing from the model. While incorporating the institution of slavery in a quantitative spatial framework is out of the scope of this paper, it is undoubtedly an interesting question left for future research.

**3.1.3. Non-farm workers.** Non-farm workers live, work, and consume at the same location. They have the same preferences as farmers. Therefore, a non-farm worker living (as well as working and consuming) at  $s$  obtains the following utility at time  $t$ :

$$U_t^N(s) = \zeta \left[ \int_0^{n_t} x_t^N(s, i)^{\frac{\epsilon-1}{\epsilon}} di \right]^{\frac{\epsilon}{\epsilon-1}} x_t^F(s)^{1-\nu},$$

where  $x_t^N(s, i)$  is the quantity of non-farm variety  $i$ , and  $x_t^F(s)$  is the quantity of the farm good consumed by the non-farm worker.  $\zeta$  is the same constant as in farmers' utility.

Non-farm workers work for firms operating in the non-farm sector, for a wage that they take as given. Therefore, in each period  $t$ , a non-farm worker at  $s$  chooses her consumption of each good to maximize her utility subject to the budget constraint

$$\int_0^{n_t} p_t^N(s, i) x_t^N(s, i) di + p_t^F(s) x_t^F(s) \leq w_t(s),$$

where  $w_t(s)$  is the non-farm wage at  $s$ .

The indirect utility of a non-farm worker living at  $s$  can then be written as

$$U_t^N(s) = \frac{w_t(s)}{P_t(s)^\nu p_t^F(s)^{1-\nu}}. \quad (3.2)$$

**3.1.4. Non-farm technology.** Non-farm varieties are produced by firms operating under monopolistic competition with free entry. This implies that each non-farm firm produces one variety  $i$  from the mass of varieties  $n_t(s)$  produced at location  $s$ . To operate for a period, the firm needs to hire  $f > 0$  workers, hence the non-farm sector is subject to internal economies of scale. With internal economies of scale, each non-farm firm will choose to operate at its efficient size which, due to CES demand, does not directly depend on the size of the local market (equation (3.8)). As a consequence, locations with a larger non-farm population will produce more varieties (equation (3.11)). In the presence of trade costs, this will make these locations more appealing for non-farm workers. This is the classical agglomeration force of [Krugman \(1991\)](#).

Once operating, a non-farm firm needs labour and the farm good to produce. The production function is CES with elasticity of substitution  $\beta$  between the two inputs,

$$q_t^N(s, i) = \left[ \ell_t^P(s, i)^{\frac{\beta-1}{\beta}} + [\iota \hat{A}_t(s, i) x_t^F(s, i)^\mu]^{\frac{\beta-1}{\beta}} \right]^{\frac{\beta}{\beta-1}},$$

where  $q_t^N(s, i)$  is the quantity of the variety produced,  $\ell_t^P(s, i)$  is the quantity of labour hired for production,  $\hat{A}_t(s, i)$  is the firm's farm good-augmenting productivity, and  $x_t^F(s, i)$  is the quantity of the farm good used.  $\iota = \mu^{-\mu} (1 - \mu)^{-(1-\mu)}$  is a constant that simplifies the subsequent formulas. Also, since varieties are symmetric within a location, I drop the index  $i$  in what follows.

An elasticity of substitution between the two inputs different from one generates a shift of the economy from using farm goods (indirectly, farm labour) to using non-farm labour, hence structural change and urbanization. In particular, we need  $\beta < 1$ , that is, complementarity between labour and the farm good. In this case, an increase in the efficiency of farm good use,  $\hat{A}_t(s)$ , allows firms to hire more workers for production. As a consequence, the non-farm population of  $s$  increases to an extent that depends on the value of parameter  $\beta$ .<sup>20</sup> This specification relates the paper to the strand of the literature originating from the seminal papers by Baumol (1967) and Ngai and Pissarides (2007), in which structural change is induced by demand complementarities and differential productivity growth across sectors. While my model does not feature demand complementarities in consumption (as the typical frameworks in this literature do), it features complementarity between the farm good and labour in non-farm production, which triggers qualitatively similar forces that lead to structural change away from the farm sector.

Firms can increase their productivity by hiring workers to innovate. In particular, I assume that the firm's period- $t$  productivity is the product of its productivity  $A_t(s)$  in the beginning of the period and its period- $t$  innovation  $\ell_t^I(s)^{1-\mu}$ . That is,

$$\hat{A}_t(s) = A_t(s)\ell_t^I(s)^{1-\mu},$$

where  $\ell_t^I(s)$  denotes the number of workers hired to innovate.<sup>21</sup> Workers can freely switch between the two tasks, innovation and production. As I show later in section "Non-farm firms' problem and its solution", a firm always has positive demand for both types of labour. Hence, wages of innovation and production workers are equal in equilibrium, and the marginal non-farm worker is indifferent between performing the two tasks.

Note that the exponents on innovation labour ( $1 - \mu$ ) and the farm good ( $\mu$ ) sum to one. This assumption helps keep the model tractable by guaranteeing constant returns to scale *after* the fixed cost has been paid. Without the assumption, one would not get closed-form solutions for firms' optimal levels of innovation, labour use, farm good use, and output price as a function of input prices and productivity (section "Non-farm firms' problem and its solution"). As the model's equilibrium conditions depend on these endogenous variables, solving the model without the assumption would require numerically solving for these variables by iterating on firms' first-order conditions at every location in every time period. This would make the calibration of the model, which relies on solving the model many times, computationally infeasible.

Trade in non-farm goods is also subject to shipping costs. Non-farm firms at  $s$  can ship their product to location  $u$  at the iceberg cost  $\tau_t^N(s, u)$ .  $\tau_t^N(\cdot, \cdot)$  is an exogenously given measurable function that is symmetric, that is,  $\tau_t^N(s, u) = \tau_t^N(u, s)$ . Note that these types of shipping costs can also change over time. Also note that I assume that non-farm firms do not have the technology to ship the farm good directly. Instead, they can trade the farm good with other trading places after transforming it into a non-farm variety that embodies both the good itself and the labour used in shipping and handling. In other words, I regard the distribution of agricultural

20. As an alternative specification, one could consider the two inputs being substitutes and productivity being labour-augmenting. This would, however, imply that larger cities, everything else fixed, grow faster than smaller ones, which is contrary to the convergence in city sizes found in the data (fact 5 in Section 2).

21. Examples of 19th-century innovations that increased the efficiency of farm good use in two important non-farm sectors, food distribution and food processing, are refrigeration, grain elevators, and canning. Although refrigerated railroad cars and cold-storage warehouses only appeared at a large scale after 1860, innovations aimed at using natural ice blocks for refrigeration predated the Civil War (Anderson, 1953). To store grain in a much more efficient way, a businessman in Buffalo, NY, set up the first steam-powered grain elevator in 1842, and Chicago had a complex system of elevators by the end of the 1850s (Cronon, 1991). The canning industry became firmly established on the East Coast and started moving to the Midwest by the 1840s (Pearson, 2016).

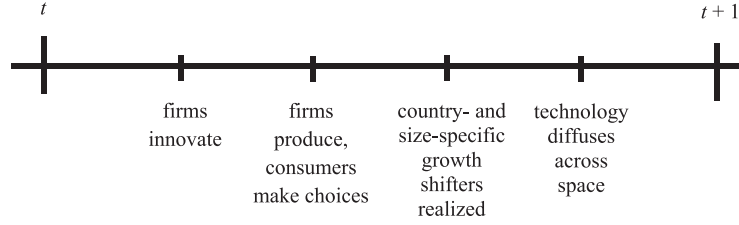


FIGURE 4  
Timing of events in the model

products as a non-farm activity, which corresponds to the way farm and non-farm activities are measured in the data.

**3.1.5. Evolution of productivity.** To incorporate the heterogeneity in growth between cities and towns observed in the data, I allow productivity growth between periods  $t$  and  $t+1$  at location  $s$  to depend not only on firms' period- $t$  innovation but also on size-dependent dynamic externalities  $g(L_t^N(s))$ , where  $L_t^N(s)$  denotes the size of the non-farm population at  $s$ . Thus, I assume that the non-farm productivity of  $s$  evolves according to the equation

$$\tilde{A}_{t+1}(s) = A_t(s)[\ell_t^I(s)^{1-\mu} + g(L_t^N(s))], \quad (3.3)$$

where  $g(\cdot)$  is an exogenously given measurable function. Guided by evidence that cities grow faster than towns but the relationship between growth and size is close to Gibrat's Law in both groups, I choose the functional form of the dynamic externality such that  $g(L_t^N(s)) = \gamma$  if  $L_t^N(s) \geq \lambda$ , and  $g(L_t^N(s)) = 0$  if  $L_t^N(s) < \lambda$ . Hence,  $\gamma > 0$  is a parameter that drives the difference in growth rates between *cities* (locations with non-farm population above  $\lambda$ ) and *towns* (locations with non-farm population below  $\lambda$ ). In the calibration of the model, I set  $\lambda = 10,000$ , the threshold size between cities and towns.

I also assume that non-farm technology can diffuse across U.S. locations between any two subsequent time periods, as in [Desmet and Rossi-Hansberg \(2014\)](#). That is, I allow a non-farm firm in period  $t+1$  to benefit from its most productive neighbours in period  $t$ , but the farther away a neighbour is, the less the firm can benefit. To model this process in the simplest possible way, I assume that firms at location  $s$  cannot only use their own technology  $\tilde{A}_{t+1}(s)$  in period  $t+1$  but can also borrow technology  $e^{-\delta|u-s|}\tilde{A}_{t+1}(u)$  from another location  $u \in S$ .  $|u-s|$  denotes the great-circle distance between  $u$  and  $s$ , and  $\delta > 0$  drives the strength of technology diffusion. The technology that firms at location  $s$  *actually use* in period  $t+1$  is the best of all these available technologies:

$$A_{t+1}(s) = \max_u e^{-\delta|u-s|}\tilde{A}_{t+1}(u). \quad (3.4)$$

Figure 4 shows the timing of events that follows from the above assumptions. Firms at location  $s$  start with productivity  $A_t(s)$  in period  $t$ , and decide how much to innovate. Next, they produce the non-farm good; in doing so, they use the productivity shifted up by their innovation,  $\hat{A}_t(s) = A_t(s)\ell_t^I(s)^{1-\mu}$ . At the same time, consumers choose their sector, location(s), production and consumption quantities, and markets clear. After all these, dynamic externalities are realized and increase productivity to  $\tilde{A}_{t+1}(s)$ , given by equation (3.3). Finally, technology diffuses across space, which leads to a productivity level  $A_{t+1}(s)$  at location  $s$ , given by equation (3.4). The process starts again next period, with this new productivity level at  $s$ .

I assume that total factor productivity in farming does not change over time. That is,  $B_t(r) = B(r)$ , where  $B(\cdot)$  is an exogenous measurable function. This assumption is motivated by evidence showing that only a tiny fraction of U.S. output growth can be explained by growth in agricultural TFP in the 19th century (Mundlak, 2005). It is also in line with the broader evidence that average TFP growth in agriculture was below average TFP growth in non-agriculture in advanced economies before the Second World War (Herrendorf *et al.*, 2014).<sup>22</sup> It is important to note, however, that average labour productivity on farms, defined as real output per farmer, did increase between 1790 and 1860, due to the occupation of new land in the Midwest and the fact that new land was more productive than the land used before. This increase in average labour productivity, calculated from the estimates by Towne and Rasmussen (1960), is matched almost exactly by the model. Hence, assuming an increase in farm TFP during the period would lead to a larger increase in labour productivity in the model than in the data.

**3.1.6. Non-farm firms' problem and its solution.** Non-farm firms choose the path of their production and innovation to maximize the present discounted value of their profits, taking into account the technology and market constraints described in sections "Non-farm technology" and "Evolution of productivity". This dynamic optimization problem simplifies to a sequence of static problems, as in Desmet and Rossi-Hansberg (2014). To see why, note that technology diffusion is perfect locally: equation (3.4) implies that a firm's innovation in period  $t$  becomes freely available for all other firms at the same location next period. As a result, all local firms have the same technology in the beginning of period  $t + 1$ , irrespectively of their choice in period  $t$ . Since they also face the same prices, their profits are identical. However, free entry drives down this common level of profits to zero. Therefore, a firm choosing its innovation and production in period  $t$  knows that it can only make zero profits in the future, thus the present discounted value of its profits equals its period- $t$  profits. In other words, the firm solves a static profit maximization problem in period  $t$ .

Once firms solve a static problem, monopolistic competition, CES demand, and iceberg shipping costs imply that they charge a constant markup over their unit variable cost:

$$p_t^N(s) = \frac{\epsilon}{\epsilon - 1} \bar{c}_t(s) w_t(s), \quad (3.5)$$

where  $\bar{c}_t(s)$  denotes the firm's unit variable cost per wage. Also, free entry drives down profits to zero. That is,

$$p_t^N(s) q_t^N(s) - \bar{c}_t(s) w_t(s) q_t^N(s) - w_t(s) f = 0,$$

which implies, using equation (3.5), that the firm's output is

$$q_t^N(s) = (\epsilon - 1) f \bar{c}_t(s)^{-1}. \quad (3.6)$$

As the production function is CES, the unit variable cost per wage can be expressed as

$$\bar{c}_t(s) = \left[ 1 + A_t(s)^{\beta-1} \left[ \frac{p_t^F(s)}{w_t(s)} \right]^{(1-\beta)\mu} \right]^{\frac{1}{1-\beta}} \quad (3.7)$$

22. Agricultural TFP started growing faster than non-agricultural TFP after the Second World War. In the model, this could be captured by letting  $B_t(r)$  grow exogenously, or by allowing for endogenous innovation in the farm sector. How agricultural TFP growth in the late-20th century may have affected urbanization and city development is an interesting question left for further research.

and optimal factor quantities can be obtained from Shephard's Lemma and equation (3.6) as

$$\ell_t^P(s) = (\epsilon - 1)f\bar{c}_t(s)^{\beta-1}, \quad (3.8)$$

$$\ell_t^I(s) = (1 - \mu)(\epsilon - 1)fA_t(s)^{\beta-1} \left[ \frac{P_t^F(s)}{w_t(s)} \right]^{(1-\beta)\mu} \bar{c}_t(s)^{\beta-1}, \quad (3.9)$$

and

$$x_t^F(s) = \mu(\epsilon - 1)fA_t(s)^{\beta-1} \left[ \frac{P_t^F(s)}{w_t(s)} \right]^{(1-\beta)\mu-1} \bar{c}_t(s)^{\beta-1}. \quad (3.10)$$

Finally, market clearing for non-farm labour pins down the number of non-farm goods produced at the location:

$$n_t(s) = \frac{L_t^N(s)}{\ell_t^P(s) + \ell_t^I(s) + f} = \frac{L_t^N(s)}{[(\epsilon - 1)(\mu\bar{c}_t(s)^{\beta-1} + 1 - \mu) + 1]f}, \quad (3.11)$$

where  $L_t^N(s)$  is the non-farm population of  $s$ , determined endogenously in equilibrium by non-farm workers' location choices.

**3.1.7. Equilibrium.** I define the equilibrium of the economy as follows.

**Definition.** Given parameters  $\alpha, \beta, \gamma, \delta, \epsilon, \lambda, \mu, \nu, f$ , the set of locations  $S$ , total population  $\bar{L}_t$ , as well as functions  $H_t, A_1, B : S \rightarrow \mathbb{R}$  and  $\tau_t^F, \tau_t^N : S^2 \rightarrow \mathbb{R}$ , an **equilibrium** of the economy is a set of functions  $L_t^F, L_t^N, p_t^F, p_t^N, q_t^N, h_t, \ell_t^P, \ell_t^I, x_t^F, \bar{c}_t, n_t, P_t, w_t, R_t, A_t : S \rightarrow \mathbb{R}$  and  $\sigma_t : S \rightarrow S$ , as well as the level of consumers' utility  $U_t$  for each time period  $t \in \{1, 2, \dots\}$  such that the following hold:

1. Farmers maximize utility, implying that farmers living at  $r$  choose their trading place  $\sigma_t(r)$ <sup>23</sup> such that

$$\sigma_t(r) = \operatorname{argmax}_{s \in S} U_t^F(r, s)$$

and their utility is equalized across locations, that is,

$$U_t^F(r, \sigma_t(r)) = U_t^F(u, \sigma_t(u)) \quad (3.12)$$

for any  $r, u \in S$ , such that  $U_t^F(\cdot, \cdot)$  is given by equation (3.1).

2. Non-farm workers maximize utility, and their utility is equalized across locations. That is,

$$U_t^N(s) = U_t^N(u) \quad (3.13)$$

for any  $s, u \in S$  with positive non-farm population, where  $U_t^N(\cdot)$  is given by equation (3.2).

3. Consumers' utility is equalized across sectors. Hence,

$$U_t = U_t^F(r, s) = U_t^N(s) \quad (3.14)$$

for any  $r, s \in S$  such that  $s = \sigma_t(r)$ .

23. Farmers at a given location are all identical, hence they choose the same trading place to ship to. Also, I only consider equilibria in which  $\sigma_t(\cdot)$  is a measurable function.

4. *Non-farm firms maximize profits, and free entry drives down profits to zero. Hence, their price is given by equation (3.5), their output by equation (3.6), their unit cost per wage by equation (3.7), and their factor use by equations (3.8)–(3.10). The number of goods produced at location  $s$  is given by equation (3.11), and the price index at  $s$  is<sup>24</sup>*

$$P_t(s) = \left[ \int_S n_t(u) p_t^N(u)^{1-\epsilon} \tau_t^N(u, s)^{1-\epsilon} du \right]^{\frac{1}{1-\epsilon}}. \quad (3.15)$$

5. *The market for the farm good clears at each trading place  $s$ . That is,*

$$\int_{\sigma_t^{-1}(s)} v \tau_t^F(r, s)^{-1} B(r) L_t^F(r)^\alpha H_t(r)^{1-\alpha} dr = x_t^F(s) n_t(s) + (1-v) \frac{w_t(s)}{p_t^F(s)} L_t^N(s), \quad (3.16)$$

where  $\sigma_t^{-1}(s)$  denotes the set of locations from which farmers ship to  $s$ , the left-hand side corresponds to supply net of farmers' demand at  $s$ , the first term on the right-hand side corresponds to firms' demand, and the second term on the right-hand side corresponds to non-farm workers' demand.

6. *The market for non-farm goods clears at each trading place  $s$ . That is,*

$$q_t^N(s) = \int_S v p_t^N(s)^{-\epsilon} \tau_t^N(s, u)^{1-\epsilon} P_t(u)^{\epsilon-1} I_t(u) du, \quad (3.17)$$

where the left-hand side corresponds to the supply of any non-farm variety produced at  $s$ , and the right-hand side corresponds to total demand for the variety.  $I_t(u)$  denotes total income of consumers at  $u$ , and equals the sum of farmers' and non-farm workers' income:

$$I_t(u) = \int_{\sigma_t^{-1}(u)} p_t^F(u) \tau_t^F(r, u)^{-1} B(r) L_t^F(r)^\alpha H_t(r)^{1-\alpha} dr + w_t(u) L_t^N(u).$$

7. *The national labour market clears. That is,*

$$\bar{L}_t = \int_S [L_t^F(r) + L_t^N(r)] dr. \quad (3.18)$$

8. *The market for land clears at each location. That is,*

$$H_t(r) = L_t^F(r) h_t(r). \quad (3.19)$$

9. *Productivity levels evolve according to equations (3.3) and (3.4).*

### 3.2. Extension to multiple countries

This section extends the model to incorporate a finite number of countries  $c \in \{1, \dots, C\}$ , such that each location  $r$  belongs to one country  $c(r)$ .<sup>25</sup> Consumers can move freely within countries, but not across countries. Thus, countries' population levels  $\bar{L}_{ct}$  are given exogenously. Trade can take place both within and across countries, but is subject to shipping costs.

24. I use continuous space notation for simplicity, but note that all the results carry over to discrete space. In that case, integrals taken over space should be replaced by sums.

25. I assume that the area of each country  $c$  is a Borel measurable subset of  $S$ .



In the multi-country model, I allow productivity growth between periods  $t$  and  $t + 1$  at location  $s$  to also depend on country-specific exogenous growth shifters  $f_{c(s),t}$ . Thus, I assume that the non-farm productivity of  $s$  evolves according to the equation

$$\tilde{A}_{t+1}(s) = A_t(s)[\ell_t^I(s)^{1-\mu} + f_{c(s),t} + g(L_t^N(s))]. \quad (3.20)$$

All other parts of the model are unchanged. [Supplementary Material, Appendix A.1](#) defines the equilibrium of the multi-country model and shows that equilibrium conditions in the multi-country model are the same as in the single-country model, except that utility equalization is only imposed within countries, national labour markets clear by country, and productivity is assumed to evolve according to equation (3.20) instead of equation (3.3).

### 3.3. The role of farm hinterlands

In this section, I formalize the hinterland hypothesis and show that the model yields it as a prediction. To accomplish this, I first define the size of a location's *effective hinterland*:

**Definition.** Consider a location  $s \in S$  with positive non-farm population  $L_t^N(s)$  at time  $t$ . The size of the **effective hinterland** of  $s$  is given by

$$\tilde{H}_t(s) = \frac{\int_{\sigma_t^{-1}(s)} L_t^F(r) dr}{L_t^N(s)}.$$

My measure of effective hinterland size captures the size of location  $s$ 's farm input market,  $\int_{\sigma_t^{-1}(s)} L_t^F(r) dr$ , relative to the size of its non-farm labour market,  $L_t^N(s)$ . The more important role the farm input plays in  $s$ 's local economy, the larger  $\tilde{H}_t(s)$ . This is the sense in which this measure captures the local importance of  $s$ 's farm hinterland.

In the next proposition, I show that farm population in the hinterland of location  $s$ , the non-farm population of  $s$ , and innovation at  $s$  all increase in  $\tilde{H}_t(s)$ . In other words, I show that the model yields the hinterland hypothesis as a prediction.

**Proposition 1** (Hinterland hypothesis). Consider a location  $s$  with positive non-farm population  $L_t^N(s)$  and effective hinterland size  $\tilde{H}_t(s)$  at time  $t$ . Everything else fixed, an increase in  $\tilde{H}_t(s)$  raises

- (i) the farm population of any location in  $s$ 's hinterland  $r \in \sigma_t^{-1}(s)$ , and thus the total farm population of the hinterland,
- (ii) the non-farm population of  $s$ , and thus the likelihood that  $s$  is a city,
- (iii) innovation at  $s$ , and thus total innovation and growth in the economy.

*Proof.* See [Supplementary Material, Appendix A.6](#).

The intuition behind Proposition 1 is as follows. At locations where  $\tilde{H}_t(s)$  is large, local non-farm firms have a high farm input to labour input ratio. As the two inputs are complements, everything else fixed, such firms will hire more non-farm labour in absolute terms than firms at locations with smaller effective hinterland size. At the same time, firms at locations with large  $\tilde{H}_t(s)$  also have an incentive to make their dominant input (the farm input) as efficiently used as

possible. In the model, they can do that by conducting innovation, which increases the efficiency with which the farm input is used in their production process.<sup>26</sup>

Proposition 1 suggests that changes in geography that allow a location to expand its effective hinterland can speed up the settlement of farmers near the location, the settlement of non-farm population at the location (thus city development), and local economic growth. What does this mean in practice? For example, if railroads had increased the sizes of cities' potential farm hinterlands, particularly in the West, then they might have contributed to the westward expansion, city development, and growth. Changing U.S. political borders, by increasing the amount of land available in the West, might have played a similar role. By contrast, international trade, which favoured Eastern locations as they were located closer to foreign markets, might have slowed down the westward expansion. At the same time, it might have fostered city development and growth in the East. The following corollary formalizes these statements by showing how railroads, borders, or international trade affect spatial and aggregate development in the model.

**Corollary of Proposition 1** (Implications of the hinterland hypothesis). *Suppose that (1) the construction of railroads, or (2) the expansion of U.S. political borders, or (3) a decrease in international trade costs increases the size of the effective hinterland of location  $s$ . Then such changes in geographic fundamentals raise*

- (i) *the farm population of any location in  $s$ 's hinterland  $r \in \sigma_t^{-1}(s)$ , and thus the total farm population of the hinterland,*
- (ii) *the non-farm population of  $s$ , and thus the likelihood that  $s$  is a city,*
- (iii) *innovation at  $s$ , and thus total innovation and growth in the economy.*

*Proof.* See [Supplementary Material, Appendix A.6](#).

An unappealing feature of the above corollary is that it assumes that railroads, borders, or international trade expand the hinterland of location  $s$  but does not show that this is the case. Unfortunately, the complexity of the general model does not allow me to prove that such changes in geography around  $s$  indeed raise  $s$ 's effective hinterland size. That said, I present a simplified special case of the model in [Supplementary Material, Appendix A.3](#) in which I can characterize the relationship between geography and effective hinterland size. The key simplification in this special case is that the hinterland around  $s$  is homogenous in its geographic characteristics. Such an assumption allows me to formally prove that  $s$ 's effective hinterland size increases with railroad construction, political border expansion, or falling international trade costs around  $s$  (Proposition 2 in [Supplementary Material, Appendix A.3](#)).

### 3.4. Solving the model

In this section, I show that the model displays a recursive structure. In particular, the following lemma shows that the evolution of productivity between periods  $t$  and  $t + 1$  only depends on the spatial distribution of population in period  $t$ :

26. The result that innovation is increasing in the ratio of  $\int_{\sigma_t^{-1}(s)} L_t^F(r) dr$  to  $L_t^N(s)$  also implies that innovation exhibits convergence: everything else fixed, locations with a large non-farm population innovate less. This is what allows the model to replicate the fact that growth exhibits convergence both among cities and among towns (fact 5 of Section 2).

**Lemma 1.** *The amount of innovation at location  $s$  in period  $t$  is given by*

$$\ell_t^I(s)^{1-\mu} = \rho \left[ \left( 1 + \frac{L_t^N(s)}{\int_{\sigma_t^{-1}(s)} L_t^F(r) dr} \right)^{-1} - (1-\nu) \right]^{1-\mu}, \quad (3.21)$$

where  $\rho = [(\epsilon(1-\mu)/\mu\nu)f]^{1-\mu}$  is a constant.

*Proof.* See [Supplementary Material, Appendix A.6](#).

[Supplementary Material, Appendix A.2](#) shows that the distribution of population in any period  $t$  can be obtained by solving a system of three equations. This result, together with Lemma 1, suggests the following way of solving the equilibrium forward. Having the initial distribution of productivity  $A_1(\cdot)$ , one can calculate the population distribution in period 1. Then one can use equations (3.21), (3.20), and (3.4) to update productivity levels at each location, and obtain  $A_2(\cdot)$ .  $A_2(\cdot)$ , in turn, can be used to calculate the population distribution in period 2, which allows one to update productivities again, and so on. This is the procedure I use to simulate the model when taking it to the data, which I describe in detail in the next section.<sup>27</sup>

#### 4. EMPIRICAL STRATEGY

This section describes the empirical strategy I follow to take the model to the data. First, I discretize the U.S. into a fine spatial grid and incorporate the rest of the world in the analysis. Next, I show how evidence from [Fogel \(1964\)](#) and [Donaldson and Hornbeck \(2016\)](#) can be used to calculate (relative) shipping costs, and how data on agricultural yields can be used to calculate the spatial distribution of farm productivity within the U.S. Finally, I calibrate the productivity of foreign markets, the initial distribution of non-farm productivity, the scale of shipping costs, and the values of structural parameters such that the simulated model matches two sets of moments in the data: moments of the U.S. economy and Europe in the initial period; and moments of aggregate and city growth in the U.S. between 1790 and 1820.

The advantage of this empirical strategy is twofold. First, it leaves the evolution of individual locations' population and sectoral specialization untargeted, both before and after 1820. This allows me to test the model by checking how well it predicts the evolution of the U.S. population distribution and the formation of cities in Section 5.1.

The second advantage of my empirical strategy is that it mitigates the concern that endogenous railroad and canal placement may bias the identification of structural parameters, since no railroads or navigable canals were built due to the lack of technology before the 1820s. Whenever railroad and canal construction technology is available, there might be a structural relationship that links the location of railroads and canals to economic outcomes. The theory developed in this paper is consistent with such a relationship, but does not model it explicitly, and ignoring

27. For any given population distribution, equations (3.21), (3.20), and (3.4) pin down a unique distribution of productivity next period. Therefore, uniqueness of the equilibrium depends only on whether the period- $t$  population distribution is unique for given period- $t$  productivities. In economic geography models, the equilibrium usually ceases to be unique if forces of agglomeration are very strong relative to forces of dispersion. Although the complex structure of the model does not allow me to theoretically characterize multiplicity of equilibria, solving the period- $t$  equilibrium with different initial guesses on the population distribution has always led to the same equilibrium in the simulations. This suggests that the model is likely to feature a unique equilibrium for the values of parameters used in the calibration. [Supplementary Material, Appendix A.4](#) provides intuition for why the structure of the model does not allow for a theoretical characterization of equilibrium uniqueness.

the relationship in the calibration may affect the identified values of parameters. However, this is not true for a period in which the technology is unavailable, hence the relationship between railroad placement and economic outcomes is irrelevant.

#### 4.1. *Setting up a spatial grid*

The unit of observation I choose is a *cell* in a 20 by 20 arc minute grid of the U.S. Although the model allows for both discrete and continuous space, computational tractability makes discretization of the data necessary. A discretization to 20 by 20 arc minutes means that each cell in the grid is approximately 20 by 20 miles large, and the entire territory of the U.S. in the final period consists of 7641 such grid cells.

Each grid cell  $r$  is characterized by three geographic attributes.  $H_t(r)$  tells us what fraction of the cell is covered by land *and* is part of the U.S. in period  $t$ .<sup>28</sup>  $WR_t(r)$  is a dummy variable taking the value of one if part of cell  $r$  is a navigable river, canal, lake, or the sea in period  $t$ . Finally,  $RR_t(r)$  is a dummy variable that equals one if there was a railroad passing through the cell in period  $t$ . Data on water come from the ESRI Map of U.S. Major Waters, while railroad data come from historical railroad maps available online at [oldrailhistory.com](http://oldrailhistory.com).<sup>29</sup>

#### 4.2. *Incorporating international trade*

I assume that the world not only consists of a single country (the U.S.) but also includes a point-like country representing foreign markets that the U.S. trades with. This international dimension is expected to be quantitatively important for the results since economic historians estimate that exports constituted approximately 9% of U.S. GDP in the 1790s. Although this ratio decreased later, it never went substantially below 5% (Lipsey, 1994).

I identify the foreign country with the European continent for two reasons. First, evidence suggests that a vast majority, about 60–75%, of U.S. exports went to Europe between 1790 and 1860 (Lipsey, 1994). Second, as explained in Section 4.5, simulation of the model requires data on the foreign country's GDP, farm and non-farm populations, and high-quality data on these are not available outside Europe during the period of investigation.

#### 4.3. *Calculating shipping costs*

I use grid cells' geographic attributes to calculate bilateral shipping costs across cells. Shipping costs largely depend on the mode of transportation that locations have access to. Based on evidence from Fogel (1964), Donaldson and Hornbeck (2016) argue that water transportation was the least expensive mode of shipping goods in the 19th century, followed by rail transportation. Wagon transportation was an order of magnitude more expensive than these other modes.

Motivated by this evidence, I assume that the cost of shipping through a cell takes the following form. The cost of shipping through cell  $r$  in period  $t$  is (1) normalized to one if  $WR_t(r) = 1$ , that is, the cell has a waterway in it; (2) equals  $\varsigma^R > 1$  if  $WR_t(r) = 0$  and  $RR_t(r) = 1$ , that is, the cell does not have access to water but does have railroads; and (3) equals  $\varsigma^I > \varsigma^R$  if  $WR_t(r) = RR_t(r) = 0$ , that is, the cell has neither water nor railroad access at time  $t$ . Once

28. This allows me to incorporate changes in U.S. political borders between 1790 and 1860. Note that, as agents' dynamic problems reduce to a sequence of static problems in the model, agents' expectations about future border changes do not influence the results.

29. I incorporate railroads and navigable canals starting from the year of their construction. Details of this procedure are provided in [Supplementary Material, Appendix B](#).

the cost of shipping through each cell is known, one can calculate bilateral costs by searching for the minimum-cost route of getting from a cell  $r$  to another cell  $s$ . I apply the fast marching algorithm to determine these minimum costs.

I base the values of railroad and wagon costs relative to water costs,  $\varsigma^R$  and  $\varsigma^I$ , on the relative freight rate estimates of [Donaldson and Hornbeck \(2016\)](#), who, in turn, borrow the estimates from [Fogel \(1964\)](#). Since these estimates suggest a water freight rate of 49 cents per ton-mile, a rail rate of 63 cents per ton-mile and a wagon rate of 23.1 dollars per ton-mile, this amounts to setting  $\varsigma^R = 0.63/0.49$  and  $\varsigma^I = 23.1/0.49$ .

As I normalized the cost of shipping along water to one, these calculations only provide me with the costs of shipping between any pair of locations relative to the water cost. In other words, my calculation misses a scaling factor in shipping costs. Moreover, it is reasonable to assume that this scaling factor may differ between the shipping of farm and non-farm products. As a result, I set the iceberg cost of shipping between U.S. locations  $r$  and  $s$  equal to

$$\tau_t^F(r, s) = e^{\phi^F \log \varsigma_t(r, s)}$$

and

$$\tau_t^N(r, s) = e^{\phi^N \log \varsigma_t(r, s)}$$

for farm and non-farm goods, respectively, where  $\varsigma_t(r, s)$  denotes the relative shipping cost calculated above between  $r$  and  $s$  in period  $t$ .<sup>30</sup> In what follows, I treat  $\phi^F$  and  $\phi^N$  like the structural parameters of the model, and calibrate them to match moments of the U.S. economy that are intuitively related to the scale of shipping costs in Section 4.5.

Unlike [Donaldson and Hornbeck \(2016\)](#), my model also includes trade with Europe. I calculate the shipping costs between U.S. locations and Europe in the following way. For any U.S. location  $u$  that is on the Atlantic coast, I measure the great-circle distance between  $u$  and the port of London in miles,  $|u - e|$ . I choose London since it was the largest port of the European continent at the time, and the U.K. was the most important trading partner of the U.S. within Europe ([Lipsey, 1994](#)). Next, I set up the cost of shipping from  $u$  to Europe, analogously to within-U.S. shipping costs, as

$$\tau_t^N(u, e) = e^{\phi^E \log |u - e|},$$

where  $\phi^E$  is a parameter driving the scale of international shipping costs. Akin to  $\phi^F$  and  $\phi^N$ , I treat  $\phi^E$  like a structural parameter and calibrate it to match the U.S. exports to GDP ratio in 1790. Finally, for any U.S. location  $s$  that is not on the Atlantic coast, I set the shipping cost with Europe to the minimum of the product of shipping costs to locations along the coast and shipping costs between those locations and Europe:

$$\tau_t^N(s, e) = \min_{u \in \text{Atlantic coast}} \tau_t^N(s, u) \tau_t^N(u, e).$$

#### 4.4. Calculating the distribution of farm productivity

I use high-resolution data on agricultural yields to calculate the spatial distribution of farm productivity  $B(\cdot)$ . I collect data on potential yields of the six main 19th-century U.S. crops (cereals,

30. As the exponents are always non-negative, this functional form also guarantees that all iceberg trade costs are larger or equal to one.

cotton, sugar cane, tobacco, white potato, and sweet potato) from the Food and Agriculture Organization's Global Agro-Ecological Zones database (FAO GAEZ). The earliest period for which the yields are available in this dataset is between 1961 and 1990.

As soil and climate conditions might have changed over the last two hundred years, the ideal data to use would be high-resolution yield data for the 19th century. Unfortunately, I am not aware of any yield data at the 20 by 20 arc minute spatial resolution that predates the FAO data. Hence, I follow the literature that has been using the FAO data to proxy historical agricultural yields, such as [Galor and Ozak \(2016\)](#) and [Mayshar \*et al.\* \(2022\)](#), and apply the filters “low input level” and “rain-fed” to be as close as possible to historical conditions.<sup>31</sup> Proxying 19th-century yields with the ones for the period 1961–90 involves the assumption that the yields of different crops have not changed, or only by the same factor across different locations. I need to make this assumption to proceed with the analysis.

Since different locations specialized in growing different crops, the potential yield of sugar cane is likely to be irrelevant for farm productivity in regions growing cereals, and vice versa. To solve this issue, I turn to the 1860 Census of Agriculture, which provides information on the output of different crops at the county level, for the entire territory of the U.S. For each county, I determine its main crop as the one having the largest share in the county's value of farm output.<sup>32</sup> Next, I discretize counties into grid cells and assign the main crop to each cell. Finally, I use the spatial distribution of crop-level productivities coming from FAO to assign the productivity of their main crop to grid cells.

#### 4.5. Calibration

It remains to choose the initial distribution of non-farm productivity across U.S. locations, the farm and initial non-farm productivity of Europe, the country-specific growth shifters, the scale parameters of shipping costs, and the values of the five remaining structural parameters.

**4.5.1. Structural parameters borrowed from the literature.** I borrow the values of four parameters—the labour share in farming  $\alpha$  and the non-farm share in consumption  $\nu$ , consumers' elasticity of substitution among non-farm varieties  $\epsilon$ , and the strength of technology diffusion  $\delta$ —from the literature. In particular, I follow [Caselli and Coleman \(2001\)](#) and set the labour share in farming to  $\alpha = 0.71$ .<sup>33</sup> To choose the non-farm share in consumption, recall that the model regards agricultural products traded across trading places, and not between a trading place and a farm, as non-farm goods. Therefore, the non-farm consumption share should not only include manufacturing and services but also interregionally traded farm goods. Guided by this and the estimates of 19th-century non-food consumption shares by [Lebergott \(1996\)](#) and [Lindstrom \(1979\)](#) which range between 55% and 75%, I set the non-farm consumption share to the upper bound of these estimates,  $\nu = 0.75$ .

The elasticity of substitution among non-farm varieties  $\epsilon$  drives the elasticity of trade with respect to trade costs. Based on the estimate of [Donaldson and Hornbeck \(2016\)](#) for late 19th-century trade across U.S. counties, I set this elasticity to eight, which implies a value of  $\epsilon = 9$ .<sup>34</sup>

31. [Costinot \*et al.\* \(2016\)](#) and [Dingel \*et al.\* \(2021\)](#) are other papers using the FAO data, but for modern times.

32. Data are not available for a few counties in the West; to these counties, I assign the main crop of their closest neighbours.

33. [Caselli and Coleman \(2001\)](#) find a labour share of 0.6, a land share of 0.19, and a capital share of 0.21 in farming. Since my model does not include capital use, I allocate the share of capital equally between labour and land.

34. This mapping between the trade elasticity in my model and the elasticity estimated by [Donaldson and Hornbeck \(2016\)](#) is perfect under the assumption that all cross-county trade took place across trading places and not between



Finally, to obtain the value of the technology diffusion parameter  $\delta$ , I rely on evidence provided by [Comin \*et al.\* \(2013\)](#) on the spatial diffusion of technology. [Comin \*et al.\* \(2013\)](#) estimate the value of  $\delta$  for various 20th-century technologies, and find an average of  $\delta = 0.0025$ . They also argue that technology diffusion depends crucially on the frequency of human interactions across space, and that human interactions have become more frequent as newer technologies replaced older ones. Guided by these findings, I use a conservative estimate of  $\delta = 0.006$ , which implies a spatial decay of 45% in productivity over 100 km. Given the uncertainty about the specific value of  $\delta$ , I check the robustness of my quantitative results to different values of this parameter in Section 6.1.

**4.5.2. Non-farm productivity in Europe.** To choose the initial non-farm productivity of Europe  $A_1(e)$  and the European growth shifters  $f_{et}$ ,<sup>35</sup> note that, by equation (3.20), this is the same problem as choosing the non-farm productivity of Europe every period,  $A_t(e)$ . Also note that, combining equations (3.7), (3.10), (3.11), and (3.16), it is possible to express the non-farm productivity of Europe as

$$A_t(e) = \epsilon^{\frac{1}{\beta-1}} \tilde{B}(e)^{-\mu} L_t^F(e)^{(1-\alpha)\mu} \left[ \frac{\mu(\epsilon - 1)}{v(L_t^F(e)/L_t^N(e)) - (1 - v)} - \mu - \epsilon(1 - \mu) \right]^{\frac{1}{1-\beta}}, \quad (4.1)$$

where  $\tilde{B}(e) = B(e)H(e)^{1-\alpha}$  is a combination of European farm productivity and land. Hence, the non-farm productivity of Europe can be obtained from the continent's land-adjusted farm productivity, farm population, and non-farm population.

Although the literature does not have precise estimates of farm and non-farm populations for the first half of the nineteenth century, there exist good estimates for urban and rural populations ([Bairoch and Goertz, 1985](#)). Having those estimates, I follow [Allen \(2000\)](#) and assume that 75% of rural population was employed in farming, while 25% of rural population and 100% of urban population were employed in the non-farm sector. This provides me with estimates of European farm and non-farm populations.<sup>36</sup> Plugging these estimates into equation (4.1), I can back out the non-farm productivity of Europe in every period  $t$  as a function of land-adjusted farm productivity  $\tilde{B}(e)$  and structural parameters.

**4.5.3. Initial non-farm productivity in the U.S..** Five cities already existed in the U.S. in 1790: New York City, Philadelphia, Boston, Charleston, and Baltimore. Through the lens of the model, this means that these five locations might have accumulated non-farm productivity prior to my initial period, which should be taken into account in the calibration. As a result, I allow the non-farm productivity of these five locations to differ from the non-farm productivity of other U.S. locations: I assume that non-farm productivity was initially distributed such that  $A_1(s) = \bar{A}$  if cell  $s$  is one of the five pre-existing cities, and  $A_1(s) = \underline{A}$  otherwise. I treat  $\bar{A}$  and  $\underline{A}$  like the structural parameters of the model, and calibrate them to match moments of the U.S. population distribution in 1790, as explained below.

**4.5.4. Calibration.** We are left with choosing the values of non-farm productivity parameters  $\bar{A}$  and  $\underline{A}$ , the scale parameters of shipping costs  $\phi^F$ ,  $\phi^N$ , and  $\phi^E$ , Europe's land-adjusted

a trading place and its farm hinterland. Although the lack of data does not allow me to test this assumption empirically, a vast majority of trade across grid cells is indeed non-farm good trade in the model simulations.

35. I normalize the U.S. growth shifter to zero in each period.

36. According to the estimates, the share of the non-farm sector in European employment was 32.8% in 1790. This share increased to 39.1% by 1860.

TABLE 4  
Calibration summary

Parameter	Target
$f = 0.783$	Annual growth rate of U.S. real GDP per capita before 1820 (Weiss, 1992)
$\beta = 0.85$	Average log decennial growth rate of cities, 1790–1820
$\mu = 0.78$	Farm employment share, 1800 (Weiss, 1992)
$\lambda = 10,000$	Threshold between cities and towns (fact 5 in Section 2)
$\gamma = 0.008$	Difference in average log decennial growth rate between cities and towns, 1790–1820
$\bar{A} = 23.6$	Total population of city cells, 1790
$\underline{A} = 18.5$	Total population of town cells, 1790
$\tilde{B}(e) = 0.438$	Europe's real GDP (relative to U.S.), 1790
$\phi^F = 0.09$	Convergence in decennial city growth, 1790–1820
$\phi^N = 0.052$	Ratio of non-farm population living in cities to total population, 1790
$\phi^E = 0.098$	U.S. export to GDP ratio, 1790 (Lipsey, 1994)

Notes: Each row in this table corresponds to one of the eleven calibrated structural parameters of the model of Section 3. The first column shows the value of the parameter chosen in the baseline calibration. The second column specifies the moment targeted with the parameter.

farm productivity  $\tilde{B}(e)$ , and the five remaining structural parameters: the fixed cost in non-farming  $f$ , the elasticity of substitution between labour and the farm good in non-farming  $\beta$ , the share of the farm input in non-farming  $\mu$ , as well as  $\lambda$  and  $\gamma$  that drive dynamic externalities. I choose the values of these eleven parameters such that simulated moments of the U.S. and European economies in the initial period and aggregate U.S. growth and urbanization until 1820 equal the corresponding moments in the data.

Finding the vector of model parameters  $\pi$  that rationalize the data moments requires me to choose the  $\pi$  that minimizes the loss function

$$\text{Loss} = \sum_{j=1}^{11} |M_j^m(\pi) - M_j^d|,$$

where  $j$  is the index of moment  $j$ ,  $M_j^m(\pi)$  denotes the value of this moment in the model for a given parameter vector  $\pi$ , and  $M_j^d$  denotes the value of the same moment in the data. Given that I choose the same number of moments (eleven) as the number of parameters, I conjecture that the loss function reaches its minimum at zero. In other words, my aim is to find a set of parameters at which the model moments exactly coincide with the corresponding data moments. Although the complex non-linear structure of the model does not allow me to prove this conjecture, I find that it is true in practice: at the calibrated values of structural parameters, the value of each model moment coincides with its counterpart in the data.

In what follows, I describe the moments I choose to match for each model parameter. Although the identification of each parameter depends on the values of other parameters, it is still intuitive to think of the calibration as if each parameter were chosen to match one moment between the model and the data. Table 4 provides a brief summary of the calibration.

As for the fixed cost in non-farming  $f$ , Lemma 1 establishes a positive relationship between this parameter and the growth rate of productivity, as an increase in  $f$  raises innovation uniformly across space. Therefore, I choose the value of  $f$  to match the fact that U.S. real GDP per capita grew by about 0.5% per year before 1820 (Weiss, 1992). This procedure pins down  $f = 0.783$ .

To find the elasticity of substitution  $\beta$ , recall that it plays a key role in generating city growth, as discussed in section “Non-farm technology”. The lower the value of  $\beta$ , the higher the degree of complementarity between productivity and non-farm labour, hence the more productivity growth draws non-farm population into cities. Thus, I choose the value of  $\beta$  such that average

city growth between 1790 and 1820 (in logs) is the same in the model as in the data. This suggests setting  $\beta = 0.85$ .

To choose the farm input share in the non-farm sector  $\mu$ , note that an increase in  $\mu$  increases demand for the farm input, and therefore farm labour. Thus, I calibrate  $\mu$  to match the share of farm employment in total employment, which equalled 74% in 1800 (Weiss, 1992).

The presence of dynamic externalities in cities implies that productivity growth differs between cities and towns in the model. Evidence on city and town growth from the data (fact 5 in Section 2) suggests that the threshold between these two types of locations lied at 10,000 inhabitants, which suggests setting  $\lambda = 10,000$ . The difference between the average decennial population growth rates of cities above 10,000 inhabitants and towns between 2500 and 10,000 inhabitants is 0.08 log points in the data between 1790 and 1820. I choose the value of  $\gamma$  such that the decennial growth rates of cities and towns (locations above versus below a non-farm population of  $\lambda$ ) differ by the same number in the model. This implies  $\gamma = 0.008$ . In Section 6.1, I check the robustness of my quantitative results to shutting down dynamic externalities in cities, *i.e.* to setting  $\gamma = 0$  instead.

To calibrate the parameter driving the non-farm productivity of pre-existing cities,  $\bar{A}$ , I measure the total population of cells with pre-existing cities. I choose the value of  $\bar{A}$  such that the population of these cells is identical between the model and the data, thus capturing the initial stock of productivity accumulated at these locations before 1790. I use a similar procedure to calibrate the value of non-farm productivity outside pre-existing cities,  $\underline{A}$ . In particular, I choose  $\underline{A}$  such that the total population of town cells (cells with a non-farm settlement above 2500 inhabitants) coincides in the model and in the data. Intuitively, increasing  $\underline{A}$  makes the non-farm sector outside cities more productive, hence attracting workers from the farm sector and increasing the population of non-farm towns.

As a higher value of Europe's land-adjusted farm productivity  $\tilde{B}(e)$  implies higher real GDP in Europe, I choose the value of this parameter to match Europe's real GDP (relative to that of the U.S.) in 1790.

To identify the parameters driving the scale of shipping costs,  $\phi^F$ ,  $\phi^N$ , and  $\phi^E$ , note first that Lemma 1 establishes a relationship between the growth rate of a location  $s$  and the size of  $s$ 's effective hinterland,  $\int_{\sigma_t^{-1}(s)} L_t^F(r) dr / L_t^N(s)$ . In a world in which effective hinterland size is uniform across space, all locations grow at the same rate and city growth exhibits no convergence. In another extreme case in which the same number of farmers sell at each location, growth rates only depend on non-farm population and therefore city growth exhibits strong convergence. As the scale of farm shipping costs  $\phi^F$  is a crucial parameter influencing effective hinterland size, it is intuitive to use cities' degree of convergence in the data to calibrate the value of this parameter. Matching the degree of convergence, that is, the coefficient in a regression of log city growth on log city size, between the model and the data implies setting  $\phi^F = 0.09$ .

To calibrate the scale of non-farm shipping costs  $\phi^N$ , note that lower non-farm costs imply a larger role of non-farm trade in the economy, hence a higher concentration of people in the largest clusters of non-farm population: cities. Therefore, I calibrate the value of  $\phi^N$  such that the model replicates the ratio of non-farm population living in cities to total U.S. population in 1790. This procedure leads to a value of  $\phi^N = 0.052$ . The fact that  $\phi^N / \phi^F$  is less than one is in line with the observation that transporting manufacturing goods was substantially cheaper during the pre-Civil War era than transporting agricultural products (Herrendorf *et al.*, 2012).

The scale parameter of international shipping costs,  $\phi^E$ , drives the importance of international trade in the model. As a result, I calibrate the value of  $\phi^E$  to match the U.S. export to GDP ratio of 9% in 1790 (Lipsey, 1994). This procedure yields  $\phi^E = 0.098$ .

Finally, I need to decide on the length of a time period. Given that railroad data come at a frequency of five years, I choose this as the length of one period. This implies that the simulated

population distribution of every second period can be compared to the decennial census data. It also implies that the model needs to be simulated for 15 consecutive periods, starting from 1790 and ending in 1860.

## 5. RESULTS

### 5.1. *Model fit*

To assess the quantitative performance of the theoretical framework, this section studies the evolution of the spatial distribution of economic activity predicted by the model and compares it to the one seen in the data. As the evolution of individual locations' population and sectoral specialization is not targeted in the calibration, this can be viewed as a test of the model.

Since the model is simulated over 20 by 20 arc minute grid cells, comparing simulation results to the evolution of actual population requires me to assign census data on county populations to these cells. Therefore, I distribute the population of each county across the grid cells it occupies based on the share of land belonging to each cell. In other words, I assume that population density was uniform within each county. Although it is unlikely that the distribution was exactly uniform, this assumption must lead to little bias since the most densely populated counties were small, usually fully contained in a single cell.

Column (1) of Table 5 shows that the correlation between cells' population levels predicted by the model and those in the data is high. Column (2) shows the same for population density (that is, population per unit of land). Even though I only match city and town cells' total population and the share of non-farm population living in cities, the correlation coefficient is already above 0.35 in 1790. As the dynamics of the model start playing out, correlation increases further to values around 0.7.<sup>37</sup> Moreover, columns (3) and (4) show that the correlation is high even if I calculate it between *changes* in population levels relative to the initial period, rather than between the levels of population or density in a given year. Thus, the model truly captures the movement of people to the same locations as in the data. In Figure 5, one can see that the location of regions with highest population density in the last period, such as the Northeast seashore and the areas around the Erie Canal and the Ohio River, coincides in the model and in the data. One can even spot the emergence of population clusters around Detroit, Minneapolis, and St. Louis in the model.

Besides featuring a high correlation with the data, the model can also replicate the first-order qualitative features of the evolution of U.S. population and economic activity. In line with the data, the model predicts a substantial westward expansion of population. By 1860, the population share of the Midwest increases to about 30% both in the model and in the data (Table 15 in [Supplementary Material, Appendix E](#)).

To assess performance of the model in replicating urbanization as well as the locations and sizes of cities, I define a city in the model as a grid cell with non-farm population above 10,000 inhabitants. In any time period, I measure urbanization as the ratio of non-farm population living in cities to total U.S. population, just like in the data. Figure 6 presents the evolution of urbanization in the model. The left panel of Figure 6 shows that the fraction of U.S. population living in cities started to increase rapidly after 1830 both in the model and in the data. In 1860, 14.8%

37. In the slow migration process presented in Section 2, the correlation is one in 1790 as the data are matched exactly in the initial period. However, the correlation drops quickly over time, and is already below 0.3 in 1860. I also simulate a process in which I extrapolate the 1790–1820 population growth of every cell to the period between 1820 and 1860. In this latter case, the correlation is below 0.35 by 1840 and below 0.05 by 1860. That is, these simple benchmark processes do poorly in replicating the population movements seen in the data.

TABLE 5  
*Correlation of population between the model and the data*

<i>t</i>	Correlation, period <i>t</i>		Correlation, change from 1790 to <i>t</i>	
	Population (1)	Population density (2)	Population (3)	Population density (4)
1790	0.379	0.366	—	—
1800	0.439	0.441	0.383	0.363
1810	0.538	0.539	0.597	0.563
1820	0.582	0.593	0.626	0.611
1830	0.680	0.669	0.741	0.708
1840	0.684	0.597	0.657	0.546
1850	0.706	0.619	0.662	0.556
1860	0.735	0.644	0.701	0.606

*Notes:* Column (1) of this table shows the correlation between grid cells' population in the data and their population in the baseline simulation of the model, by census year, from 1790 until 1860. Column (2) shows the same between population per unit of land instead of population. Column (3) shows the correlation between the change in population from 1790 until the given year in the data and the same object in the model. Column (4) shows the same between population per unit of land instead of population. Source: U.S. census.

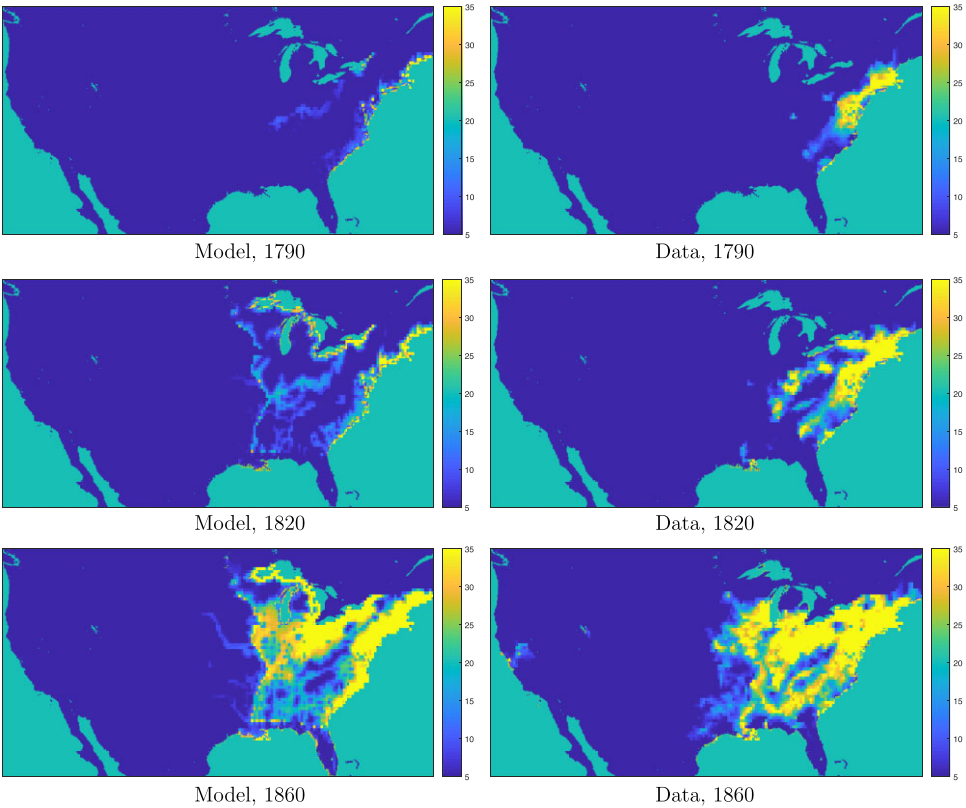


FIGURE 5  
Population distribution: model versus data

*Notes:* The three left panels of the figure present population (number of inhabitants) per square mile in each 20 by 20 arc minute grid cell of the U.S. in the baseline model simulation in 1790, 1820, and 1860, respectively. The three right panels present the same in the data. The scale is bottom-coded at 5 and top-coded at thirty-five for visibility. Source: U.S. census.

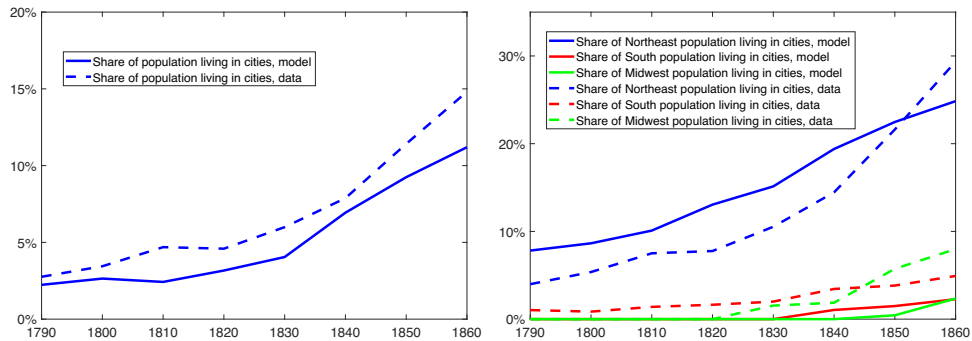


FIGURE 6  
Urbanization in the model

*Notes:* In the left graph, the solid blue line shows the ratio of non-farm population living in cities (cells with more than 10,000 non-farm inhabitants) to total U.S. population in the baseline model simulation for each census year. The dotted blue line shows the fraction of U.S. population living in cities (settlements above 10,000 inhabitants) in the data for each census year. In the right graph, the solid blue, red, and green lines show the ratio of non-farm population living in cities to the total population of the region in the three large U.S. regions: the Northeast, the South, and the Midwest, respectively, in the baseline model simulation. The dotted blue, red, and green lines show the fraction of population living in cities (settlements above 10,000 inhabitants) in the Northeast, South, and Midwest, respectively, in the data. Source: U.S. census.

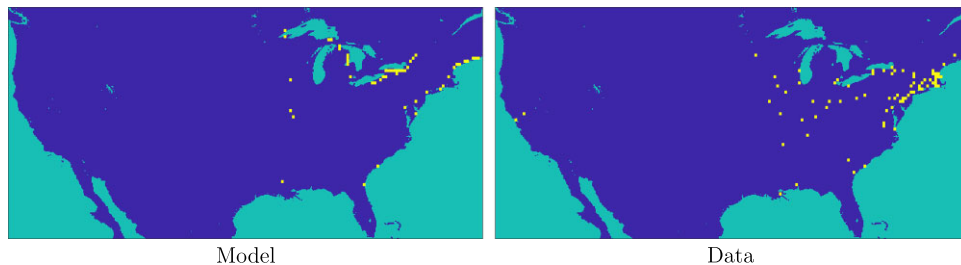


FIGURE 7  
City locations by 1860: model versus data

*Notes:* In the left map, yellow 20 by 20 arc minute grid cells are those that have a city in them in 1860 in the baseline model simulation. In the right map, yellow grid cells are those that have a city in them forming by 1860 in the data. Source: U.S. census.

of U.S. population lived in cities in the data, while the corresponding number is 11.2% in the model. The right panel of Figure 6 shows that urbanization is mainly due to the Northeast, where the share of people living in cities in 1860 was more than ten times as high as the U.S. urbanization rate in 1790. In the model, this is largely due to the productivity advantage of pre-existing Northeast cities, which allows them to remain strong over the whole period of investigation.

Figure 7 presents the set of cells in which cities form by 1860 in the model, and contrasts it with the data. As the location of cities other than the five pre-existing ones is completely untargeted in the calibration, it is natural that the model does not match city locations exactly. Nonetheless, it is reassuring that both the number of cells that contain cities (fifty in the model and seventy-five in the data) and the geographic location pattern of cities look similar. In particular, cities are not all clustered together but rather spread out across the Northeast, the Midwest, and the South, though the model stops short at predicting cities in the central parts of the Midwest and the South.

In the model, the fact that cities are spread out in space is tightly linked to the key role that farm hinterlands play in city formation. Close to cities, workers have an incentive to specialize in farm activities and ship their product to the city rather than create another concentration of



TABLE 6  
Clustering of cities: model versus data

	Model	Data
Average distance from cities in the region (miles)		
– Northeast	443.6	315.8
– Rest of the U.S.	810.0	1057.4
Average distance from the closest city in the region (miles)		
– Northeast	50.2	50.2
– Rest of the U.S.	148.2	142.7

*Notes:* The first two numbers in the “Model” column correspond to the unweighted average of pairwise distances between 1860 cities in the Northeast and in the rest of the U.S., respectively, in the baseline model simulation. The last two numbers of the “Model” column correspond to the unweighted average of an 1860 city’s distance from the closest 1860 city in the Northeast and in the rest of the U.S., respectively, in the baseline model simulation. The “Data” column reports the corresponding numbers in the data (also reported in Table 3). Source: U.S. census.

TABLE 7  
Fraction of cities forming at trading routes: model versus data

	Fraction of cells	Fraction of land	Fraction of cities	
			Model	Data
Water	39.1%	35.6%	100.0%	98.6%
Confluence	23.0%	19.9%	91.1%	85.5%
Railroad	24.3%	23.2%	71.1%	95.7%

*Notes:* The first column of this table shows the number of 20 by 20 arc minute grid cells located at water, a confluence, or a railroad relative to the total number of cells in the U.S. The second column shows the fraction of U.S. land that belongs to cells located at water, a confluence, or a railroad. The third column shows the fraction of 1860 U.S. cities that formed in cells located at water, at a confluence, or at a railroad in the baseline model simulation. The last column shows the corresponding numbers in the data (also reported in Table 1). A cell is located at water if the cell, or the cell next to it, has a navigable river, canal, lake, or the sea in 1860. It is located at a confluence if the cell, or a cell next to it, is surrounded by at least three cells with water in 1860. It is located at a railroad if the cell, or a cell next to it, was part of the railroad network in 1860. Source: U.S. census, ESRI Map of U.S. Major Waters, and *oldrailhistory.com*.

non-farm activity. However, one would expect this effect to be somewhat muted in the Northeast, where spatial technology diffusion from pre-existing high-productivity cities gives an advantage to pursuing non-farm activities nearby. Thus, everything else fixed, one would expect cities to cluster more in the Northeast than in the rest of the U.S. Section 2 shows that this is true in the data (fact 4), while Table 6 verifies it in the model: average distance from other cities and average distance from the closest neighbour were both smaller in the Northeast, both qualitatively and quantitatively similar to the data.

Just like in the data, most cities form near trading routes in the model (Table 7). In particular, all cities form at water (98.6% in the data), 91.1% form at confluences (85.5% in the data), and 71.1% at railroads (95.7% in the data), all substantially higher than the fraction of cells located at these types of trading routes. In the model, the disproportionate emergence of cities near trading routes is due to trade playing an important role in city formation. Cities formed at trading routes to benefit from these locations’ better access to other locations.

The model is also quite successful at predicting the history of the largest U.S. cities. In 1860, Boston is the largest city in the model (762,100 inhabitants), while it was the third-largest in the data (279,200 inhabitants). New York was the largest city in the data (slightly above 1

million inhabitants) and the second-largest in the model (569,000 inhabitants).<sup>38</sup> Looking at the other three cities already existing in 1790, Baltimore and Philadelphia both exceed 200,000 inhabitants, making them the third- and fourth-largest cities in 1860 both in the model and in the data. Charleston, on the other hand, remains below 150,000 both in the data and in the model simulation.

Unfortunately, the lack of high-resolution 19th-century spatial data prevents me from testing the model on most variables other than population, city locations, and city sizes. One aggregate variable that can be compared between the model and the data is exports. Although the 1790 U.S. export to GDP ratio is targeted in the calibration, the evolution of this variable after 1790 is not. Figure 17 (Supplementary Material, Appendix E) presents the evolution of U.S. exports to GDP in the model, and contrasts them with historical estimates from Lipsey (1994). One can see that the model is able to predict the U-shaped pattern in this variable: exports as a fraction of GDP first declined, then increased over the course of the 19th century. It is true, however, that the model overpredicts the increase in exports to GDP during the second half of the period. In section “The effect of international trade”, I discuss potential consequences of this.

How surprising is it that the model of Section 3 features a good fit to the dynamic evolution of the U.S. population distribution and to the locations and sizes of cities? This question cannot be answered without a benchmark to which the model’s fit is compared. One natural benchmark is the model developed in Desmet *et al.* (2018) (“DNR”). DNR is also a dynamic quantitative model of economic geography, and therefore can be used to predict the evolution of the spatial distribution of population. Unlike the model of Section 3, DNR features moving costs across locations, which, if anything, should make it better suited to matching the slow migration of population toward the initially low-density West.<sup>39</sup> Unlike the model of Section 3, however, DNR features one sector. As a result, besides serving as a benchmark, a comparison with the DNR model can also clarify whether a two-sector framework consistent with the hinterland hypothesis is crucial to quantitatively replicate the data.

In Supplementary Material, Appendix D, I develop a version of the DNR model that I can take to the data and compare to the performance of the model of Section 3. I conclude that the DNR model is unable to replicate the first-order features of the data—specifically, the westward expansion and the emergence of cities outside the Northeast. This underscores that a two-sector model consistent with the hinterland hypothesis is necessary to explain the key patterns of the evolution of the U.S. population distribution and city formation.

## 5.2. Counterfactual: the effect of railroads

This section evaluates the first implication of the hinterland hypothesis: the prediction that railroads fostered city development and aggregate U.S. growth. How much railroads contributed to American economic growth and output has been the subject of long debates in economic history. In his book, Fogel (1964) attempts to account for the various local and aggregate effects of late-nineteenth century railroads, and finds that the overall impact of railroads on agricultural

38. Glaeser (2011) argues that the leading position of New York over other major ports in the Northeast depended on its deep and protected harbour that could accommodate big clipper ships. As this force is outside my model, the fact that New York only becomes the second-largest city in the model may suggest that it could not have become the largest city without this advantage.

39. It needs to be noted, however, that DNR restricts the functional form of moving costs such that they do not have a bilateral, distance-related component. Supplementary Material, Appendix C presents another benchmark model in which I allow for distance-dependent moving costs and show that this benchmark model is also unable to replicate several important features of the data.

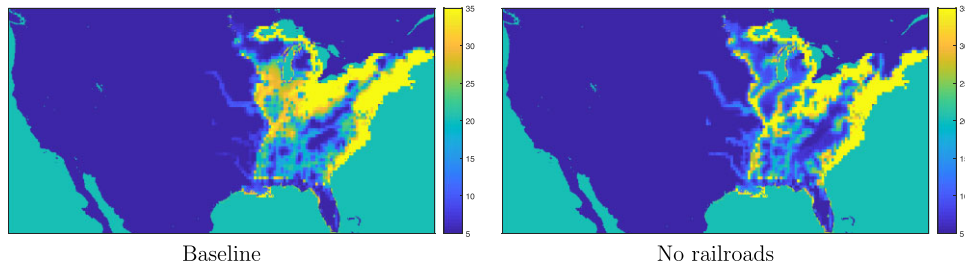


FIGURE 8

Population distribution in 1860: baseline versus no railroads

*Notes:* The left map presents population (number of inhabitants) per square mile in each 20 by 20 arc minute grid cell of the U.S. in the baseline model simulation in 1860. The right map presents the same in the counterfactual with no railroads. The scale is bottom-coded at five and top-coded at thirty-five for visibility.

output was rather modest. [Donaldson and Hornbeck \(2016\)](#) revisit the topic through the lens of market access. They argue that railroads decreased land shipping costs dramatically. Hence, they made it possible to source agricultural products from locations from which shipping was prohibitively expensive earlier. This led to an increase in market access and a decrease in agricultural prices at most U.S. locations, but especially at those near railroads. As a result, welfare and output increased, the latter being 3.22% higher in 1890 in the presence of railroads than in their absence. Relative to [Donaldson and Hornbeck \(2016\)](#), my model also allows me to account for the impact of railroads on U.S. output and welfare through railroads changing the dynamic evolution of cities.

To assess the impact of railroads on city formation, growth, and output, I simulate the model under the assumption that no railroads were built during my period of investigation. The absence of railroads directly increases shipping costs in 1.0% of grid cells in 1840, in 3.0% of the cells in 1850, and in 8.2% of the cells in 1860. As I argue, however, the absence of railroads affects a much larger share of U.S. locations indirectly through its effects on trade and the formation of cities.

The results indicate that the U.S. economy would look fairly different in the absence of railroads. Figure 8 plots the 1860 distribution of population in the no-railroad counterfactual, as well as in the baseline calibration. As can be seen from the figure, the lack of railroads makes population shift toward navigable rivers, whereas regions far from water remain lower-density, especially in the Midwest. Many of the large cities remain substantially smaller without railroads. The grid cell in which Philadelphia is located loses 26% of its population relative to the baseline, while Chicago loses 12%, New York loses 17%, and Boston loses as much as 62% of its population in 1860.

Without railroads, the total number of people living in cities would be 1.8% lower. These results show that the effect of railroads on city growth, through railroads expanding cities' farm hinterland, is quantitatively large. In the absence of railroads, hinterlands are smaller, hence non-farm firms' incentives to innovate are lower. This implies slower city growth and less urbanization.

The substantial impact that railroads have on the development of cities translates to a large effect on output and consumers' welfare. Figure 18 (Supplementary Material, Appendix E) shows the evolution of U.S. real GDP per capita, which also equals U.S. consumers' per-period utility in the model. The absence of railroads decreases the growth rate of the U.S. economy between 1830 and 1860 by 25%, from 0.71% to 0.54% per year. As a result, real GDP is 5.1%

lower in 1860 than in the presence of railroads. Even if the growth rates in the two scenarios were equal after 1860, this would imply a permanent loss of 5.1% in consumers' long-run welfare.

A substantial fraction of these gains arises due to railroads changing the incentives to innovate in the non-farm sector and, hence, the formation and development of cities. To show this, I consider an alternative version of the model of Section 3 in which productivity growth in the non-farm sector is uniform across space:

$$A_{t+1}(s) = \tilde{f} A_t(s)$$

for any U.S. location  $s$  and time period  $t$ . In the model of Section 3, railroads increase the incentives to innovate at locations whose hinterlands expand, implying faster non-farm productivity growth at these locations. In the model with uniform productivity growth, this channel is absent as non-farm productivity growth is the same across locations.

To make the model with uniform productivity growth comparable to the model of Section 3, I calibrate the value of productivity growth parameter  $\tilde{f}$  to match the growth rate of U.S. real GDP before 1820. Recall that this moment was also targeted in the baseline model calibration. Hence, the model with uniform productivity growth features the same growth rate as the baseline prior to railroad construction.

Next, I simulate the model with uniform productivity growth both in the presence and in the absence of railroads. I find that the absence of railroads leads to a 3.4% loss in real GDP in 1860 in the model with uniform productivity growth. Thus, without differences in growth rates across space, arising from differences in innovation and dynamic externalities in cities, the effect of railroads on output would decrease from 5.1% to 3.4%, that is, by 34%. It is, however, important to note that cities still form and develop in the model with uniform productivity growth, due to population growth and consumers relocating to places with better access to trade, and consumers moving to newly forming cities benefit from agglomeration. In other words, the 34% number should be viewed as a lower bound on the contribution of city formation to the effect of railroads on U.S. output. In line with the first implication of the hinterland hypothesis, city formation and development seems to be an influential factor driving the impact of railroads on economic growth, output, and welfare.

### 5.3. *Further counterfactuals*

The remaining two implications of the hinterland hypothesis have to do with how other elements of U.S. geography—political border changes and international trade—shaped U.S. growth and city formation. To evaluate each of these two implications, I conduct two further counterfactuals in this section.

**5.3.1. The effect of political border changes.** In 1790, the political border of the U.S. mostly coincided with the Mississippi River in the West and with the current Georgia–Florida border in the South. Areas farther West and South were under French, Spanish, or British rule. The total land area of the U.S. was 864,746 square miles. By 1860, the U.S. reached the Pacific coast in the West, covering 2,969,640 square miles, almost equal to its current contiguous area. This means that the land area of the country increased by a staggering 211.8%. According to the second implication of the hinterland hypothesis, this process could have led to faster city development and aggregate growth through increasing the size of cities' potential hinterlands.

To see how much the expansion of U.S. political borders changed the process of city formation and development prior to the Civil War, I conduct a counterfactual in which I keep U.S. land area at its initial level. That is, I assume that grid cells that were not part of the U.S. in 1790

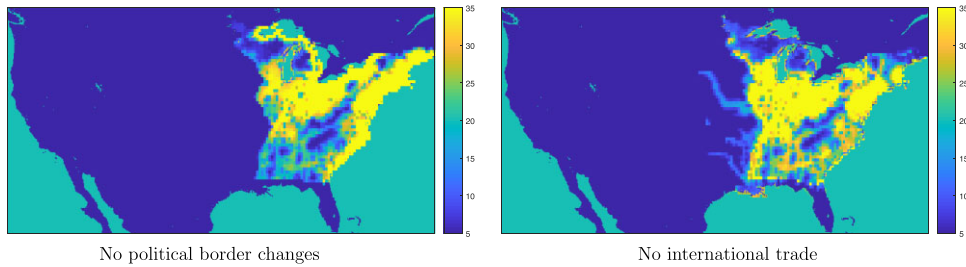


FIGURE 9

Further counterfactuals: population distribution in 1860

*Notes:* The left map presents population (number of inhabitants) per square mile in each 20 by 20 arc minute grid cell of the U.S. in the counterfactual with no political border changes (section “The effect of political border changes”); the right map presents the same in the counterfactual with no international trade (section “The effect of international trade”). The scale is bottom-coded at five and top-coded at thirty-five for visibility.

never became part of the country. I compare the results of this counterfactual to the baseline calibration, in which I change U.S. political borders as in the data.

The left map of Figure 9 shows the 1860 population distribution without political border changes. The differences relative to the baseline (left map of Figure 8) are surprisingly small. Some clusters along the Mississippi and in northern Florida do not get to form without changing borders, but the population distribution in the rest of the country hardly changes. Overall, real GDP per capita is 2.6% lower in 1860 without changes in political borders. This number may not be small in absolute terms, but it is certainly small relative to the 211.8% increase in land area that the border changes led to.

The reason for the small effect of political border changes is that the region West of the Mississippi had few navigable rivers and few railroads in 1860. The first transcontinental railroad crossing the region was built in 1869, opening up the land to the settlement of population. With the transport infrastructure in place before 1860, the region could hardly serve as a farm hinterland. At the same time, given all the transport improvements after the Civil War, pre-1860 political border changes might have brought substantial gains *after* 1860—but not during my period of investigation.

**5.3.2. The effect of international trade.** My final counterfactual investigates the third implication of the hinterland hypothesis by quantifying the effect of international trade on U.S. city formation and growth. The U.S. was a relatively export-oriented country in the first half of the nineteenth century. For instance, the share of U.S. trade in world trade was well above the share of U.S. output in world output (Lipsey, 1994). The U.S. only ceased to be an open economy for a few months starting from December 1807, when Congress imposed an almost complete embargo on foreign trade at the request of President Thomas Jefferson (Irwin, 2005). The embargo was lifted in March 1809.

To study the overall impact of international trade, I simulate the model in an extreme counterfactual scenario. In particular, I make the assumption that President Jefferson’s trade embargo remained in place forever, implying that the U.S. was in autarky from period 5 (1810) until period 15 (1860). To implement this counterfactual, I raise the cost of trading with Europe for every U.S. location to infinity during this entire period.

The right map of Figure 9 shows the 1860 population distribution under autarky. Comparing it to the baseline (left map of Figure 8), one can conclude that international trade was a factor mitigating westward expansion. Under autarky, the spatial distribution of population shifts from regions close to the Atlantic coast toward regions farther West. Total population of the Midwest,

for instance, is 17.3% higher under autarky. The intuition for these results is straightforward. As the Eastern regions of the U.S. were geographically closer to Europe, they could reap the benefits from trading with Europe.

The lack of foreign trade hurts smaller cities in the Northeast particularly severely. In 1860, the cell in which Providence, RI, is located loses 70% of its population, while Syracuse, NY, loses as much as 77% of its population relative to the baseline. Boston is among the biggest losers as well: its population decreases by 70%. Somewhat surprisingly at first glance, some other large cities in the Northeast gain in population. Baltimore and Philadelphia increase their populations by 76% and 30%, respectively. However, this is not surprising in light of the quantitative economic geography literature, which has found that losses in market access tend to hurt small cities more, and population might reallocate toward larger cities as a result (Redding and Sturm, 2008). Nevertheless, small cities' losses seem to dominate in the aggregate. In 1860, the urbanization rate decreases from 11.2% under international trade to 6.1% under autarky according to the model.

These losses in urbanization translate to substantial losses in growth. Between the period right before the embargo (1805) and 1860, the growth rate of U.S. real GDP per capita is 19% lower under autarky. Real GDP in 1860 is 5.8% lower. Based on these findings, one can conclude that international trade slowed down westward expansion but fostered the development of cities in the Northeast, in line with the third implication of the hinterland hypothesis. Due to its favourable effect on Northeast cities, international trade significantly increased aggregate growth in the pre-Civil War U.S.

One should take the quantitative findings of section "The effect of international trade" with a grain of salt, however. As discussed in Section 5.1, the model overestimates the share of international trade in U.S. GDP throughout most of the period of investigation. It is therefore likely that the model overstates the role of trade relative to the data. Although there is no reason to believe that the model is wrong about the effects of trade in a qualitative sense, the numbers found above should be interpreted as upper bounds on the effects of trade on the pre-Civil War U.S. economy.

## 6. ROBUSTNESS

### 6.1. *Technology diffusion and dynamic externalities*

To study whether the assumptions of spatial technology diffusion and dynamic externalities in cities have a significant impact on my quantitative results, I conduct three robustness exercises in this section. In one exercise, I strengthen spatial technology diffusion by decreasing the diffusion parameter  $\delta$  from its baseline value, 0.006, to 0.003. In a second exercise, I weaken technology diffusion by increasing  $\delta$  to 0.009. In the third exercise, I eliminate dynamic externalities in cities by setting their scale parameter  $\gamma$  to zero.

For each of the three exercises, I simulate the model in all the periods between 1790 and 1860, and measure the fit of the model to the same untargeted moments of the population distribution and city formation as in the baseline calibration. Finally, I conduct the three counterfactuals described in Sections 5.2 and 5.3: one without railroad construction; one without political border changes; and one in which President Jefferson's trade embargo lasts all the way until 1860.

Table 8 confirms the robustness of my findings to the alternative values of technology diffusion and dynamic externality parameters. Each of the last three columns in the table corresponds to a robustness exercise, while the first column repeats the results obtained under the baseline calibration. The correlation of cell populations between the model and the data, the population



TABLE 8  
*Robustness: technology diffusion and dynamic externalities*

	Baseline	$\delta = 0.003$	$\delta = 0.009$	$\gamma = 0$
Correlation of population with the data, 1860	0.735	0.564	0.735	0.639
Regions' shares in total U.S. population, 1860				
– Northeast	41.7%	41.6%	41.7%	35.8%
– South	26.4%	26.4%	26.4%	28.8%
– Midwest	31.1%	31.2%	31.1%	34.5%
– West	0.8%	0.9%	0.8%	0.9%
Fraction of cities forming at				
– Water	100.0%	100.0%	100.0%	100.0%
– Confluence	91.1%	89.7%	91.1%	93.5%
– Railroad	71.1%	77.6%	71.1%	71.0%
Average distance between cities (miles), 1860				
– Northeast	443.6	423.2	443.6	421.1
– Rest of the U.S.	810.0	895.3	810.0	763.5
Difference in average log decennial growth rate between cities and towns, 1790–1820	0.08	–0.18	0.06	–0.05
Counterfactuals: change in U.S. real GDP per capita relative to the baseline, 1860				
– No railroads	–5.1%	–4.5%	–5.1%	–5.2%
– No political border changes	–2.6%	–2.6%	–2.6%	–2.6%
– No international trade	–5.8%	–6.4%	–5.8%	–5.8%

*Notes:* The “Baseline” column presents a set of statistics obtained in the baseline model simulation: the correlation between grid cells’ population in the model and in the data in 1860; the share of U.S. population living in either of the four large regions; the fraction of 1860 cities that formed at water, at a confluence, or at a railroad; the unweighted average of pairwise distances between 1860 cities in the Northeast and in the rest of the U.S., respectively; the difference in average log decennial population growth rate between cities and towns over the period 1790–1820; and the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.2 and 5.3) relative to the baseline. Columns  $\delta = 0.003$  and  $\delta = 0.009$  present the same statistics in two alternative model simulations in which the technology diffusion parameter  $\delta$  is set to 0.003 and 0.009, respectively. Column  $\gamma = 0$  presents the same statistics in an alternative model simulation in which the dynamic externality parameter  $\gamma$  is set to zero.

shares of the three large regions, the fraction of cities forming at trading routes and the clustering of cities, measured by the average distance between them, all look very similar to the baseline. Intuitively, the difference in cross-city average distance between the Northeast and the rest of the U.S. is larger if technology diffusion is strong, as more cities form closer to high-productivity cities that spread out their technology in the Northeast.<sup>40</sup> Conversely, the difference between the Northeast and the rest of the country is smaller if technology diffusion is weak. Also, the difference between cities’ and towns’ growth rates, which is the moment I targeted when choosing  $\gamma$  in the baseline, is no longer matched if dynamic externalities are absent in cities ( $\gamma = 0$ ). In fact, cities grow slower than towns in this case, due to the convergence in growth rates.

The last four rows of Table 8 present the counterfactual results in the three robustness exercises, and compare them to the baseline. One can see that the counterfactual changes in U.S. real GDP per capita are almost identical between the baseline and the three robustness exercises.

40. In this exercise ( $\delta = 0.003$ ), the number of cities forming in the Northeast by 1860 increases by one-third, from 15 to 20.

## 6.2. *Endogenous railroads*

Clearly, the placement of railroads is not exogenous. As discussed in Section 4, this cannot bias the identification of the model's structural parameters, since I calibrate these parameters exclusively to moments before railroad construction begins in the 1820s. It cannot bias the results of the counterfactual conducted in Section 5.2 either, as that counterfactual compares the U.S. economy with the actual set of railroads to an economy with no railroads whatsoever.

This argument, however, does not apply to the counterfactuals presented in Section 5.3. Take section "The effect of political border changes" as an example. In that section, I compare the U.S. economy with the observed changes in political borders to a counterfactual economy in which political borders did not change. Both in the baseline and in the counterfactual, I assume that the actual set of railroads were built. But is this a realistic assumption in the counterfactual? One could think that railroad placement, if it is endogenous, must have been affected by the location of U.S. borders. Hence, the counterfactual with no political border changes should feature an endogenous counterfactual set of railroads, not the actual ones.

To see whether the results of Section 5.3 are robust to this issue, I develop a procedure of endogenous railroad placement. In various models of endogenous transport infrastructure investment (Fajgelbaum and Schaal, 2020; Allen and Arkolakis, 2021; Santamaria, 2022), the optimal amount of investment allocated to a location is an increasing function of how much the location trades with others. This is intuitive: a larger volume of trade implies that there is more demand for infrastructure investment at the location.<sup>41</sup>

Motivated by this relationship between trade and optimal infrastructure investment, I consider the following extension of the model of Section 3. Before 1830, I assume that the technology to build railroads was not available, just like in the baseline model. In 1830, I first simulate the model without any railroad. I select the grid cell that has largest volume of trade with others from the set of cells without a navigable waterway. I assign a railroad to this cell, then recalculate all shipping costs, and determine the volume of trade for every cell.<sup>42</sup> I again select the cell with the largest volume of trade from the set of cells without a railroad or a waterway. I continue this procedure until the number of railroad cells equals the number of cells that received an actual railroad in 1830 (excluding the cells that had both railroad and water access in the data). Taking past railroad investments as given, I repeat the procedure for the next period, then for the next period, and so on.

The left map of Figure 10 displays the set of grid cells to which my railroad placement procedure assigns an endogenous railroad by 1860, while the right map of Figure 10 shows the actual set of railroad cells (excluding those with both railroad and water access) in the data. Even though my railroad placement procedure is undoubtedly a stylized one, it can successfully predict a dense railroad network in the Eastern regions of the U.S. That said, the procedure somewhat underpredicts the set of railroads built in the South. According to the model, these areas do not have enough trading activity to justify the construction of railroads. Nonetheless, the stylized railroad placement framework can correctly predict 51.4% of the grid cells in which railroads get constructed by 1860, despite the fact that these cells only constitute 4.2% of the total number of cells.

41. The other major factor affecting the allocation of investment in these models is the location-specific cost of building infrastructure. In my framework, I abstract from these costs, which is identical to assuming that they do not vary across locations. Although the cost of building railroads undoubtedly varies across space, I find that my framework can predict the actual location of railroads well even if I do not take these costs into account.

42. I select from the set of cells without a waterway since the cost of railroad transportation was slightly higher than the cost of water transportation. Hence, there would be no incentive to build a railroad in a cell that already has water access.

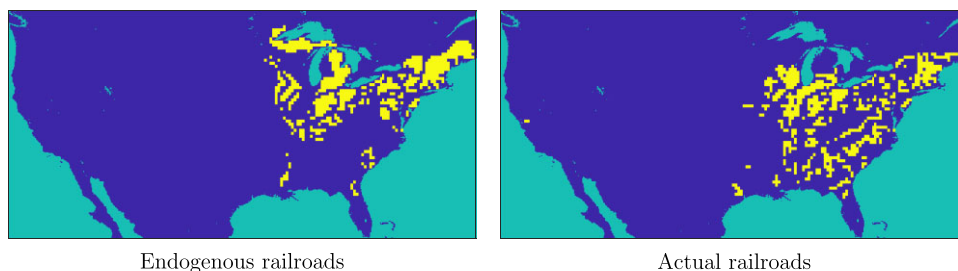


FIGURE 10

Location of endogenous and actual railroads in 1860

Notes: In the left map, yellow 20 by 20 arc minute grid cells are those into which the model with endogenous railroads places a railroad by 1860. In the right map, yellow grid cells are those that have an actual railroad in 1860. Source: *oldrailhistory.com*.

TABLE 9

*Counterfactual results under endogenous railroads*

Counterfactual	Actual railroads	Endogenous railroads
No political border changes	−2.6%	−2.5%
No international trade	−5.8%	−5.7%
No railroads	−5.1%	−5.7%

Notes: The first column of this table shows the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.2 and 5.3) relative to the baseline model simulation. The second column shows the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.2 and 5.3) relative to a new baseline model simulation such that railroads in both the counterfactuals and the new baseline are placed endogenously.

To study how robust the counterfactual results of Section 5.3 are to endogenous railroad placement, I conduct each of the two counterfactuals in the extended model with endogenous railroads. In other words, I allow railroad placement to react to political border changes or international trade. The only variable I keep fixed between the counterfactuals and the baseline is the total number of cells that receive a railroad in a given period. The first two rows of Table 9 present the changes in U.S. real GDP per capita in each of the counterfactuals, and contrast them with the corresponding real GDP changes found without endogenous railroads. The differences are small, suggesting that the effects of political border changes and international trade are robust to incorporating the endogeneity of railroad placement.

In the last row of Table 9, I compare the equilibrium with endogenous railroads to the no-railroad counterfactual. Not surprisingly, endogenous railroads lead to a somewhat larger increase in U.S. real GDP than actual railroads, as they are built where demand for them is the highest according to the model. The difference, however, is not large. Endogenous railroads induce a 5.7% increase in 1860 U.S. real GDP, while the actual set of railroads increases real GDP by 5.1%. This result is in line with the fact that the placement of endogenous railroads quite closely follows the placement of actual railroads.

## 7. CONCLUSION

In this paper, I argue that farm hinterlands played an important role in U.S. city formation and aggregate growth prior to the Civil War. The relationship I identify between hinterlands, city formation, and endogenous growth constitutes a novel mechanism through which geography can affect growth in economies at an early stage of development. This immediately implies that studying the drivers of pre-Civil War U.S. growth is not the only possible application of the quantitative framework proposed in the paper. Studying today's developing economies is one

particularly fruitful direction for further research. China and India have seen massive changes in their geography due to recent large-scale transport infrastructure improvements (Faber, 2014; Ghani *et al.*, 2016). How did these transport improvements affect the formation and growth of cities? How much does their potential contribution to city development add to the returns to these infrastructure improvements? What would be the effect of proposed infrastructure projects, or the effect of further economic integration between these countries and their foreign markets?

Another question is how the mechanism identified in the paper may interact with institutions, which have been viewed by many economists as a crucial determinant of development (North, 2003). The quantitative analysis of the paper is conducted under the assumption that market institutions were available at every location in the U.S. and over the entire period of investigation. This assumption might be realistic in a U.S. context, but not necessarily in the case of other economies. Institutions may indeed develop endogenously over time, similar to local innovations. Can institutions, and their endogenous evolution, shape the effect of geography on cities and growth? My hope is that the setup proposed in this paper opens a door to research that can answer these, as well as other interesting questions.

*Acknowledgments.* The earlier title of this paper is “City location and economic development”. I am grateful to Esteban Rossi-Hansberg for invaluable guidance and support. I am also grateful to Judit Badics, Klaus Desmet, Cécile Gaubert, Ed Glaeser, Gene Grossman, Oleg Itskhoki, Réka Juhász, Miklós Koren, Jan de Loecker, Ildikó Magyari, Robert Margo, Thierry Mayer, Eduardo Morales, Fernando Parro, Luigi Pascali, Giacomo Ponzetto, Stephen Redding, Jacques Thisse, Felix Tintelnot, Áron Tóbiás, Sharon Traiberman, Jaume Ventura, four anonymous referees, as well as seminar and conference participants for helpful comments and suggestions. I thank Adrien Bilal and Charly Porcher for help with optimizing the simulation of the model. I acknowledge financial support from the International Economics Section at Princeton University, as well as from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563 and CEX2019-000915-S), through the Juan de la Cierva Grant (FJCI-2017-34728), the Spanish State Research Agency and the European Social Fund (RYC2019-027620-I/AEI/10.13039/501100011033), and from the Generalitat de Catalunya, through CERCA and SGR Programme (2017-SGR-1393). All errors are my own.

### Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

### Data Availability Statement

The data and programmes underlying this article are available on Zenodo at <https://doi.org/10.5281/zenodo.7419812>.

## REFERENCES

- ALLEN, R. (2000), “Economic Structure and Agricultural Productivity in Europe, 1300–1800”, *European Review of Economic History*, **3**, 1–25.
- ALLEN, T. and ARKOLAKIS, C. (2014), “Trade and the Topography of the Spatial Economy”, *Quarterly Journal of Economics*, **129**, 1085–1140.
- ALLEN, T. and ARKOLAKIS, C. (2021), “The Welfare Effects of Transportation Infrastructure Improvements” (Mimeo).
- ANDERSON, O. E. (1953), *Refrigeration in America: A History of a New Technology and Its Impact* (Princeton, New Jersey: Princeton University Press).
- ATAK, J., BATEMAN, F., HAINES, M., *et al.* (2010), “Did Railroads Induce or Follow Economic Growth? Urbanization and Population Growth in the American Midwest, 1850–1860”, *Social Science History*, **34**, 171–197.
- BAIROCH, P. and GOERTZ, G. (1985), “Factors of Urbanization in the Nineteenth Century Developed Countries: A Descriptive and Econometric Analysis”, *Urban Studies*, **23**, 285–305.
- BAUMOL, W. (1967), “Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis”, *American Economic Review*, **57**, 415–426.
- BERTINELLI, L. and BLACK, D. (2004), “Urbanization and Growth”, *Journal of Urban Economics*, **56**, 80–96.
- BLACK, D. and HENDERSON, V. (1999), “A Theory of Urban Growth”, *Journal of Political Economy*, **107**, 252–284.
- BUERA, F. and KABOSKI, J. (2012a), “The Rise of the Service Economy”, *American Economic Review*, **102**, 2540–2569.
- BUERA, F. and KABOSKI, J. (2012b), “Scale and Origins of Structural Change”, *Journal of Economic Theory*, **147**, 684–712.

- CALIENDO, L., DVORKIN, M. and PARRO, F. (2019), "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock", *Econometrica*, **87**, 741–835.
- CALIENDO, L., PARRO, F., ROSSI-HANSBERG, E., *et al.* (2018), "The Impact of Regional and Sectoral Productivity Changes on the U.S. Economy", *Review of Economic Studies*, **85**, 2042–2096.
- CASELLI, F. and COLEMAN, J. (2001), "The U.S. Structural Transformation and Regional Convergence: A Reinterpretation", *Journal of Political Economy*, **109**, 584–616.
- CAZIER, L. (1976), *Surveys and Surveyors of the Public Domain, 1785–1975* (Ann Arbor, Michigan: University of Michigan Library).
- COMIN, D., DMITRIEV, M. and ROSSI-HANSBERG, E. (2013), "The Spatial Diffusion of Technology" (Mimeo).
- COSTINOT, A., DONALDSON, D. and SMITH, C. (2016), "Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 Million Fields Around the World", *Journal of Political Economy*, **124**, 205–248.
- CRONON, W. (1991), *Nature's Metropolis: Chicago and the Great West* (New York: W. W. Norton & Company).
- DAVID, P. (1969), "Transport Innovation and Economic Growth: Professor Fogel On and Off the Rails", *The Economic History Review*, **22**, 506–525.
- DAVID, P. (2005), "Real Income and Economic Welfare Growth in the Early Republic" (Working Paper, Stanford University).
- DESMET, K., NAGY, D. and ROSSI-HANSBERG, E. (2018), "The Geography of Development", *Journal of Political Economy*, **126**, 903–983.
- DESMET, K. and RAPPAPORT, J. (2017), "The Settlement of the United States, 1800–2000: The Long Transition to Gibrat's Law", *Journal of Urban Economics*, **98**, 50–68.
- DESMET, K. and ROSSI-HANSBERG, E. (2014), "Spatial Development", *American Economic Review*, **104**, 1211–1243.
- DINGEL, J., MENG, K. and HSIANG, S. (2021), "Spatial Correlation, Trade, and Inequality: Evidence from the Global Climate" (Mimeo).
- DIXIT, A. and STIGLITZ, J. (1977), "Monopolistic Competition and Optimum Product Diversity", *American Economic Review*, **67**, 297–308.
- DONALDSON, D. and HORNBECK, R. (2016), "Railroads and American Economic Growth: A 'Market Access' Approach", *Quarterly Journal of Economics*, **131**, 799–858.
- EATON, J. and ECKSTEIN, Z. (1997), "Cities and Growth: Theory and Evidence from France and Japan", *Regional Science and Urban Economics*, **27**, 443–474.
- ECKERT, F. and PETERS, M. (2018), "Spatial Structural Change" (Mimeo).
- EECKHOUT, J. (2004), "Gibrat's Law for (All) Cities", *American Economic Review*, **94**, 1429–1451.
- FABER, B. (2014), "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System", *Review of Economic Studies*, **81**, 1046–1070.
- FAJGELBAUM, P. and REDDING, S. (2022), "Trade, Structural Transformation and Development: Evidence from Argentina 1869–1914", *Journal of Political Economy*, **130**, 1249–1318.
- FAJGELBAUM, P. and SCHAAL, E. (2020), "Optimal Transport Networks in Spatial Equilibrium", *Econometrica*, **88**, 1411–1452.
- FOGEL, R. (1964), *Railroads and American Economic Growth: Essays in Econometric History* (Baltimore: The Johns Hopkins Press).
- FOGEL, R. (1979), "Notes on the Social Saving Controversy", *The Journal of Economic History*, **39**, 1–54.
- FOOTE, S. (1974), *The Civil War: A Narrative* (New York: Random House).
- GALOR, O. and OZAK, O. (2016), "The Agricultural Origins of Time Preference", *American Economic Review*, **106**, 3064–3103.
- GHANI, E., GOSWAMI, A. and KERR, W. (2016), "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing", *The Economic Journal*, **126**, 317–357.
- GLAESER, E. (2011), *Triumph of the City* (London: Macmillan).
- HELSLEY, R. and STRANGE, W. (1994), "City Formation with Commitment", *Regional Science and Urban Economics*, **24**, 373–390.
- HENDERSON, V. and VENABLES, A. (2009), "The Dynamics of City Formation", *Review of Economic Dynamics*, **12**, 233–254.
- HERRENDORF, B., ROGERSON, R. and VALENTINYI, Á. (2014), "Growth and Structural Transformation", in Aghion, P. and Durlauf, S. (eds) *Handbook of Economic Growth* (Vol. 2 Chap. 6) (North Holland: Elsevier) 855–941.
- HERRENDORF, B., SCHMITZ, J. and TEIXEIRA, A. (2012), "The Role of Transportation in U.S. Economic Development: 1840–1860", *International Economic Review*, **53**, 693–715.
- IOANNIDES, Y. and OVERMAN, H. (2003), "Zipf's Law for Cities: An Empirical Examination", *Regional Science and Urban Economics*, **33**, 127–137.
- IRWIN, D. (2005), "The Welfare Cost of Autarky: Evidence from the Jeffersonian Trade Embargo, 1807–1809", *Review of International Economics*, **13**, 631–645.
- KIM, S. and MARGO, R. (2004), "Historical Perspectives on U.S. Economic Geography", in Henderson, J. V., and Thisse, J. (eds) *Handbook of Regional and Urban Economics* (Vol. 4, Chap. 66) (North Holland: Elsevier) 2981–3019.
- KRUGMAN, P. (1991), "Increasing Returns and Economic Geography", *Journal of Political Economy*, **99**, 483–499.
- LEBERGOTT, S. (1996), *Consumer Expenditures: New Measures and Old Motives* (Princeton, New Jersey: Princeton University Press).

- LINDERT, P. and WILLIAMSON, J. (2013), "American Incomes Before and After the Revolution", *The Journal of Economic History*, **73**, 725–765.
- LINDSTROM, D. (1979), "American Economic Growth Before 1840: New Evidence and New Directions", *The Journal of Economic History*, **39**, 289–301.
- LIPSEY, R. (1994), "U.S. Foreign Trade and the Balance of Payments, 1800–1913" (NBER Working Paper 4710).
- MAYSHAR, J., MOAV, O., NEEMAN, Z., *et al.* (2022), "The Origin of the State: Land Productivity or Appropriability?", *Journal of Political Economy*, **130**, 1091–1144.
- MICHAELS, G., RAUCH, F. and REDDING, S. (2012), "Urbanization and Structural Transformation", *Quarterly Journal of Economics*, **127**, 535–586.
- MONTE, F., REDDING, S. and ROSSI-HANSBERG, E. (2018), "Commuting, Migration and Local Employment Elasticities", *American Economic Review*, **108**, 3855–3890.
- MUNDLAK, Y. (2005), "Economic Growth: Lessons from Two Centuries of American Agriculture", *Journal of Economic Literature*, **43**, 989–1024.
- NGAI, L. and PISSARIDES, C. (2007), "Structural Change in a Multisector Model of Growth", *American Economic Review*, **97**, 429–443.
- NORTH, D. (2003), *The Role of Institutions in Economic Development: Gunnar Myrdal Lecture* (Geneva: U.N. Palais de Nations).
- PEARSON, G. (2016), "The Democratization of Food: Tin Cans and the Growth of the American Food Processing Industry, 1810–1940" (Doctoral thesis) 2756.
- REDDING, S. (2016), "Goods Trade, Factor Mobility and Welfare", *Journal of International Economics*, **101**, 148–167.
- REDDING, S. and STURM, D. (2008), "The Costs of Remoteness: Evidence from German Division and Reunification", *American Economic Review*, **98**, 1766–1797.
- REDDING, S. and TURNER, M. (2015), "Transportation Costs and the Spatial Organization of Economic Activity", in Duranton, G., Henderson, V. and Strange, W. (eds) *Handbook of Regional and Urban Economics* (Vol. 5, Chap. 20) (North Holland: Elsevier) 1339–1398.
- ROSSI-HANSBERG, E. and WRIGHT, M. (2007), "Urban Structure and Growth", *Review of Economic Studies*, **74**, 597–624.
- SANTAMARIA, M. (2022), "Reshaping Infrastructure: Evidence from the Division of Germany" (Mimeo).
- SWISHER, S. (2014), "Reassessing Railroads and Growth: Accounting for Transport Network Endogeneity" (Mimeo).
- TOWNE, M. and RASMUSSEN, W. (1960), "Farm Gross Product and Gross Investment in the Nineteenth Century", in Parker, W. (ed.) *Trends in the American Economy in the Nineteenth Century*, NBER Studies in Income and Wealth (Vol. 24) (Princeton, New Jersey: Princeton University Press) 255–316.
- TREW, A. (2020), "Endogenous Infrastructure Development and Spatial Takeoff in the First Industrial Revolution", *American Economic Journal: Macroeconomics*, **12**, 44–93.
- TURNER, F. (1894), *The Significance of the Frontier in American History* (Madison, Wisconsin: State Historical Society of Wisconsin).
- WEISS, T. (1992), "U.S. Labor Force Estimates and Economic Growth, 1800–1860", in Gallman, R. and Wallis, J. (eds) *American Economic Growth and Standards of Living Before the Civil War* (Chicago: University of Chicago Press), 19–78.