Hinterlands, city formation and growth: Evidence from the U.S. westward expansion*

Dávid Krisztián Nagy†

October 2, 2022

Abstract

I study how geography shaped city formation and aggregate development in the United States 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.

1 Introduction

In the very end of the 18th century, the United States was still a rural, pre-industrial country. The average income of a U.S. person was below average income in England and

*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 support from the International Economics Section at Princeton University. All errors are my own.

†CREi, Universitat Pompeu Fabra and Barcelona GSE. Email: dnagy@crei.cat.

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 favored 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.[1] 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 labor to produce a homogeneous good under constant returns to scale, and sells the good to consumers and to the non-farm sector. Firms in the non-farm sector combine the farm good with labor 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 labor 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

[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.

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, Nagy and Rossi-Hansberg (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 United States. 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 implcation 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 modeling 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,b) 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.[2]

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, Parro, Rossi-Hansberg and Sarte (2018), Fajgelbaum and Redding (2022), Monte, Redding and Rossi-Hansberg (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, Rauch and Redding (2012).[3] 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

[2]See Herrendorf, Rogerson and Valentinyi (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, Dvorkin and Parro (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 labor mobility but an endogenous dynamic evolution of productivity.

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 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.[4] 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.[5] 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 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 United States. 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

[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, Schmitz and Teixeira (2012), Santamaria (2019), 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.

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 2,500 inhabitants. I classify the settlements above 10,000 as *cities*, and those between 2,500 and 10,000 as *towns*.[6] 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 the rail network in 1835, 1840, 1845, 1850 and 1860. 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 modeling the Civil War is outside the scope of my theoretical framework, I focus on the years 1790 to 1860 throughout the paper.

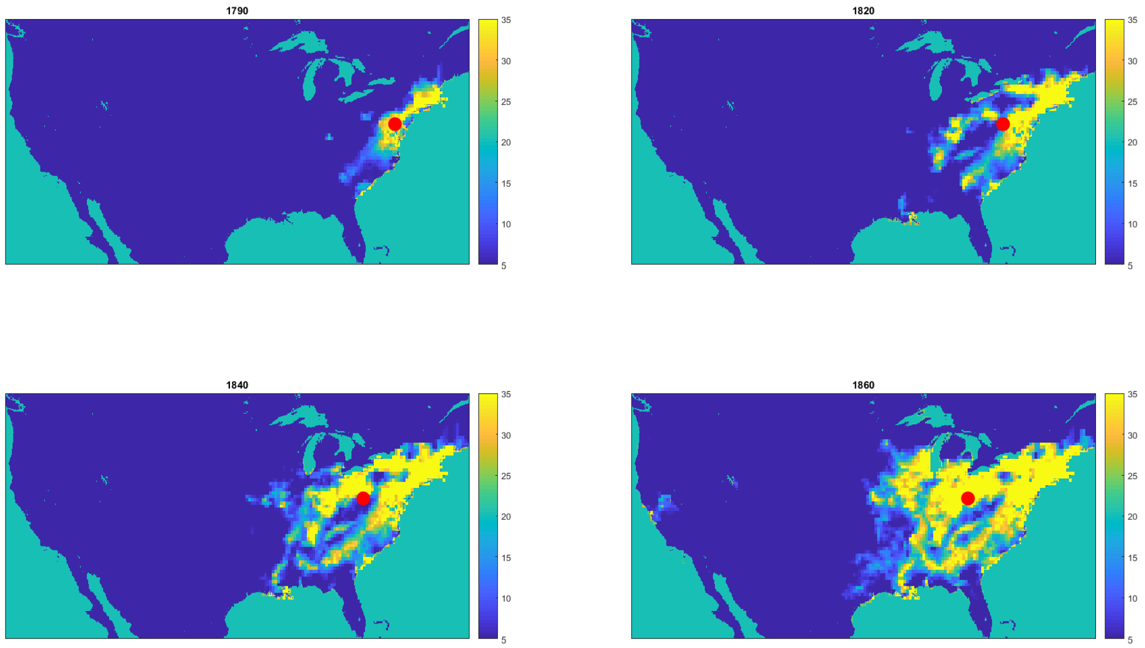
Fact 1: Westward expansion

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).[7] Population growth was, however, far from

[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).

[7]Although modeling 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.

Figure 1: Evolution of the U.S. population distribution



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 United States in 1790, 1820, 1840 and 1860, respectively. The red dot represents the mean center of U.S. population. The scale is bottom-coded at 5 and top-coded at 35 for visibility. Source: U.S. census.

uniform across space. Population in the West increased much more rapidly, as evident from Figure 1 which shows population density and the mean center of U.S. population in 1790, 1820, 1840 and 1860. By the end of the period, the mean center 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 United States: the Northeast, the South and the Midwest.[8] 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 Appendix E).[9]

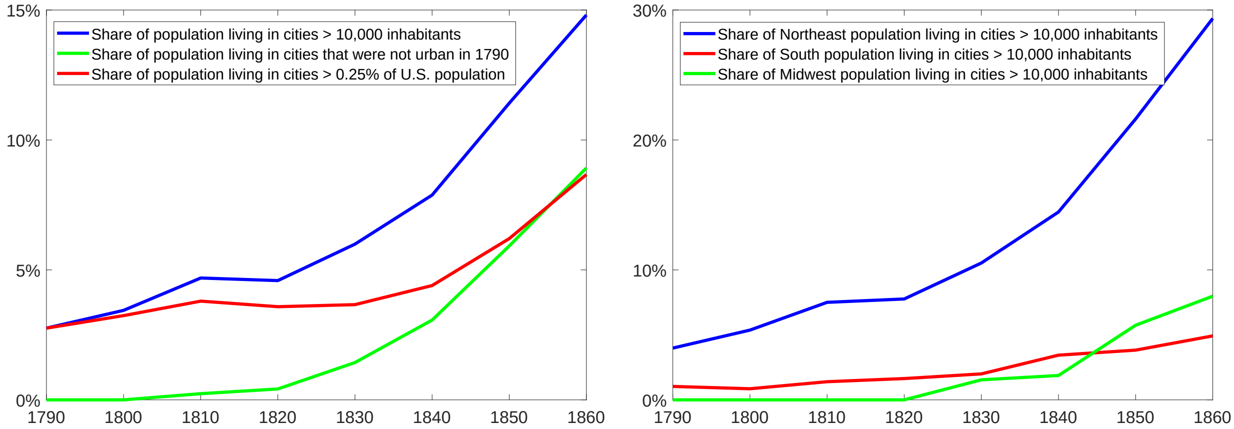
Fact 2: Rapid urbanization

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,

[8] I follow Caselli and Coleman (2001) when defining the geographic boundaries of these regions. See 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 2: Urbanization in the pre-Civil War U.S.



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 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.

The fact that the number of cities above 10,000 inhabitants rose from five (New York City, Philadelphia, Boston, Charleston and Baltimore) to 99 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 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 (Appendix E) shows the locations of the 94 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,

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%

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 3 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*.

I discretize the territory of the U.S. into 20 by 20 arc minute grid cells,[10] 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,[11] and consider specifications of the form

$$newcity_i = \beta_0 + \beta_1 tradeopp_i + \beta_2 amen_i + \beta_3 prod_i + \epsilon_i \tag{1}$$

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 (1) as a

[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 to 1990 yields of the major crops grown in 19th-century U.S. Appendix B describes each of these variables in detail. Section 4.4 further discusses the assumption that 1961 to 1990 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**			0.018**		
	(0.003)			(0.002)		
confluence		0.032**			0.029**	
		(0.004)			(0.004)	
railroad			0.035**			0.030**
			(0.004)			(0.004)
Prod & amenities	No	No	No	Yes	Yes	Yes
No. of observations	7641	7641	7641	7641	7641	7641

This table presents the results of estimating equation (1). The unit of observation is a 20 by 20 arc minute grid cell in the 1860 United States. 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 & 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 3 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 5%; **: significant at 1%. Source: U.S. census, ESRI Map of U.S. Major Waters, *oldrailhistory.com* and FAO GAEZ.

linear probability model.[12]

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 β_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 to 4 percent 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 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 (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 of a theoretical framework in which incentives to trade, and not other forces, attract people to locations with good trading opportunities.

[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

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 United States, 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.

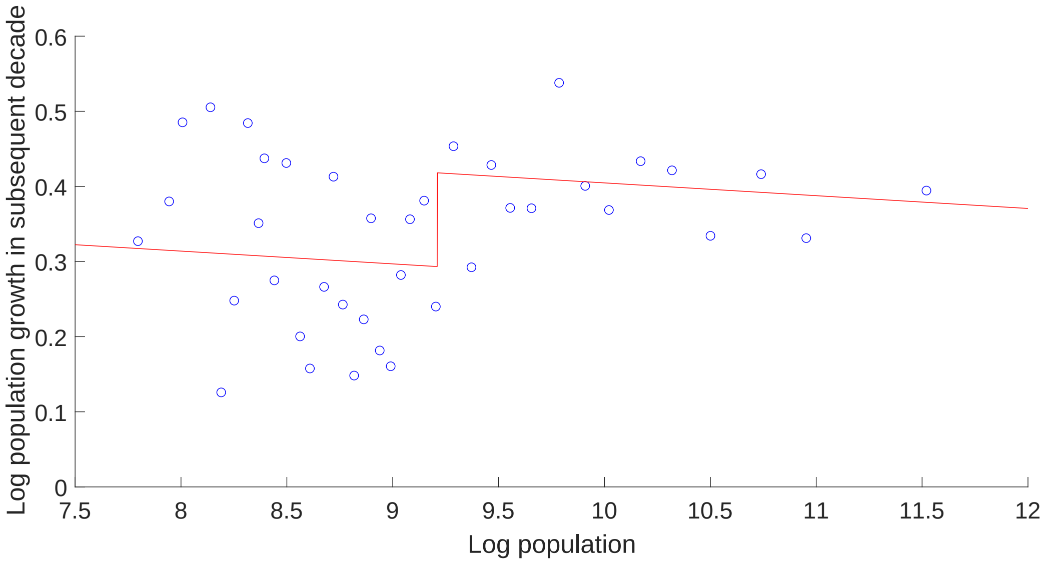
Fact 4: Cities clustered more in the Northeast

Besides cities' locations relative to trading routes, cities' locations relative to each other might also be of interest. A quick glance at Figure 16 (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 neighbor.

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 neighbor 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.[13]

[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



This figure presents a binscatter of log decennial growth against log population among settlements above 2,500 inhabitants in the United States 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 40 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.

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 2,500 and 10,000 inhabitants). Figure 3 plots a binscatter of the log decennial growth rate of these settlements against their log size in the beginning of the decade.[14] 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:

$$ln(growth_{st}) = \gamma_0 + \gamma_1 ln(size_{st}) + \gamma_2 city_{st} + \eta_{st}$$

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 (Appendix E) presents the results of the above regression specification. The estimated coefficient on $city_{st}$ is indeed significantly different from zero. Columns (2) to (4) consider thresholds different from 10,000, and find no

[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 2,500 inhabitants.

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).

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, Eeckhout 2004, Ioannides and Overman 2003). 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 United States – 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 $t \in \{1, 2, ..., \}$, an exogenously given mass of $\bar{L}_t$ consumers live in the U.S.[15] Each consumer owns one unit of labor 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.[16] Section 3.1.2 describes the choices and constraints of those choosing to work in the farm sector ("farmers"), while Section 3.1.3 describes the choices and constraints of those choosing to work in the non-farm sector.

3.1.2 Farmers

Farmers pick a *residential location*,[17] 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.[18]

Farmers order non-farm varieties according to 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:

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

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

[15]Note that I allow U.S. populaton to chage over time, which allows me to incorporate birth, death and immigration. Modeling 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).

[17]I abstract from slavery. I discuss this modeling 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. 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.

labor 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 labor used (which equals one due to the inelastic supply of labor), 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

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

where $p_t^N(s,i)$ is the price of non-farm variety i and $p_t^F(s)$ is the price of the farm good at the trading place, and $R_t(r)$ 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 $y_t(r)$, 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.[19]

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

$$y_t(r) = R_t(r)\, h_t(r)$$

As a result, the farmer's nominal income that she can spend on consumption goods equals her production revenues, $p_t^F(s)\,\tilde{q}_t^F(r,s)$. 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:

$$U_t^F(r,s) = \frac{p_t^F(s)\,\tilde{q}_t^F(r,s)}{P_t(s)^{\nu}\, p_t^F(s)^{1-\nu}} = \frac{p_t^F(s)\,\tau_t^F(r,s)^{-1} B_t(r)\left[\frac{H_t(r)}{L_t^F(r)}\right]^{1-\alpha}}{P_t(s)^{\nu}\, p_t^F(s)^{1-\nu}} \tag{2}$$

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

$$P_t(s) = \left[\int_0^{n_t} p_t^N(s,i)^{1-\epsilon}\, di\right]^{\frac{1}{1-\epsilon}}.$$

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 $B_t(r)$. In practice, however, I lack data on actual farm productivity and match $B_t(r)$ 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 South during slavery. In fact, the model underestimates population and economic activity in the South compared to the data (Table 15 in 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.

[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.

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:

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

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. ζ 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

$$U_t^N(s) = \frac{w_t(s)}{P_t(s)^\nu \, p_t^F(s)^{1-\nu}}. \tag{3}$$

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 9). As a consequence, locations with a larger non-farm population will produce more varieties (equation 12). 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 labor and the farm good to produce. The production function is CES with elasticity of substitution β between the two inputs,

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

where $q_t^N(s,i)$ is the quantity of the variety produced, $\ell_t^P(s,i)$ is the quantity of labor 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 labor) to using non-farm labor, hence structural change and urbanization. In particular, we need $\beta < 1$, that is, complementarity between labor 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 β.[20] 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 labor 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.[21] Workers can freely switch between the two tasks, innovation and production. As I show later in Section 3.1.6, a firm always has positive demand for both types of labor. 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 labor $(1-\mu)$ and the farm good (μ) 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, labor use, farm good use and

[20] As an alternative specification, one could consider the two inputs being substitutes and productivity being labor-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).

output price as a function of input prices and productivity (Section 3.1.6). 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 labor used in shipping and handling. In other words, I regard the distribution of agricultural 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

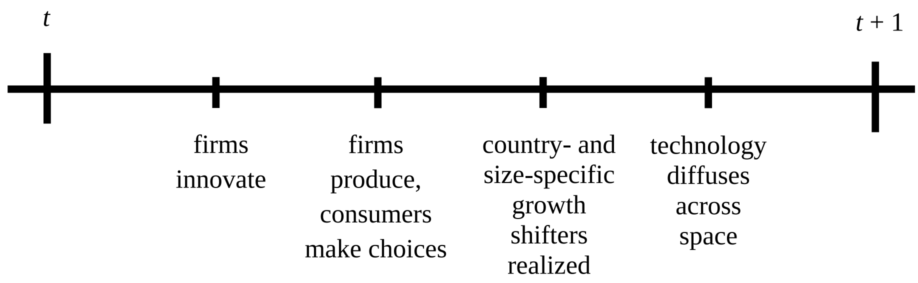
To incorporate the heterogeneity in growth between cities and towns observed in the data, I allow productivity growth between periods t and $t+1$ at location s to depend not only on firms' period-t innovation, but also on size-dependent dynamic externalities $g\left(L_t^N(s)\right)$, where $L_t^N(s)$ 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

$$\tilde{A}_{t+1}(s) = A_t(s)\left[\ell_t^I(s)^{1-\mu} + g\left(L_t^N(s)\right)\right] \tag{4}$$

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 $g\left(L_t^N(s)\right) = \gamma$ if $L_t^N(s) \geq \lambda$, and $g\left(L_t^N(s)\right) = 0$ 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 λ) and *towns* (locations with non-farm population below λ). 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 neighbors in period t, but the farther away a neighbor 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

Figure 4: Timing of events in the model



actually use in period $t+1$ is the best of all these available technologies:

$$A_{t+1}(s) = max_u e^{-\delta|u-s|} \tilde{A}_{t+1}(u) \tag{5}$$

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 (4). Finally, technology diffuses across space, which leads to a productivity level $A_{t+1}(s)$ at location s, given by (5). 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).[22] It is important to note, however, that average labor 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 labor 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 labor productivity in the model than in the data.

[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.

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 3.1.4 and 3.1.5. 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 (5) 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,

$$p_t^N(s) = \frac{\epsilon}{\epsilon - 1}\bar{c}_t(s)\, w_t(s) \tag{6}$$

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 (6), that the firm's output is

$$q_t^N(s) = (\epsilon - 1)\, f \bar{c}_t(s)^{-1}. \tag{7}$$

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

$$\bar{c}_t(s) = \left[1 + A_t(s)^{\beta-1}\left[\frac{p_t^F(s)}{w_t(s)}\right]^{(1-\beta)\mu}\right]^{\frac{1}{1-\beta}} \tag{8}$$

and optimal factor quantities can be obtained from Shephard's Lemma and equation (7) as

$$\ell_t^P(s) = (\epsilon - 1)\, f \bar{c}_t(s)^{\beta-1}, \tag{9}$$

$$\ell_t^I(s) = (1-\mu)(\epsilon - 1)\, f A_t(s)^{\beta-1}\left[\frac{p_t^F(s)}{w_t(s)}\right]^{(1-\beta)\mu} \bar{c}_t(s)^{\beta-1} \tag{10}$$

and

$$x_t^F(s) = \mu(\epsilon - 1)\, f A_t(s)^{\beta-1}\left[\frac{p_t^F(s)}{w_t(s)}\right]^{(1-\beta)\mu-1} \bar{c}_t(s)^{\beta-1}. \tag{11}$$

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

$$n_t(s) = \frac{L_t^N(s)}{\ell_t^P(s) + \ell_t^I(s) + f} = \frac{L_t^N(s)}{\left[(\epsilon - 1)\left(\mu \bar{c}_t(s)^{\beta-1} + 1 - \mu\right) + 1\right] f} \tag{12}$$

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 α, β, γ, δ, ϵ, λ, μ, ν, f, the set of locations S, total population $\bar{L}_t$, as well as functions $H_t, A_1, B : S \to \mathbb{R}$ and $\tau_t^F, \tau_t^N : S^2 \to \mathbb{R}$, an* ***equilibrium*** *of the economy is a set of functions $L_t^F, L_t^N, p_t^F, p_t^N, q_t^N, h_t, \ell_t^P, \ell_t^I, x_t^F, \bar{c}_t, n_t, P_t, w_t, R_t, A_t : S \to \mathbb{R}$ and $\sigma_t : S \to S$, as well as the level of consumers' utility U_t for each time period $t \in \{1, 2, ...\}$ such that the following hold:*

1. *Farmers maximize utility, implying that farmers living at r choose their trading place $\sigma_t(r)$[23] such that*

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

and their utility is equalized across locations, that is,

$$U_t^F(r, \sigma_t(r)) = U_t^F(u, \sigma_t(u)) \tag{13}$$

for any $r, u \in S$, such that $U_t^F(\cdot, \cdot)$ is given by (2).

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

$$U_t^N(s) = U_t^N(u) \tag{14}$$

for any $s, u \in S$ with positive non-farm population, where $U_t^N(\cdot)$ is given by (3).

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

$$U_t = U_t^F(r, s) = U_t^N(s) \tag{15}$$

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

4. *Non-farm firms maximize profits, and free entry drives down profits to zero. Hence, their price is given by (6), their output by (7), their unit cost per wage by (8), and*

[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.

their factor use by (9) to (11). The number of goods produced at location s is given by (12), and the price index at s is

$$P_t(s) = \left[\int_S n_t(u)\, p_t^N(u)^{1-\epsilon}\, \tau_t^N(u,s)^{1-\epsilon}\, du\right]^{\frac{1}{1-\epsilon}}.^{24} \tag{16}$$

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

$$\int_{\sigma_t^{-1}(s)} \nu \tau_t^F(r,s)^{-1} B(r)\, L_t^F(r)^{\alpha} H_t(r)^{1-\alpha}\, dr = x_t^F(s)\, n_t(s) + (1-\nu)\frac{w_t(s)}{p_t^F(s)} L_t^N(s) \tag{17}$$

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,*

$$q_t^N(s) = \int_S \nu p_t^N(s)^{-\epsilon}\, \tau_t^N(s,u)^{1-\epsilon}\, P_t(u)^{\epsilon-1}\, I_t(u)\, du \tag{18}$$

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 labor market clears. That is,*

$$\bar{L}_t = \int_S \left[L_t^F(r) + L_t^N(r)\right] dr. \tag{19}$$

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

$$H_t(r) = L_t^F(r)\, h_t(r). \tag{20}$$

9. *Productivity levels evolve according to equations (4) and (5).*

[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.

3.2 Extension to multiple countries

This section extends the model to incorporate a finite number of countries $c \in \{1, ..., C\}$, such that each location r belongs to one country $c(r)$.[25] 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.

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

$$\tilde{A}_{t+1}(s) = A_t(s) \left[\ell_t^I(s)^{1-\mu} + f_{c(s),t} + g\left(L_t^N(s)\right) \right]. \tag{21}$$

All other parts of the model are unchanged. 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 labor markets clear by country, and productivity is assumed to evolve according to equation (21) instead of equation (4).

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 labor 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*

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

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 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 labor input ratio. As the two inputs are complements, everything else fixed, such firms will hire more non-farm labor 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.[26]

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 favored 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,

[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).

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 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 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 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:

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

$$\ell_t^I(s)^{1-\mu} = \rho \left[\left(1 + \frac{L_t^N(s)}{\int_{\sigma_t^{-1}(s)} L_t^F(r)\, dr} \right)^{-1} - (1-\nu) \right]^{1-\mu} \tag{22}$$

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

Proof. See Appendix A.6.

□

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 (22), (21) and (5) 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.[27]

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 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.

[27] For any given population distribution, equations (22), (21) and (5) 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. Appendix A.4 provides intuition for why the structure of the model does not allow for a theoretical characterization of equilibrium uniqueness.

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 United States. 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.[28] $WR_t(r)$ 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*.[29]

4.2 Incorporating international trade

I assume that the world not only consists of a single country (the United States), 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% to 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

[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 Appendix B.

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 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, ς^R and ς^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.[30] In what follows, I treat ϕ^F and ϕ^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 United Kingdom was the most important trading partner of the U.S. within Europe (Lipsey, 1994). Next, I set up the

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

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 ϕ^E is a parameter driving the scale of international shipping costs. Akin to ϕ^F and ϕ^N, I treat ϕ^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, 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 to 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.[31] Proxying 19th-century yields with the ones for the period 1961 to 1990 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.[32] 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.

[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 neighbors.

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.

Structural parameters borrowed from the literature. I borrow the values of four parameters – the labor share in farming α and the non-farm share in consumption ν, consumers' elasticity of substitution among non-farm varieties ϵ, and the strength of technology diffusion δ – from the literature. In particular, I follow Caselli and Coleman (2001) and set the labor share in farming to $\alpha = 0.71$.[33] 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 ϵ 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 $\epsilon = 9$.[34]

Finally, to obtain the value of the technology diffusion parameter δ, I rely on evidence provided by Comin, Dmitriev and Rossi-Hansberg (2013) on the spatial diffusion of technology. Comin et al. (2013) estimate the value of δ 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 kilometers. Given the uncertainty about the specific value of δ, I check the robustness of my quantitative results to different values of this parameter in Section 6.1.

Non-farm productivity in Europe. To choose the initial non-farm productivity of

[33] Caselli and Coleman (2001) find a labor 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 labor 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 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.

Europe $A_1(e)$ and the European growth shifters f_{et},[35] note that, by equation (21), this is the same problem as choosing the non-farm productivity of Europe every period, $A_t(e)$. Also note that, combining equations (8), (11), (12) and (17), it is possible to express the non-farm productivity of Europe as

$$A_t(e) = \epsilon^{\frac{1}{\beta-1}} \tilde{B}(e)^{-\mu} L_t^F(e)^{(1-\alpha)\mu} \left[\frac{\mu(\epsilon-1)}{\nu \frac{L_t^F(e)}{L_t^N(e)} - (1-\nu)} - \mu - \epsilon(1-\mu) \right]^{\frac{1}{1-\beta}} \tag{23}$$

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 was employed in the non-farm sector. This provides me with estimates of European farm and non-farm populations.[36] Plugging these estimates into equation (23), 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.

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.

Calibration. We are left with choosing the values of non-farm productivity parameters $\bar{A}$ and $\underline{A}$, the scale parameters of shipping costs ϕ^F, ϕ^N and ϕ^E, Europe's land-adjusted farm productivity $\tilde{B}(e)$, and the five remaining structural parameters: the fixed cost in non-farming f, the elasticity of substitution between labor and the farm good in non-farming β, the share of the farm input in non-farming μ, as well as λ and γ that drive dynamic

[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 to 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 avg log decennial growth rate b/w cities and towns, 1790 to 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 to 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)

Each row in this table corresponds to one of the 11 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.

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 π that rationalize the data moments requires me to choose the π that minimizes the loss function

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

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 π, 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 β, recall that it plays a key role in generating city growth, as discussed in Section 3.1.4. The lower the value of β, the higher the degree of complementarity between productivity and non-farm labor, hence the more productivity growth draws non-farm population into cities. Thus, I choose the value of β 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 μ, note that an increase in μ increases demand for the farm input, and therefore farm labor. Thus, I calibrate μ to match the share of farm employment in total employment, which equaled 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 2,500 and 10,000 inhabitants is 0.08 log points in the data between 1790 and 1820. I choose the value of γ such that the decennial growth rates of cities and towns (locations above versus below a non-farm population of λ) 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 2,500 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, ϕ^F, ϕ^N and ϕ^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 ϕ^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 ϕ^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 ϕ^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 ϕ^N/ϕ^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, ϕ^E, drives the importance of international trade in the model. As a result, I calibrate the value of ϕ^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

Table 5: Correlation of population between the model and the data

t	Correlation, period t		Correlation, change from 1790 to t	
	Population (1)	Population density (2)	Population (3)	Population density (4)
1790	0.379	0.366	–	–
1800	0.440	0.441	0.383	0.363
1810	0.538	0.540	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

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.

the most densely populated counties were small, usually fully contained in a single cell.

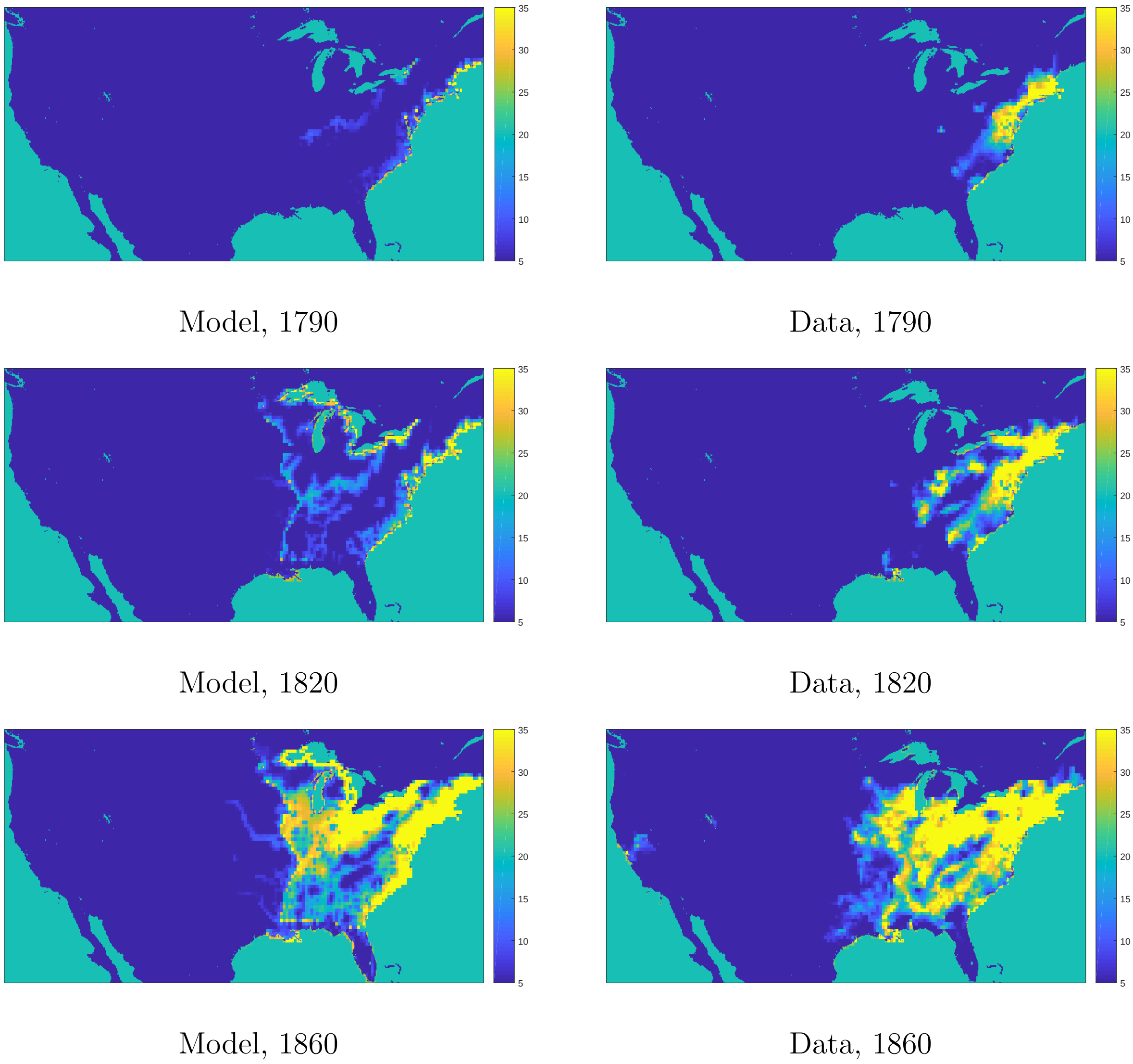
Column (1) of Table 5 shows that the correlation between cells' population levels predicted by the model and those in the data is high. Column (2) shows the same for population density (that is, population per unit of land). Even though I only match city and town cells' total population and the share 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.[37] 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 Appendix E).

To assess performance of the model in replicating urbanization as well as the locations

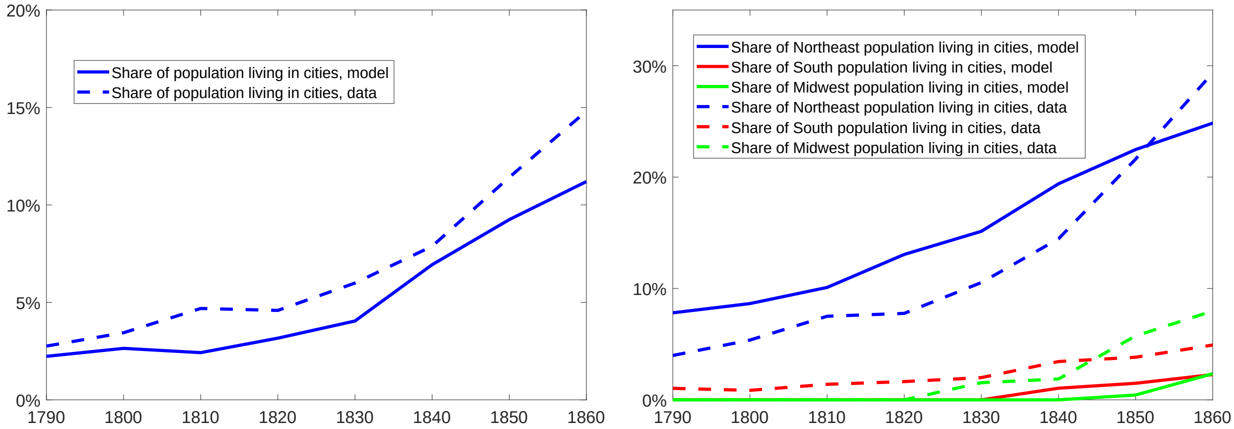
[37] In the slow migration process presented in Section 2, the correlation is one in 1790 as the data are matched exactly in the initial period. However, the correlation drops quickly over time, and is already below 0.3 in 1860. I also simulate a process in which I extrapolate the 1790 to 1820 population 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.

Figure 5: Population distribution: model versus data



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 35 for visibility. Source: U.S. census.

Figure 6: Urbanization in the model



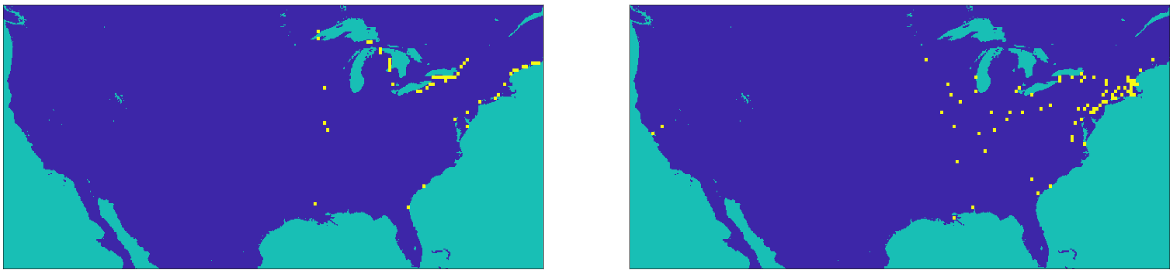
In the left graph, the solid blue line shows the ratio of non-farm population living in cities (cells with more than 10,000 non-farm inhabitants) to total U.S. population in the baseline model simulation for each census year. The dotted blue line shows the fraction of U.S. population living in cities (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.

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% 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 (50 in the model, 75 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 non-farm activity. However, one would expect this effect to be

Figure 7: City locations by 1860: model versus data



Model Data

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.

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

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 United States, 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.

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 United States. 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 neighbor 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-

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%

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 3 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.*

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).[38] 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 (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 5.3.2, 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

[38] Glaeser (2011) argues that the leading position of New York over other major ports in the Northeast depended on its deep and protected harbor 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.

West.[39] 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 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

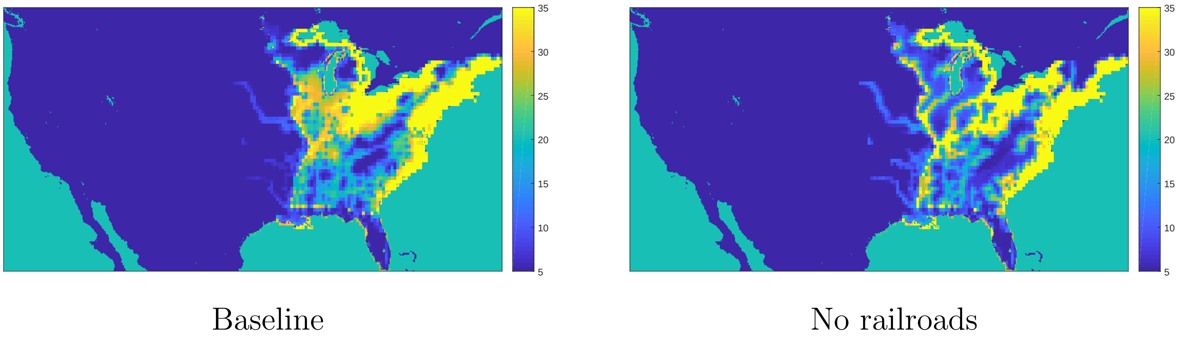
This section evaluates the first implication of the hinterland hypothesis: the prediction that railroads fostered city development and aggregate U.S. growth. How much railroads contributed to American economic growth and output has been the subject of long debates in economic history. In his book, Fogel (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 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 counter-

[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. 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



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 5 and top-coded at 35 for visibility.

factual, 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 (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

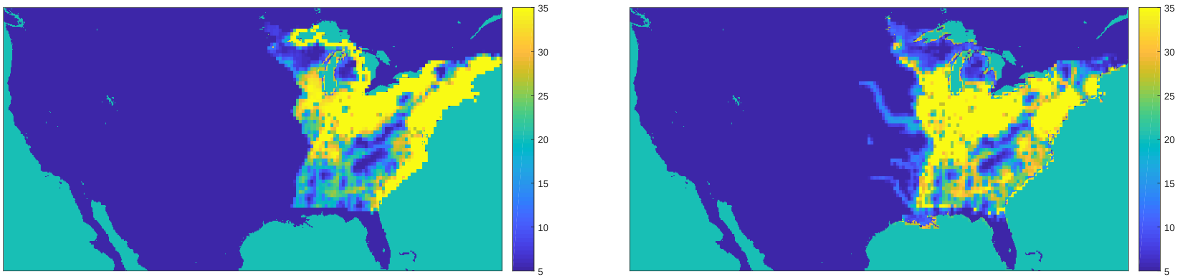
The remaining two implications of the hinterland hypothesis have to do with how other elements of U.S. geography – political border changes and international trade – shaped U.S. growth and city formation. To evaluate each of these two implications, I conduct two further counterfactuals in this section.

5.3.1 The effect of political border changes

In 1790, the political border of the United States 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

Figure 9: Further counterfactuals: population distribution in 1860



No political border changes | No international trade

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 5.3.1); the right map presents the same in the counterfactual with no international trade (Section 5.3.2). The scale is bottom-coded at 5 and top-coded at 35 for visibility.

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 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 United States 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 favorable effect on Northeast cities, international trade significantly increased aggregate growth in the pre-Civil War United States.

One should take the quantitative findings of Section 5.3.2 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 δ from its baseline value, 0.006, to 0.003. In a second exercise, I weaken technology diffusion by increasing δ to 0.009. In the third exercise, I eliminate dynamic externalities in cities by setting their scale parameter γ 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 5.3 and 5.4: 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 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.[40] 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 γ 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.

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%	93.6%
– 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 to 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%

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 United States, respectively; the difference in average log decennial population growth rate between cities and towns over the period 1790 to 1820; and the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.3 and 5.4) 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 δ 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 γ is set to zero.

Figure 10: Location of endogenous and actual railroads in 1860

Endogenous railroads

Actual railroads

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*.

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 5.3.1 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 (Allen and Arkolakis 2021, Fajgelbaum and Schaal 2020, Santamaria 2019), 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.[41]

Motivated by this relationship between trade and optimal infrastructure investment, I

[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.

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.[42] 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 United States. 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.

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

[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.

Table 9: Counterfactual results under endogenous railroads

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

The first column of this table shows the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.3 and 5.4) relative to the baseline model simulation. The second column shows the change in U.S. real GDP per capita in the three counterfactuals (Sections 5.3 and 5.4) relative to a new baseline model simulation such that railroads in both the counterfactuals and the new baseline are placed endogenously.

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.

References

- Allen, R. (2000): Economic structure and agricultural productivity in Europe, 1300–1800. *European Review of Economic History*, vol. 3, 1–25.
- Allen, T. and Arkolakis, C. (2014): Trade and the topography of the spatial economy. *Quarterly Journal of Economics*, vol. 129(3), 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 University Press.
- Arkolakis, C., Costinot, A. and Rodríguez-Clare, A. (2012): New trade models, same old gains? *American Economic Review*, vol. 102(1), 94–130.
- Atack, J., Bateman, F., Haines, M. and Margo, R. (2010): Did railroads induce or follow economic growth? Urbanization and population growth in the American Midwest, 1850-1860. *Social Science History*, vol. 34(2), 171–197.
- Bairoch, P. and Goertz, G. (1985): Factors of urbanization in the nineteenth century developed countries: A descriptive and econometric analysis. *Urban Studies*, vol. 23, 285–305.
- Baumol, W. (1967): Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review*, vol. 57(3), 415–426.
- Bertinelli, L. and Black, D. (2004): Urbanization and growth. *Journal of Urban Economics*, vol. 56, 80–96.
- Black, D. and Henderson, V. (1999): A theory of urban growth. *Journal of Political Economy*, vol. 107, 252–284.
- Bleakley, H. and Lin, J. (2012): Portage and path dependence. *Quarterly Journal of Economics*, vol. 127(2), 587–644.
- Buera, F. and Kaboski, J. (2012a): The rise of the service economy. *American Economic Review*, vol. 102, 2540–2569.
- Buera, F. and Kaboski, J. (2012b): Scale and origins of structural change. *Journal of Economic Theory*, vol. 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*, vol. 87(3), 741–835.

- Caliendo, L., Parro, F., Rossi-Hansberg, E., and Sarte, P. (2018): The impact of regional and sectoral productivity changes on the U.S. economy. *Review of Economic Studies*, vol. 85(4), 2042–2096.

- Caselli, F. and Coleman, J. (2001): The U.S. structural transformation and regional convergence: A reinterpretation. *Journal of Political Economy*, vol. 109(3), 584–616.

- Cazier, L. (1976): Surveys and surveyors of the public domain, 1785–1975. 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*, vol. 124, 205–248.

- Cronon, W. (1991): Nature's metropolis: Chicago and the Great West. W. W. Norton & Company.

- David, P. (1969): Transport innovation and economic growth: Professor Fogel on and off the rails. *The Economic History Review*, vol. 22(3), 506–525.

- David, P. (2005): Real income and economic welfare growth in the early republic. *Stanford University Working Paper*.

- Desmet, K., Kopp, R., Kulp, S., Nagy, D., Oppenheimer, M., Rossi-Hansberg, E. and Strauss, B. (2021): Evaluating the economic cost of coastal flooding. *American Economic Journal: Macroeconomics*, vol. 13(2), 444–486.

- Desmet, K., Nagy, D., and Rossi-Hansberg, E. (2018): The geography of development. *Journal of Political Economy*, vol. 126(3), 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*, vol. 98, 50–68.

- Desmet, K. and Rossi-Hansberg, E. (2013): Urban accounting and welfare. *American Economic Review*, vol. 103, 2296–2327.

- Desmet, K. and Rossi-Hansberg, E. (2014): Spatial development. *American Economic Review*, vol. 104(4), 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*, vol. 67(3), 297–308.
- Donaldson, D. and Hornbeck, R. (2016): Railroads and american economic growth: A "market access" approach. *Quarterly Journal of Economics*, vol. 131(2), 799–858.
- Eaton, J. and Eckstein, Z. (1997): Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics*, vol. 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*, vol. 94, 1429–1451.
- Faber, B. (2014): Trade integration, market size, and industrialization: Evidence from China's National Trunk Highway System. *Review of Economic Studies*, vol. 81(3), 1046–1070.
- Fajgelbaum, P. and Redding, S. (2022): Trade, structural transformation and development: Evidence from Argentina 1869-1914. *Journal of Political Economy*, vol. 130(5), 1249–1318.
- Fajgelbaum, P. and Schaal, E. (2020): Optimal transport networks in spatial equilibrium. *Econometrica*, vol. 88(4), 1411–1452.
- Fogel, R. (1964): Railroads and American economic growth: Essays in econometric history. The Johns Hopkins Press.
- Fogel, R. (1979): Notes on the social saving controversy. *The Journal of Economic History*, vol. 39(1), 1–54.
- Foote, S. (1974): The civil war: A narrative. Random House.
- Fujita, M., Krugman, P., and Venables, A. (1999): The spatial economy. Cities, regions, and international trade. The MIT Press.
- Galor, O. and Ozak, O. (2016): The agricultural origins of time preference. *American Economic Review*, vol. 106(10), 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*, vol. 126, 317–357.
- Glaeser, E. (2011): Triumph of the city. Macmillan.

- Helsley, R. and Strange, W. (1994): City formation with commitment. *Regional Science and Urban Economics*, vol. 24(3), 373–390.

- Henderson, V. and Venables, A. (2009): The dynamics of city formation. *Review of Economic Dynamics*, vol. 12, 233–254.

- Herrendorf, B., Rogerson, R., and Valentinyi, Á. (2014): Growth and structural transformation. *Handbook of Economic Growth*, vol. 2, ch. 6, 855–941.

- Herrendorf, B., Schmitz, J., and Teixeira, A. (2012): The role of transportation in U.S. economic development: 1840–1860. *International Economic Review*, vol. 53(3), 693–715.

- Ioannides, Y. and Overman, H. (2003): Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics*, vol. 33, 127–137.

- Irwin, D. (2005): The welfare cost of autarky: Evidence from the Jeffersonian trade embargo, 1807–09. *Review of International Economics*, vol. 13(4), 631–645.

- Kim, S. and Margo, R. (2004): Historical perspectives on U.S. economic geography. *Handbook of Regional and Urban Economics*, vol. 4, ch. 66, 2981–3019.

- Krugman, P. (1991): Increasing returns and economic geography. *Journal of Political Economy*, vol. 99(3), 483–499.

- Lebergott, S. (1996): Consumer expenditures: New measures and old motives. Princeton University Press.

- Lindert, P. and Williamson, J. (2013): American incomes before and after the revolution. *The Journal of Economic History*, vol. 73(3), 725–765.

- Lindstrom, D. (1979): American economic growth before 1840: New evidence and new directions. *The Journal of Economic History*, vol. 39(1), 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., and Pascali, L. (2022): The origin of the state: Land productivity or appropriability? *Journal of Political Economy*, forthcoming.

- Michaels, G., Rauch, F., and Redding, S. (2012): Urbanization and structural transformation. *Quarterly Journal of Economics*, vol. 127, 535–586.

- Monte, F., Redding, S., and Rossi-Hansberg, E. (2018): Commuting, migration and local employment elasticities. *American Economic Review*, vol. 108(12), 3855–3890.

- Mundlak, Y. (2005): Economic growth: Lessons from two centuries of American agriculture. *Journal of Economic Literature*, vol. 43, 989–1024.
- Nagy, D. (2020): Trade and urbanization: Evidence from Hungary. *American Economic Journal: Microeconomics*, forthcoming.
- Ngai, L. and Pissarides, C. (2007): Structural change in a multisector model of growth. *American Economic Review*, vol. 97, 429–443.
- North, D. (2003): The role of institutions in economic development. Gunnar Myrdal Lecture, 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*, vol. 101, 148–167.
- Redding, S., Sturm, D. (2008): The costs of remoteness: Evidence from German Division and Reunification. *American Economic Review*, vol. 98(5), 1766–1797.
- Redding, S. and Turner, M. (2015): Transportation costs and the spatial organization of economic activity. *Handbook of Urban and Regional Economics*, vol. 5, ch. 20, 1339–1398.
- Rossi-Hansberg, E. and Wright, M. (2007): Urban structure and growth. *Rewiew of Economic Studies*, vol. 74(2), 597–624.
- Santamaria, M. (2022): Reshaping infrastructure: Evidence from the division of Germany. Mimeo.
- Starrett, D. (1978): Market allocations of location choice in a model with free mobility. *Journal of Economic Theory*, vol. 17, 21–37.
- 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 *Trends in the American Economy in the Nineteenth Century*, ed. Parker, W. NBER Studies in Income and Wealth, 24. Princeton University Press.
- Trew, A. (2020): Endogenous infrastructure development and spatial takeoff in the First Industrial Revolution. *American Economic Journal: Macroeconomics*, vol. 12(2), 44–93.

- Turner, F. (1894): The significance of the frontier in American history. State Historical Society of Wisconsin.
- Weiss, T. (1992): U.S. labor force estimates and economic growth, 1800-1860. In *American Economic Growth and Standards of Living before the Civil War*, 19–78. University of Chicago Press.

A Theory appendix

This appendix supplements the theoretical framework presented in Section 3 in six respects. First, Section A.1 defines the equilibrium of the multi-country model. Next, Section A.2 shows how the population distribution in period t in the multi-country model can be obtained by solving a system of three equations, and describes an algorithm to solve these equations. Section A.3 presents a special case of the model with a homogenous hinterland around a non-farm location. Section A.4 discusses why multiplicity of equilibria may arise in the model, as well as why the complex structure of the model does not allow for a theoretical characterization of equilibrium uniqueness. Section A.5 presents a version of the model in which consumption happens at the residential location, as well as a condition on shipping costs that guarantees isomorphism between this alternative model and the model of Section 3. Finally, Section A.6 provides proofs.

A.1 Definition of the equilibrium in the multi-country model

Definition. *Given parameters α, β, γ, δ, ϵ, λ, μ, ν, f, the set of locations S, countries' population $\bar{L}_{ct}$ and growth shifters f_{ct}, as well as functions $H_t, A_1, B : S \to \mathbb{R}$ and $\tau_t^F, \tau_t^N : S^2 \to \mathbb{R}$, an* ***equilibrium*** *of the economy is a set of functions $L_t^F, L_t^N, p_t^F, p_t^N, q_t^N, h_t, \ell_t^P, \ell_t^I, x_t^F, \bar{c}_t, n_t, P_t, w_t, R_t, A_t : S \to \mathbb{R}$ and $\sigma_t : S \to S$, as well as countries' utility levels U_{ct} for each time period $t \in \{1, 2, ...\}$ such that the following hold:*

1. *Farmers maximize utility, implying that farmers living at r choose their trading place $\sigma_t(r)$ such that*
$$\sigma_t(r) = argmax_{s \in S_c} U_t^F(r, s)$$
and their utility is equalized across locations within each country, that is,
$$U_t^F(r, \sigma_t(r)) = U_t^F(u, \sigma_t(u)) \tag{24}$$
for any $r, u \in S_c$, such that $U_t^F(\cdot, \cdot)$ is given by (2).
2. *Non-farm workers maximize utility, and their utility is equalized across locations within each country, that is,*
$$U_t^N(s) = U_t^N(u) \tag{25}$$

for any $s, u \in S_c$ with positive non-farm population, where $U_t^N(\cdot)$ is given by (3).

3. *Consumers' utility is equalized across sectors within each country. Hence,*

$$U_{ct} = U_t^F(r, s) = U_t^N(s) \tag{26}$$

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

4. *Non-farm firms maximize profits, and free entry drives down profits to zero. Hence, their price is given by (6), their output by (7), their unit cost per wage by (8), and their factor use by (9) to (11). The number of goods produced at location s is given by (12), and the price index at s is*

$$P_t(s) = \left[\int_S n_t(u)\, p_t^N(u)^{1-\epsilon}\, \tau_t^N(u, s)^{1-\epsilon}\, du\right]^{\frac{1}{1-\epsilon}}. \tag{27}$$

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

$$\int_{\sigma_t^{-1}(s)} \nu \tau_t^F(r, s)^{-1} B(r)\, L_t^F(r)^{\alpha} H_t(r)^{1-\alpha}\, dr = x_t^F(s)\, n_t(s) + (1-\nu)\frac{w_t(s)}{p_t^F(s)} L_t^N(s) \tag{28}$$

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 location s. That is,*

$$q_t^N(s) = \int_S \nu p_t^N(s)^{-\epsilon}\, \tau_t^N(s, u)^{1-\epsilon}\, P_t(u)^{\epsilon-1}\, I_t(u)\, du \tag{29}$$

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. *National labor markets clear. That is,*

$$\bar{L}_{ct} = \int_{S_c} \left[L_t^F(r) + L_t^N(r)\right] dr \tag{30}$$

where S_c denotes the set of locations that belong to country c.

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

$$H_t(r) = L_t^F(r)\, h_t(r)\,. \tag{31}$$

9. Productivity levels evolve according to equations (21) and (5).

The equilibrium of the multi-country model differs in three respects from the equilibrium of the single-country model, which was defined in Section 3.1.7. First, as consumers cannot move across countries, utility equalization conditions (24), (25) and (26) are only imposed across locations that are in the same country. Second, the national labor market clearing condition (30) is imposed by country, rather than across all locations. Finally, equation (21) replaces equation (4) as the law of motion for local productivity.

A.2 Solving the equilibrium population distribution in period t

To simplify the model's period-t equilibrium conditions – that is, the system of equations (2), (3), (6) to (12) and (24) to (31) –, note first that farmers living at a trading place always want to trade there. The intuition for this result is that, if farmers living at s preferred some other trading place u to s, then, by the triangle inequality of shipping costs, farmers living at any other location would also prefer u to s. Hence, s would not even arise as a trading place.

But what is the trading place chosen by farmers who do not live at one? Clearly, farmers living at a location r in country c choose the trading place that maximizes their utility (2). Note also that, by (26), the utility of a farmer living at a trading place $s \in S_c$ equals the utility of a non-farmer (3),

$$\frac{p_t^F(s)\, B(s) \left[\frac{H_t(s)}{L_t^F(s)}\right]^{1-\alpha}}{P_t(s)^{\nu}\, p_t^F(s)^{1-\nu}} = \frac{w_t(s)}{P_t(s)^{\nu}\, p_t^F(s)^{1-\nu}}$$

from which the price of the farm good can be expressed as

$$p_t^F(s) = B(s)^{-1} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{1-\alpha} w_t(s)\,. \tag{32}$$

Plugging this back into (2), one obtains

$$U_t^F(r,s) = \tau_t^F(r,s)^{-1} \frac{B(r)}{B(s)} \left[\frac{H_t(r)}{H_t(s)}\right]^{1-\alpha} \left[\frac{L_t^F(r)}{L_t^F(s)}\right]^{-(1-\alpha)} U_{ct}$$

where I used (26) again to substitute $U_{ct} = \frac{w_t(s)}{P_t(s)^{\nu} p_t^F(s)^{1-\nu}}$. The trading place s which is optimal for farmers at location r is the one that maximizes the above expression, thus it is

$$\sigma_t(r) = argmax_{s \in S_c} \tau_t^F(r,s)^{-1} B(s)^{-1} H(s)^{-(1-\alpha)} L_t^F(s)^{1-\alpha}. \tag{33}$$

Once we know who trades where, utility equalization relates the farm population of any location to the farm population of its trading place. To see this, consider a location r together with its trading place $\sigma_t(r)$. By (24), farmers living at these two places have the same utility, that is,

$$U_t^F(r, \sigma_t(r)) = U_t^F(\sigma_t(r), \sigma_t(r)).$$

Substituting for $U_t^F(\cdot,\cdot)$ using (2), one can express the farm population of r as

$$L_t^F(r) = \tau_t^F(r, \sigma_t(r))^{-\frac{1}{1-\alpha}} \frac{H_t(r)}{H_t(\sigma_t(r))} \left[\frac{B(r)}{B(\sigma_t(r))}\right]^{\frac{1}{1-\alpha}} L_t^F(\sigma_t(r)) \tag{34}$$

We are only left with finding the distribution of farmers across trading places since equations (33) and (34) pin down farm population at any location r *conditional* on this distribution. To obtain the farm population of trading places, I use the price index (27), the non-farm market clearing condition (29), and utility equalization. The result is stated in the following lemma.

Lemma 2. *In any period t, the distribution of farm population is the solution to the following system of equations:*

$$B(s)^{-\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}} \bar{c}_t(s)^{\frac{(\epsilon-1)^2}{(2\epsilon-1)}} = \kappa U_{c(s),t}^{-\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}.$$

$$\int_S U_{c(u),t}^{-\frac{(\epsilon-1)^2}{2\epsilon-1}} \frac{B(u)^{\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}} \left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}} \bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{(2\epsilon-1)}}}{(1-\nu)[\epsilon(1-\mu)+\mu] + (\epsilon-1)\mu\left[1-\nu\bar{c}_t(u)^{\beta-1}\right]} \left[\int_{\sigma_t^{-1}(u)} L_t^F(r)\,dr\right] \tau_t^N(u,s)^{1-\epsilon}\,du \tag{35}$$

where $L_t^F(r)$ as a function of $L_t^F(u)$ is given by (34), $\sigma_t^{-1}(u)$ is given by (33), $\bar{c}_t(s)$ is given by (8), and $\kappa = \epsilon^{1-\epsilon}(\epsilon-1)^{\epsilon-2}\mu^{-1}f^{-1}$ is a constant.

Proof. See Appendix A.6.

□

Next, non-farm population at trading places can be obtained using the distribution of farm population and farm market clearing (28). Combining (28) with (11) and (12) yields

$$L_t^N(s) = \frac{\nu}{(1-\nu)+\mu(\epsilon-1)\left[\mu+\epsilon(1-\mu)+\frac{\epsilon}{\bar{c}_t(s)^{1-\beta}-1}\right]^{-1}} \int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr \tag{36}$$

as the non-farm population of location s. Note that, by (36), the non-farm population of places to which no one ships farm goods is zero. Since the supply of farm goods is zero at these locations, non-farm production cannot take place, hence non-farm workers do not move to these locations in equilibrium. On the other hand, equation (34) implies that every location with a positive amount of land has some farm population. Pulling these results together, one can conclude that there are two types of locations in equilibrium: locations that have both farm and non-farm population and serve as trading places for farmers, and locations that only have farm population and do not serve as trading places. Whether a given location belongs to the first or the second type is an endogenous outcome shaped by the fundamentals of the model, including geography.

Finally, countries' utility levels U_{ct} can be obtained by imposing national labor market conditions (30). Hence, we have the following lemma.

Lemma 3. *Given the values of parameters, geography S, countries' population $\bar{L}_{ct}$ and growth shifters f_{ct}, functions $H_t(\cdot)$, $B(\cdot)$, $\tau_t^F(\cdot,\cdot)$ and $\tau_t^N(\cdot,\cdot)$ and the distribution of non-farm productivity $A_t(\cdot)$, the system of three equations (30), (35) and (36) determines the spatial distribution of population and countries' utility levels in period t.*

As a result of Lemma 3, obtaining the population distribution in period t requires solving the system of equations (30), (35) and (36). In the two-country world considered in this paper, I solve this system by an iteration algorithm, similar to the one applied in Nagy (2020). The algorithm consists of the following steps.

1. Calculate the non-farm productivity of Europe, using equation (23).

2. Guess country utility levels $U_{US,t}$ and U_{et}.

3. Guess the farm population trading at each location, $\int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr$.

4. Guess the farm population living at each trading location, $L_t^F(s)$.

5. Use equation (8) to calculate $\bar{c}_t(s)$. Then calculate the right-hand side of equation (35). Setting the left-hand side equal to the right-hand side, update $\bar{c}_t(s)$. Then use equation (8) again to update $L_t^F(s)$. Keep updating $L_t^F(s)$ until convergence.

6. Use the values of $L_t^F(s)$, as well as equations (33) and (34) to calculate optimal trading places and the farm population of any location r. Using these, update the farm population trading at each location, $\int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr$, and continue from step 4. Keep updating $\int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr$ until convergence.

7. Solve equation (36) for non-farm populations $L_t^N(s)$.

8. Check if national labor market clearing conditions (30) hold. If not, modify $U_{US,t}$ and U_{et}, and continue from step 3.

Although the complex structure of the model does not allow me to derive conditions under which the algorithm converges to the equilibrium distribution of population, simulation results suggest that the algorithm displays good convergence properties unless either agglomeration or dispersion forces are very strong. In particular, the algorithm always converges to the equilibrium in a broad neighborhood around the parameter values chosen in the calibration.

A.3 A special case of the model with a homogenous hinterland

This appendix section presents a special case of the model in which I can explicitly characterize how changes in geography around a location (in the form of railroad construction, an expansion of political borders, or a decrease in international trade costs) affect the size of the location's effective hinterland, $\tilde{H}_t(s)$.

The key simplification relative to the general model of Section 3 is that I assume a geographically homogenous hinterland around location s. Specifically, I assume that in a Borel measurable set of locations around s, denoted by $S_t(s) \subset S$ such that $s \in S_t(s)$,[43]

- farm productivity is homogenous: for any location $r \in S_t(s)$, $B(r) = B(s)$;
- non-farm productivity is zero everywhere except at s, guaranteeing that no other location in $S_t(s)$ specializes in non-farm production;
- farm trade costs are uniform: between s and any location $r \in S_t(s)$, $\tau_t^F(r, s) = \tau_t(s)$.

I also assume that farm trade costs between locations in $S_t(s)$ and locations in the rest of the economy are infinitely high. This guarantees that the locations in $S_t(s)$, and only those locations, serve as the farm hinterland of s.

As for non-farm trade costs, I assume that s can trade with another location e, which represents foreign markets and looks identical to s in every respect. This means that, among other things, e is also surrounded by a hinterland identical to $S_t(s)$. Trading non-farm goods between s and e is subject to an international trade cost $\tau_t^N(s, e) = b_t(s)$. Trading non-farm goods between s or e and any other location in the economy is subject to infinitely high costs. Thus, s and e ony trade non-farm goods with each other in equilibrium.

I need to impose a few further simplifications to keep analytical tractability. First, I need to assume that the share of farm goods in consumption is zero ($\nu = 1$), implying that farm goods are only used as intermediates in the non-farm sector. While a farm good share different from zero might be important for the quantitative results of this paper, the key mechanism of the model hinges on the use of farm goods in non-farm production (see the discussion in Section 3.3), which is present even when $\nu = 1$. Second, I need to assume that the economy-wide utility level, $\bar{U}_t$, is unaffected by what happens to the locations in $S_t(s)$.

[43]With a slight abuse of notation, I also use $S_t(s)$ to denote the mass of locations in set $S_t(s)$.

This assumption can be interpreted as $S_t(s)$ being small relative to the entire economy. Finally, I need to restrict the non-farm productivity of s to be low enough initially. In particular, I need to assume that $A_t(s)$ is such that non-farm firms' unit costs satisfy

$$\bar{c}_t(s)^{1-\beta} - 1 > \frac{\frac{1}{(1-\alpha)\mu} - (1-\beta)}{\epsilon - 1 - \frac{1}{(1-\alpha)\mu}}.$$

This means that s cannot be too developed in the beginning. This assumption suits locations in the U.S. West well in my historical context.

Under these assumptions, the next proposition shows that the size of s's effective hinterland increases when railroad construction, an expansion of political borders, or a fall in international trade costs affects s and its hinterland.

Proposition 2. *In the model with a homogenous hinterland, the size of the effective hinterland of location s, $\tilde{H}_t(s)$, increases if*

1. *railroad construction lowers the cost of shipping farm goods between s and its hinterland, $\tau_t(s)$, or*
2. *political border changes increase the number of locations in s's hinterland, $S_t(s)$, or*
3. *the cost of international trade between s and the rest of the world, $b_t(s)$, decreases.*

Proof. See Appendix A.6.

□

Once Proposition 2 holds, Proposition 1 implies that changes in geography in the form of railroad construction, border expansion, or increasing international trade bring more farm workers to the hinterland of s, and more non-farm workers and innovation to s itself. In other words, these changes in geography foster farm settlement, city development and growth around s. This result is stated in the following corollary.

Corollary of Proposition 2. *In the model with a homogenous hinterland, any of the three changes in geographic fundamentals presented in parts 1 to 3 of Proposition 2 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. This is the immediate consequence of Proposition 2 and the Corollary of Proposition 1.

□

A.4 Difficulties of characterizing equilibrium uniqueness

This section of the appendix builds intuition for why it is difficult to theoretically characterize the uniqueness of the period-t equilibrium in the model of Section 3. To this end, I consider two special cases of the model in which uniqueness can be characterized. Next, I turn to the general case and point out how features of the model that are not present in the special cases prevent the theoretical characterization of uniqueness. For simplicity, I restrict my attention to the single-country version of the model in this section of the appendix. Generalizing the results to the multi-country version is straightforward.

Case 1: An economy with only a farm sector. I first consider a special case of the model in which the share of non-farm goods in consumption is zero: $\nu = 0$. This implies that there is no demand for non-farm goods in the economy. Therefore, the non-farm sector is absent and does not employ any workers, implying that the population of any location r equals the location's farm population: $L_t(r) = L_t^F(r)$.

As the farm sector produces a homogeneous good under constant returns to scale, no equilibrium of this economy can feature trade by the spatial impossibility theorem (Starrett, 1978). As locations are in autarky, the utility that a farmer residing at r obtains is given by

$$U_t(r) = x_t^F(r) = B(r)\left[\frac{H_t(r)}{L_t(r)}\right]^{1-\alpha}.$$

By free mobility, $U_t(r)$ becomes equal to the utility level of the country at each location: $U_t(r) = U_t$. Substituting this into the previous equation, one can express the equilibrium population of location r at time t as

$$L_t(r) = U_t^{-\frac{1}{1-\alpha}} B(r)^{\frac{1}{1-\alpha}} H_t(r) \tag{37}$$

while the condition $\bar{L}_t = \int_S L_t(r)\, dr$ pins down the level of utility as

$$U_t = \bar{L}_t^{-(1-\alpha)}\left[\int_S B(r)^{\frac{1}{1-\alpha}} H_t(r)\, dr\right]^{1-\alpha}. \tag{38}$$

Equations (37) and (38) uniquely pin down the population of every location. Thus, in the version of the model with only a farm sector, the equilibrium is guaranteed to be unique. This is not surprising as the farm sector does not feature any agglomeration force (such as economies of scale), while it features a force of dispersion in the form of land use. If more people move to a location, the share of land that any of them can use decreases, decreasing farm output and hence utility per individual. This dispersion force prevents the emergence of multiple equilibria.

Case 2: An economy with only a non-farm sector. I now consider a special case of the model in which the share of farm goods in consumption is zero, $\nu = 1$, and farm goods are not used as intermediates in non-farm production either, $\mu = 0$. As a result, there is no demand for farm goods and farmers, and the population of any location r equals the location's non-farm population: $L_t(r) = L_t^N(r)$.

Plugging $\nu = 1$ and $\mu = 0$ into the period-t equilibrium conditions of the model, I obtain that the spatial distribution of population at time t is the solution to the system of equations

$$\bar{c}_t(s)^{\frac{(\epsilon-1)^2}{2\epsilon-1}} = \frac{\tilde{\kappa}}{\epsilon} U_t^{1-\epsilon} \int_S \bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{2\epsilon-1}} L_t(u)\, \tau_t^N(s,u)^{1-\epsilon}\, du \tag{39}$$

where $\tilde{\kappa} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1}$, $\bar{c}_t(s) = \left[1 + A_t(s)^{\beta-1}\right]^{\frac{1}{1-\beta}}$, and the level of utility, U_t, is pinned down by the national labor market clearing condition

$$\bar{L} = \int_S L_t(r)\, dr.$$

Equation (39) is a special case of the population equations considered in Allen and Arkolakis (2014), which take the form

$$f_1(s)\, L_t(s)^{\gamma_1} = k U_t^{1-\epsilon} \int_S f_2(u)\, L_t(u)^{\gamma_2}\, du \tag{40}$$

such that $f_1(s)$ and $f_2(u)$ are exogenously given continuous functions, and k is a constant. Theorem 2 in Allen and Arkolakis (2014) shows that equation (40) has a unique positive solution for the population distribution, $L_t(\cdot)$, if $\frac{\gamma_2}{\gamma_1} \in [-1, 1]$. This condition does not hold for equation (39), as $\gamma_1 = 0$ and $\gamma_2 = 1$ in that case. Although the condition $\frac{\gamma_2}{\gamma_1} \in [-1, 1]$ is not necessary for uniqueness, Allen and Arkolakis (2014) show that one can construct geographies featuring multiple equilibria if it does not hold. Hence, one can conclude that the version of the model with only a non-farm sector tends to feature multiple equilibria in general.[44]

The fact that an economy with only a non-farm sector tends to feature equilibrium multiplicity is not surprising either. The non-farm sector itself does not feature any dispersion force such as land use, but it features the classical agglomeration force of Krugman (1991), due to transport costs and economies of scale. If more people move to a location, more firms enter the location's non-farm sector, which decreases the price index of non-farm goods. Moreover, local non-farm firms face a larger demand for their product, which allows them to pay higher nominal wages. Higher nominal wages and lower non-farm prices, in turn, attract even more people to the location. This force of circular causation can lead to various equilibria that involve the concentration of population at different locations.

[44] Allen and Arkolakis (2014) also show that the standard procedure of solving the model also becomes impossible to use when the condition guaranteeing equilibrium uniqueness does not hold. This and the presence of multiple equilibria are the reasons why I contrast the model of Section 3 with DNR, rather than with a special case of the model in which I eliminate the farm sector (and, therefore, farm hinterlands).

The general case. In the previous paragraphs, I argued that the presence of the farm sector works in favor of equilibrium uniqueness, while the presence of the non-farm sector works in favor of equilibrium multiplicity. The general model of Section 3 features both sectors and as a result, intuitively, can feature either multiplicity or uniqueness. But why is the structure of this general model too complex to prevent one from theoretically characterizing when uniqueness arises? Recall that the distribution of farm population across trading places, $L_t^F(s)$, is pinned down by the equation

$$B(s)^{-\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}\left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}\bar{c}_t(s)^{\frac{(\epsilon-1)^2}{(2\epsilon-1)}}=\kappa U_t^{1-\epsilon}.$$

$$\int_S\frac{B(u)^{\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{(2\epsilon-1)}}}{(1-\nu)\left[\epsilon(1-\mu)+\mu\right]+(\epsilon-1)\mu\left[1-\nu\bar{c}_t(u)^{\beta-1}\right]}\left[\int_{\sigma_t^{-1}(u)}L_t^F(r)\,dr\right]\tau_t^N(u,s)^{1-\epsilon}\,du \tag{41}$$

while the distribution of farmers at locations r that trade at trading place $\sigma_t(r)$ is given by

$$L_t^F(r)=\tau_t^F(r,\sigma_t(r))^{-\frac{1}{1-\alpha}}\frac{H_t(r)}{H_t(\sigma_t(r))}\left[\frac{B(r)}{B(\sigma_t(r))}\right]^{\frac{1}{1-\alpha}}L_t^F(\sigma_t(r)) \tag{42}$$

where

$$\bar{c}_t(s)=\left[1+A_t(s)^{\beta-1}B(s)^{(\beta-1)\mu}\left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)(1-\beta)\mu}\right]^{\frac{1}{1-\beta}} \tag{43}$$

as a result of (8) and (32), the non-farm population of a location s is given by

$$L_t^N(s)=\frac{\nu}{(1-\nu)+\mu(\epsilon-1)\left[\mu+\epsilon(1-\mu)+\frac{\epsilon}{\bar{c}_t(s)^{1-\beta}-1}\right]^{-1}}\int_{\sigma_t^{-1}(s)}L_t^F(r)\,dr \tag{44}$$

and the level of utility, U_t, is pinned down by national labor market clearing

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

If the distribution of farm population across trading places, $L_t^F(s)$, is unique, then the farm population distribution across other locations is uniquely determined by equation (42) and the non-farm population distribution is uniquely determined by equation (44). But does equation (41) pin down a unique distribution of $L_t^F(s)$? Equation (41) differs from equation (39) in three respects:

1. In the general model, people can consume the farm good. This feature of the general model is responsible for the terms $\left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}$ and $\left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}$ appearing on the left- and right-hand sides of equation (41), respectively.

2. In the general model, the farm good can be used as an intermediate in non-farm production. This feature of the general model is responsible for the term $\left[\frac{L_t^F(r)}{H_t(r)}\right]^{(1-\alpha)(1-\beta)\mu}$ appearing in equation (43), as well as for $\bar{c}_t(u)$ appearing in the denominator on the right-hand side of equation (41).

3. Farmers trade the farm good, and hence choose which trading place they want to trade at. This feature of the general model is responsible for the term $\int_{\sigma_t^{-1}(u)} L_t^F(r)\,dr$ on the right-hand side of equation (41).

These three additional features of the general model introduce non-linearities in equations (41) and (43), which together imply that the system is no longer a special case of equation (40).[45] As a result, one can no longer use the mathematical results underlying Theorem 2 in Allen and Arkolakis (2014) and characterize uniqueness of the equilibrium distribution of farm population.

Intuitively, even though the farm and non-farm sectors solely feature forces for and against equilibrium uniqueness, respectively, their complex interconnections make a clean theoretical characterization of uniqueness infeasible. Although land use provides incentives for farmers to spread out across space, they may be driven to different regions in two distinct equilibria that feature different spatial distributions of non-farm population, and therefore different spatial distributions of demand for their product. Alternatively, even though the agglomeration force works in favor of multiple equilibria in the non-farm sector, their indirect use of land through the farm input might still lead to a unique equilibrium distribution of non-farm population. In the simulations, I find the latter to be the case.

A.5 A model with home consumption

This section of the appendix presents a version of the model of Section 3 in which farmers consume goods at their residential location, not at the trading place. For non-farm goods, which farmers purchase from others, this assumption results in an extra shipping cost that they need to incur between the trading place and their residence. For the farm good, which they produce, the assumption leads to savings on shipping costs as farmers do not need to ship the fraction of the good that they consume to the trading place. In what follows, I describe the farmer's problem, as well as the set of equilibrium conditions that change relative to Section 3.

Farmers choose their production and consumption levels, their residence and their trading place to maximize utility subject to the constraints

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

[45] With feature 1 alone, the system would still be a special case of (40), but not with features 2 and/or 3.

and

$$\int_0^{n_t} p_t^N(s,i)\,\tau_t^N(s,r)\,x_t^N(r,s,i)\,di + R_t(r)\,h_t(r) \leq p_t^F(s)\,\tau_t^F(r,s)^{-1}\left[q_t^F(r) - x_t^F(r,s)\right] + y_t(r)$$

where the definitions of all variables are the same as in Section 3. Notice the two differences between the budget constraint presented here and the one in Section 3.1. First, the farmer needs to pay the additional cost $\tau_t^N(s,r)$ of shipping non-farm varieties home from the trading place. Second, the right-hand side has $q_t^F(r) - x_t^F(r,s)$, the difference between the quantity of the farm good produced and the quantity consumed by the farmer herself.

The farmer's indirect utility is then given by

$$\begin{aligned} U_t^F(r,s) &= \frac{p_t^F(s)\,\tau_t^F(r,s)^{-1}\,q_t^F(r)}{\left[\tau_t^N(s,r)\,P_t(s)\right]^{\nu}\left[\tau_t^F(r,s)^{-1}\,p_t^F(s)\right]^{1-\nu}} \\ &= \frac{p_t^F(s)\,\tau_t^F(r,s)^{-\nu}\,\tau_t^N(r,s)^{-\nu}\,B_t(r)\left[\frac{H_t(r)}{L_t^F(r)}\right]^{1-\alpha}}{P_t(s)^{\nu}\,p_t^F(s)^{1-\nu}}. \end{aligned} \tag{2'}$$

Combining this with equations (3), (26) and (32) implies that the trading place chosen by farmers living at location r is

$$\sigma_t(r) = argmax_{s\in S_c}\left[\frac{\tau_t^N(s,r)^{\nu}}{\tau_t^F(r,s)^{1-\nu}}\right]^{-1}\varsigma_t(r,s)^{-1}\,B(s)^{-1}\,H(s)^{-(1-\alpha)}\,L_t^F(s)^{1-\alpha} \tag{33'}$$

where I normalized $\tau_t^N(s,s)$ to one, and the farm population of r is

$$L_t^F(r) = \left[\frac{\tau_t^N(\sigma_t(r),r)^{\nu}}{\tau_t^F(r,\sigma_t(r))^{1-\nu}}\right]^{-\frac{1}{1-\alpha}}\varsigma_t(r,\sigma_t(r))^{-\frac{1}{1-\alpha}}\,\frac{H_t(r)}{H_t(\sigma_t(r))}\left[\frac{B(r)}{B(\sigma_t(r))}\right]^{\frac{1}{1-\alpha}}L_t^F(\sigma_t(r)). \tag{34'}$$

As a result, equations (35) and (36) become

$$B(s)^{-\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}\left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}\bar{c}_t(s)^{\frac{(\epsilon-1)^2}{(2\epsilon-1)}} = \kappa\bar{U}_{c(s),t}^{-\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}.$$

$$\begin{aligned} &\int_S \bar{U}_{c(u),t}^{-\frac{(\epsilon-1)^2}{2\epsilon-1}}\frac{B(u)^{\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{(2\epsilon-1)}}}{(1-\nu)\left[\epsilon(1-\mu)+\mu\right]+(\epsilon-1)\mu\left[1-\nu\bar{c}_t(u)^{\beta-1}\right]}\cdot \\ &\left[\int_{\sigma_t^{-1}(u)}\frac{\tau_t^N(u,r)^{\nu}}{\tau_t^F(r,u)^{1-\nu}}L_t^F(r)\,dr\right]\tau_t^N(u,s)^{1-\epsilon}\,du \end{aligned} \tag{35'}$$

and

$$L_t^N(s) = \frac{\nu}{(1-\nu)+\mu(\epsilon-1)\left[\mu+\epsilon(1-\mu)+\frac{\epsilon}{\bar{c}_t(s)^{1-\beta}-1}\right]^{-1}} \int_{\sigma_t^{-1}(s)} \frac{\tau_t^N(s,r)^{\nu}}{\tau_t^F(r,s)^{1-\nu}} L_t^F(r)\,dr, \tag{36'}$$

respectively.

A comparison of equations (33'), (34'), (35') and (36') to their counterparts (33), (34), (35) and (36) reveals how additional shipping costs $\tau_t^N(s,r)^{\nu}$ and shipping cost savings $\tau_t^F(r,s)^{1-\nu}$ alter the equilibrium relative to Section 3. If these shipping cost changes exactly counterbalance each other, that is, $\frac{\tau_t^N(s,r)^{\nu}}{\tau_t^F(r,s)^{1-\nu}} = 1$, then the equilibrium population, productivity and utility levels of the two models become identical. This is stated formally in the next proposition.

Proposition 3 (Isomorphism with model of Section 3)**.** *Assume* $\frac{\tau_t^N(s,r)^{\nu}}{\tau_t^F(r,s)^{1-\nu}} = 1$ *for all* $r, s \in S$. *Then the model with home consumption is isomorphic in its evolution of population, productivity and utility to the model presented in Section 3.*

In the baseline calibration of the model of Section 3, we have $\nu = 0.75$ and $\tau_t^N(s,r) = \tau_t^F(r,s)^{\phi^N/\phi^F}$. We obtain the isomorphism whenever $\frac{\tau_t^F(r,s)^{0.75\phi^N/\phi^F}}{\varsigma_t(r,s)^{0.25}} = 1$, that is, $0.75\phi^N/\phi^F - 0.25 = 0$, or $\phi^N/\phi^F = 1/3$. The value of ϕ^N/ϕ^F used in the calibration is indeed close to this value. Therefore, Proposition 3 implies that the difference between the two models should be small, and changing the assumption about where consumption happens in space is unlikely to alter the results substantially.

A.6 Proofs

Proof of Lemma 1

The firm's optimal innovation decision (10) implies

$$\ell_t^I(s)^{1-\mu} = (1-\mu)^{1-\mu}(\epsilon-1)^{1-\mu} f^{1-\mu}\left[A_t(s)^{\beta-1}\left[\frac{p_t^F(s)}{w_t(s)}\right]^{(1-\beta)\mu}\bar{c}_t(s)^{\beta-1}\right]^{1-\mu}$$

from which, using (8), we obtain

$$\ell_t^I(s)^{1-\mu} = (1-\mu)^{1-\mu}(\epsilon-1)^{1-\mu} f^{1-\mu}\left[1-\bar{c}_t(s)^{\beta-1}\right]^{1-\mu}. \tag{45}$$

Equation (36) implies that the unit cost per wage at s can be written as

$$\bar{c}_t(s) = \left[\frac{\nu\left(1+\frac{L_t^N(s)}{\int_{\sigma_t^{-1}(s)} L_t^F(r)dr}\right)}{\left(\nu+\frac{(1-\nu)\epsilon}{\mu(\epsilon-1)}\right)\frac{L_t^N(s)}{\int_{\sigma_t^{-1}(s)} L_t^F(r)dr} - \nu\left(\frac{\epsilon}{\mu(\epsilon-1)}-1\right)}\right]^{\frac{1}{1-\beta}}.$$

Plugging this into equation (45) and rearranging yields the result.

Proof of Lemma 2

Equation (27) provides the price index at any trading place s. Combining it with equations (6) and (12) yields

$$P_t(s)^{1-\epsilon} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1} \int_S \frac{\bar{c}_t(u)^{1-\epsilon}}{(\epsilon-1)\left(\mu\bar{c}_t(u)^{\beta-1}+1-\mu\right)+1} w_t(u)^{1-\epsilon} L_t^N(u)\,\tau_t^N(u,s)^{1-\epsilon}\,du. \tag{46}$$

Alternatively, one can express the price index at s from equation (3) as

$$P_t(s) = U_{c(s),t}^{-\frac{1}{\nu}} B(s)^{\frac{1-\nu}{\nu}} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}} w_t(s)$$

where I used equations (26) and (32). Plugging this into the left-hand side of equation (46) implies

$$U_{c(s),t}^{\frac{\epsilon-1}{\nu}} B(s)^{-(\epsilon-1)\frac{1-\nu}{\nu}} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)(\epsilon-1)\frac{1-\nu}{\nu}} w_t(s)^{1-\epsilon} =$$

$$\tilde{\kappa}\int_S \frac{\bar{c}_t(u)^{1-\epsilon}}{(\epsilon-1)\left(\mu\bar{c}_t(u)^{\beta-1}+1-\mu\right)+1} w_t(u)^{1-\epsilon} L_t^N(u)\,\tau_t^N(u,s)^{1-\epsilon}\,du \tag{47}$$

where $\tilde{\kappa} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1}$.

Equation (29) provides the market clearing condition for each non-farm good at any trading place s. Combining it with equations (6) and (7) yields

$$\bar{c}_t(s)^{\epsilon-1} w_t(s)^{\epsilon} = \nu\epsilon^{-\epsilon}(\epsilon-1)^{\epsilon-1} f^{-1} \int_S P_t(u)^{\epsilon-1} I_t(u)\,\tau_t^N(s,u)^{1-\epsilon}\,du \tag{48}$$

where

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

is the sum of farmers' and non-farm workers' income at trading place u. Equations (2) and (24) allow me to rewrite income as

$$I_t(u) = p_t^F(u)\,B(u)\left[\frac{H_t(u)}{L_t^F(u)}\right]^{1-\alpha} \int_{\sigma_t^{-1}(u)} L_t^F(r)\,dr + w_t(u)\,L_t^N(u)$$

from which, by equation (32), we obtain

$$I_t(u) = w_t(u)\left[\int_{\sigma_t^{-1}(u)} L_t^F(r)\,dr + L_t^N(u)\right].$$

Also, combining (28) with (11) and (12) yields

$$L_t^N(s) = \frac{\nu}{(1-\nu) + \mu(\epsilon - 1)\left[\mu + \epsilon(1-\mu) + \frac{\epsilon}{\bar{c}_t(s)^{1-\beta} - 1}\right]^{-1}} \int_{\sigma_t^{-1}(s)} L_t^F(r)\, dr$$

hence income can be written as

$$I_t(u) = w_t(u) \frac{\nu^{-1}\epsilon}{(\epsilon-1)\left(\mu c_t(u)^{\beta-1} + 1 - \mu\right) + 1} L_t^N(u)$$

and equation (48) can be written as

$$\bar{c}_t(s)^{\epsilon-1} w_t(s)^{\epsilon} = \tilde{\kappa} \int_S U_{c(u),t}^{-\frac{\epsilon-1}{\nu}} \frac{B(u)^{(\epsilon-1)\frac{1-\nu}{\nu}} \left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)(\epsilon-1)\frac{1-\nu}{\nu}}}{(\epsilon-1)\left(\mu \bar{c}_t(u)^{\beta-1} + 1 - \mu\right) + 1} w_t(u)^{\epsilon} L_t^N(u)\, \tau_t^N(s,u)^{1-\epsilon}\, du \tag{49}$$

where $\tilde{\kappa} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1}$, and I used equations (3), (26) and (32) to substitute for the price index on the right-hand side.

In what follows, I show that equations (47) and (49) reduce to a single equation. To see this, note first that, since non-farm shipping costs are symmetric, $\tau_t^N(s,u) = \tau_t^N(u,s)$. Second, guess that wages at location s take the form

$$w_t(s) = U_{c(s),t}^{\iota_1} B(s)^{\iota_2} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{\iota_3} \bar{c}_t(s)^{\iota_4}$$

where ι_1, ι_2, ι_3 and ι_4 are constants. Plugging the guess into (47) and (49), one obtains that both of these equations hold if and only if $\iota_1 = \frac{\epsilon-1}{\nu(2\epsilon-1)}$, $\iota_2 = -\frac{1-\nu}{\nu}\frac{\epsilon-1}{2\epsilon-1}$, $\iota_3 = (1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon-1}{2\epsilon-1}$, and $\iota_4 = -\frac{\epsilon-1}{2\epsilon-1}$. Thus, wages at s can be written as

$$w_t(s) = U_{c(s),t}^{\frac{\epsilon-1}{\nu(2\epsilon-1)}} B(s)^{-\frac{1-\nu}{\nu}\frac{\epsilon-1}{2\epsilon-1}} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{(1-\alpha)\frac{1-\nu}{\nu}\frac{\epsilon-1}{2\epsilon-1}} \bar{c}_t(s)^{-\frac{\epsilon-1}{2\epsilon-1}}$$

and equations (47) and (49) reduce to equation (35).

Proof of Proposition 1

Rearranging equation (36) yields location s's effective hinterland size as

$$\tilde{H}_t(s) = \frac{\int_{\sigma_t^{-1}(s)} L_t^F(r)\, dr}{L_t^N(s)} = \frac{1}{\nu}\left[(1-\nu) + \mu(\epsilon-1)\left(\mu + \epsilon(1-\mu) + \frac{\epsilon}{\bar{c}_t(s)^{1-\beta} - 1}\right)^{-1}\right]$$

from which, by rearranging,

$$\bar{c}_t(s)^{1-\beta} - 1 = \frac{\epsilon}{\frac{\mu(\epsilon-1)}{\nu\tilde{H}_t(s)-(1-\nu)} - \mu - \epsilon(1-\mu)}. \tag{50}$$

To show part i of the proposition, combine equation (50) with (8) and (32) to get

$$L_t^F(s) = \left[\frac{\epsilon}{\frac{\mu(\epsilon-1)}{\nu\tilde{H}_t(s)-(1-\nu)} - \mu - \epsilon(1-\mu)}\right]^{\frac{1}{(1-\alpha)(1-\beta)\mu}} A_t(s)^{\frac{1}{(1-\alpha)\mu}} B(s)^{\frac{1}{1-\alpha}} H_t(s) \tag{51}$$

which implies that the farm population of s is increasing in $\tilde{H}_t(s)$ as

$$\frac{dL_t^F(s)}{d\tilde{H}_t(s)} = \frac{1}{(1-\alpha)(1-\beta)\mu}\left[\frac{\epsilon}{\frac{\mu(\epsilon-1)}{\nu\tilde{H}_t(s)-(1-\nu)} - \mu - \epsilon(1-\mu)}\right]^{\frac{1}{(1-\alpha)(1-\beta)\mu}-1} A_t(s)^{\frac{1}{(1-\alpha)\mu}} \cdot$$
$$B(s)^{\frac{1}{1-\alpha}} H_t(s) \frac{\epsilon\frac{\mu(\epsilon-1)}{\left(\nu\tilde{H}_t(s)-(1-\nu)\right)^2}\nu}{\left(\frac{\mu(\epsilon-1)}{\nu\tilde{H}_t(s)-(1-\nu)} - \mu - \epsilon(1-\mu)\right)^2} > 0.$$

The farm population of any other location r in s's hinterland is also increasing in $\tilde{H}_t(s)$ since, by equation (34),

$$\frac{dL_t^F(r)}{d\tilde{H}_t(s)} = \tau_t^F(r,s)^{-\frac{1}{1-\alpha}} \frac{H_t(r)}{H_t(s)}\left[\frac{B(r)}{B(s)}\right]^{\frac{1}{1-\alpha}} \frac{dL_t^F(s)}{d\tilde{H}_t(s)} > 0.$$

To show part ii of the proposition, use the definition of $\tilde{H}_t(s)$ to obtain

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

then use (34) and (51) to get

$$L_t^N(s) = \tilde{H}_t(s)^{-1}\left[\frac{\epsilon}{\frac{\mu(\epsilon-1)}{\nu\tilde{H}_t(s)-(1-\nu)} - \mu - \epsilon(1-\mu)}\right]^{\frac{1}{(1-\alpha)(1-\beta)\mu}} A_t(s)^{\frac{1}{(1-\alpha)\mu}} \cdot$$
$$\int_{\sigma_t^{-1}(s)} \tau_t^F(r,s)^{-\frac{1}{1-\alpha}} B(r)^{\frac{1}{1-\alpha}} H_t(r)\,dr$$

from which

$$L_t^N(s) = \left[\frac{\epsilon}{\frac{\mu(\epsilon-1)}{\nu \tilde{H}_t(s)^{1-(1-\alpha)(1-\beta)\mu} - (1-\nu)\tilde{H}_t(s)^{-(1-\alpha)(1-\beta)\mu}} - (\mu + \epsilon(1-\mu)) \tilde{H}_t(s)^{(1-\alpha)(1-\beta)\mu}} \right]^{\frac{1}{(1-\alpha)(1-\beta)\mu}} \cdot A_t(s)^{\frac{1}{(1-\alpha)\mu}} \int_{\sigma_t^{-1}(s)} \tau_t^F(r,s)^{-\frac{1}{1-\alpha}} B(r)^{\frac{1}{1-\alpha}} H_t(r)\, dr \tag{52}$$

which implies $\frac{dL_t^N(s)}{d\tilde{H}_t(s)} > 0$ since $0 < (1-\alpha)(1-\beta)\mu < 1$. Thus, the non-farm population of s is increasing in $\tilde{H}_t(s)$.

To show part iii of the proposition, apply the definition of $\tilde{H}_t(s)$ in equation (22) to get

$$\ell_t^I(s)^{1-\mu} = \rho \left[\left(1 + \tilde{H}_t(s)^{-1}\right)^{-1} - (1-\nu) \right]^{1-\mu} \tag{53}$$

which implies

$$\frac{d\ell_t^I(s)^{1-\mu}}{d\tilde{H}_t(s)} = (1-\mu)\rho^{\frac{1}{1-\mu}} \ell_t^I(s)^{-\mu} \frac{\tilde{H}_t(s)^{-2}}{\left(1 + \tilde{H}_t(s)^{-1}\right)^2} > 0.$$

That is, innovation at s is increasing in $\tilde{H}_t(s)$ as well.

Proof of Proposition 2

In the special case of the model presented in Section A.3, equations (8) and (32) together imply

$$L_t^F(s) = \left[\bar{c}_t(s)^{1-\beta} - 1\right]^{\frac{1}{(1-\alpha)(1-\beta)\mu}} A_t(s)^{\frac{1}{(1-\alpha)\mu}} B(s)^{\frac{1}{1-\alpha}} \tau_t(s)^{-\frac{1}{1-\alpha}} H_t(s) \tag{54}$$

while equation (35) simplifies to

$$\bar{c}_t(s)^{\frac{(\epsilon-1)^2}{2\epsilon-1}} = \kappa \bar{U}_t^{1-\epsilon} \left(1 + b_t(s)^{1-\epsilon}\right) \frac{\bar{c}_t(s)^{-\frac{\epsilon(\epsilon-1)}{2\epsilon-1}}}{(\epsilon-1)\mu\left[1 - \bar{c}_t(s)^{\beta-1}\right]} S_t(s) L_t^F(s). \tag{55}$$

Plugging (54) into (55), rearranging and taking logs yields

$$\begin{aligned}(\epsilon-1)\log \bar{c}_t(s) = &\log\left(\frac{\kappa}{(\epsilon-1)\mu}\right) + (1-\epsilon)\bar{U}_t + \log\left(1 + b_t(s)^{1-\epsilon}\right) - \log\left[1 - \bar{c}_t(s)^{\beta-1}\right] \\ &+ \frac{1}{(1-\alpha)(1-\beta)\mu}\log\left[\bar{c}_t(s)^{1-\beta} - 1\right] + \log S_t(s) + \frac{1}{(1-\alpha)\mu}\log A_t(s) \\ &+ \frac{1}{1-\alpha}\log B(s) - \frac{1}{1-\alpha}\log \tau_t(s) + \log H_t(s).\end{aligned} \tag{56}$$

To show part 1 of the proposition, take the derivative of (56) with respect to $\log \tau_t(s)$:

$$(\epsilon-1)\frac{d\log\bar{c}_t(s)}{d\log\tau_t(s)}=-\frac{(1-\beta)\,\bar{c}_t(s)^{\beta-1}}{1-\bar{c}_t(s)^{\beta-1}}\frac{d\log\bar{c}_t(s)}{d\log\tau_t(s)}+\frac{1}{(1-\alpha)\,\mu}\frac{\bar{c}_t(s)^{1-\beta}}{\bar{c}_t(s)^{1-\beta}-1}\frac{d\log\bar{c}_t(s)}{d\log\tau_t(s)}-\frac{1}{1-\alpha}$$

Rearranging yields

$$\frac{d\log\bar{c}_t(s)}{d\log\tau_t(s)}=-\frac{1}{(1-\alpha)\left(\epsilon-1-\frac{1}{(1-\alpha)\mu}-\frac{\frac{1}{(1-\alpha)\mu}-(1-\beta)}{\bar{c}_t(s)^{1-\beta}-1}\right)}<0$$

where the inequality comes from the assumption $\bar{c}_t(s)^{1-\beta}-1>\frac{\frac{1}{(1-\alpha)\mu}-(1-\beta)}{\epsilon-1-\frac{1}{(1-\alpha)\mu}}$. Taking the derivative of (50) with respect to $\log\bar{c}_t(s)$ and rearranging, we also obtain

$$\frac{d\log\tilde{H}_t(s)}{d\log\bar{c}_t(s)}=\frac{(1-\beta)\,\bar{c}_t(s)^{1-\beta}\left(\mu(\epsilon-1)\,\tilde{H}_t(s)^{-1}-\mu-\epsilon(1-\mu)\right)^2}{\mu\epsilon(\epsilon-1)\,\tilde{H}_t(s)^{-1}}>0.$$

Pulling the two derivatives together, we get, by the chain rule,

$$\frac{d\log\tilde{H}_t(s)}{d\log\tau_t(s)}=\frac{d\log\tilde{H}_t(s)}{d\log\bar{c}_t(s)}\frac{d\log\bar{c}_t(s)}{d\log\tau_t(s)}<0$$

which proves that the size of the effective hinterland is decreasing in $\tau_t(s)$. This is part 1 of the proposition.

The proof of parts 2 and 3 is analogous, with the exception that one needs to take the derivative of (56) with respect to $\log S_t(s)$ and $\log b_t(s)$, respectively.

Proof of Corollary of Proposition 1

In the case of international trade costs, the result is an immediate consequence of Proposition 1 due to equations (51), (52) and (53).

In the case of railroad construction, we also need to take into account the direct effect it has on $\tau_t^F(r,s)$ in equations (51), (34) and (52). However, since railroads lower all $\tau_t^F(r,s)$, this direct effect amplifies the indirect effect that railroads have through increasing effective hinterland size. Hence, $L_t^F(r)$, $L_t^N(s)$ and $\ell_t^I(s)$ all increase.

In the case of the expansion of political borders, there might be a direct effect on the set of locations in s's hinterland, $\sigma_t^{-1}(s)$, and thus on $L_t^N(s)$. But since $\sigma_t^{-1}(s)$ cannot shrink by the fact that political borders expand – as well as by equation (33) –, we again get a direct effect that amplifies the indirect effect that border expansion has through increasing effective hinterland size. Hence, $L_t^F(r)$, $L_t^N(s)$ and $\ell_t^I(s)$ all increase in that case as well.

B Data appendix

This appendix describes the datasets used to document the major patterns of pre-Civil War U.S. urban history, to take the model to the data, and to evaluate the model's fit. I use geographic data on the location of the sea, navigable rivers, canals and lakes, as well as railroads to calculate shipping costs and to quantify the importance of trading routes in city location. I use census data on county, city and town locations and populations to calibrate the model and to evaluate how well the model fits the evolution of population seen in the data.

My unit of observation is a *cell* in a 20 by 20 arc minute grid of the United States. I create this grid of the U.S. using Geographic Information Software (GIS), and combine it with other sources of geographic data to determine whether any given cell is at a waterway or at a railroad, and to calculate the agricultural productivity, natural amenities, and population of the cell. In what follows, I provide additional details on this procedure.

Waterways. I use the ESRI Map of U.S. Major Waters and the 20 by 20 arc minute grid of the U.S. to determine whether a grid cell is at the sea. In particular, I regard a cell as being at the sea if a positive fraction of its area is in the sea.

I follow Donaldson and Hornbeck's (2016) definition of navigable rivers, lakes and canals, who, in turn, borrow the definition from Fogel (1964). Combining the definition with the ESRI Map of U.S. Major Waters and the 20 by 20 arc minute grid of the U.S., I classify each cell based on whether it contains a navigable waterway. As canals were gradually constructed during the 19th century, I do this classification of cells separately for every time period t, using the set of canals that were already open at t. Table 10 provides a list of navigable canals, along with their locations and opening dates.

Railroads. The website *http://oldrailhistory.com* includes maps of the U.S. railroad network in 1835, 1840, 1845, 1850 and 1860.[46] I georeference these maps to the 20 by 20 arc minute grid of the U.S., and classify each cell depending on whether it contained some railroads in any given period t between 1835 and 1860.[47]

Agricultural productivity. I collect high-resolution data on agricultural yields from the Food and Agriculture Organization's Global Agro-Ecological Zones database (FAO GAEZ). This database contains the potential yield of various crops at a 5 by 5 arc minute spatial resolution, under different irrigation and input conditions. These yields come in the form of an index whose value, in its raw form, ranges between 0 and 10,000 for each

[46] Although the first railroads started to be built in the late 1820s, there only existed a small number of short and disconnected segments in 1830. Therefore, it is reasonable to assume that no U.S. location had access to a rail network until 1830.

[47] Lacking the map of the network in 1855, I need to approximate it by the network in 1850.

Table 10: Navigable canals constructed between 1790 and 1860

Year of opening	Canal name	Canal was constructed to connect...
1823	Champlain Canal	Lake Champlain and Hudson River
1825	Erie Canal	Lake Erie and Hudson River
1827	Schuylkill Canal	Port Carbon, PA and Philadelphia
1828	Erie Canal, Oswego branch	Erie Canal and Lake Ontario
1828	Union Canal	Middletown, PA and Reading, PA
1828	Delaware and Hudson Canal	Delaware River and Hudson River
1828	Farmington Canal	New Haven, CT and interior of Connecticut
1828	Blackstone Canal	Worcester, MA and Providence, RI
1829	Chesapeake and Delaware Canal	Chesapeake Bay and Delaware River
1831	Morris Canal	Phillipsburg, NJ and Jersey City, NJ
1832	Ohio and Erie Canal	Lake Erie and Ohio River
1832	Pennsylvania Canal System, Delaware Division	Easton, PA and Bristol, PA
1832	Cumberland and Oxford Canal	lakes in Southern Maine and Portland, MA
1834	Delaware and Raritan Canal	Delaware River and New Brunswick, NJ
1834	Chenango Canal	Binghamton, NY and Utica, NY
1835	Pennsylvania Canal System	several rivers and canals in Pennsylvania
1840	Susquehanna and Tidewater Canal	Wrightsville, PA and Chesapeake Bay
1840	Pennsylvania and Ohio Canal	Ohio and Erie Canal and Beaver and Erie Canal
1840	James River and Kanawha Canal	Lynchburg, VA and Richmond, VA
1841	Genesee Valley Canal	Dansville, NY and Rochester, NY
1844	Beaver and Erie Canal	Lake Erie and Ohio River
1845	Miami and Erie Canal	Lake Erie and Cincinatti, OH
1847	Whitewater Canal	Ohio River and Lawrenceburg, IN
1848	Illinois and Michigan Canal	Lake Michigan and Illinois River
1848	Wabash and Erie Canal, section 1	Miami and Erie Canal and Terre Haute, OH
1848	Sandy and Beaver Canal	Ohio and Erie Canal and Ohio River
1850	Chesapeake and Ohio Canal	Cumberland, MD and Washington, D.C.
1853	Wabash and Erie Canal, section 2	Ohio River and Terre Haute, OH
1855	Black River Canal	Black River and Erie Canal
1858	Chemung and Junction Canals	Erie Canal and Pennsylvania Canal

Table 11: Productivity of the six main U.S. crops

Name of crop	Minimum	Maximum	Mean	Standard deviation
Cereals	0	1.716	1.481	0.303
Cotton	0	1.708	0.627	0.739
Sugar cane	0	1.715	0.138	0.427
Sweet potato	0	1.716	0.320	0.604
Tobacco	0	1.710	1.060	0.657
White potato	0	1.708	1.373	0.376

The four columns of this table show the minimum, the maximum, the unweighted mean and the unweighted standard deviation of grid cell-level yields for the six major crops in the United States. Filters "low-input level" and "rain-fed" have been applied for each crop. Source: FAO GAEZ.

crop-location pair. This makes the yields comparable across crops and across locations. To provide the best possible approximation to 19th-century productivity, I calculate the yields under the assumption of no irrigation and low input levels. I use data on the potential yields of cereals, cotton, sugar cane, sweet potato, tobacco, and white potato.[48] I aggregate the data to the 20 by 20 arc minute level by calculating the average productivity of each crop within each 20 by 20 minute cell. Table 11 provides summary statistics of productivity for each crop.

Natural amenities. The FAO GAEZ dataset also includes data on natural amenities as they heavily influence agricultural yields. I select five climate variables that are the closest to standard measures of natural amenities in the literature (see, for instance, Desmet and Rossi-Hansberg, 2013): the mean annual temperature, the annual temperature range, the number of days with minimum temperature below 5 °C, the number of days with mean temperature above 10 °C, and the annual precipitation.[49] To be as close as possible to 19th-century conditions, I use the earliest data available (1961 to 1990 for the annual temperature range, and 1960 for the other variables). I aggregate the variables to the 20 by 20 arc minute level using the same procedure as the one used for productivity. Table 12 provides summary statistics of the natural amenity variables.

County, city and town populations. The National Historical Geographic Information System (NHGIS) provides census data on county populations for 1790, 1800, 1810, 1820, 1830, 1840, 1850 and 1860, along with maps of county boundaries.[50] I use the county

[48] According to the 1860 Census of Agriculture, these were the six major crops grown in the United States.

[49] Although these variables are close to the ones considered in the literature, they do not exactly coincide with them. Therefore, as a robustness check, I collect the climate variables used in Desmet and Rossi-Hansberg (2013) from *weatherbase.com* for a 845-element subset of U.S. grid cells. Using these alternative variables does not alter the results substantially. These results are available from the author upon request.

[50] The database is available at *nhgis.org*. Source: Minnesota Population Center. *National Historical Geographic Information System: Version 2.0.* Minneapolis, MN: University of Minnesota 2011.

Table 12: Natural amenity variables

Variable	Minimum	Maximum	Mean	Standard deviation
Mean annual temperature (°C)	-2.1	24.4	10.9	5.1
Annual temperature range (°C)	5.3	37.4	24.6	5.4
Number of days with minimum temperature below 5 °C	0	251	111.2	61.2
Number of days with mean temperature above 10 °C	23.4	365	202.4	58.4
Annual precipitation (mm)	44.8	2634.8	695.2	417.4

The four columns of this table show the minimum, the maximum, the unweighted mean and the unweighted standard deviation of grid-cell level natural amenity variables for the United States. All data are for 1960, except annual temperature range which is for the period between 1961 and 1990. Source: FAO GAEZ.

population data to calculate the population of each 20 by 20 arc minute grid cell. For each census year, I transform the county map into a raster of 2 by 2 arc minute cells, and allocate the population of each county equally across the small cells it occupies. Next, I calculate the population of each 20 by 20 minute cell by summing the population levels of 2 by 2 minute cells inside the cell. Finally, I obtain city and town populations from a census database that provides the population of settlements above 2,500 inhabitants in each census year,[51] while I use Google Maps to determine the geographic location of each town and city.

Large regions. Based on the boundaries of U.S. states today, I assign each 20 by 20 arc minute grid cell to the state to which its centroid belongs. Next, I assign the cell to one of the four large U.S. regions: the Northeast, the South, the Midwest or the West, following the mapping of states to regions in Caselli and Coleman (2001). Therefore, the Northeast constitutes of Connecticut, Delaware, Massachusetts, Maryland, Maine, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island and Vermont; the South includes Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia and West Virginia; the Midwest constitutes of Iowa, Illinois, Indiana, Kansas, Michigan, Minnesota, Missouri, North Dakota, Nebraska, Ohio, South Dakota and Wisconsin; and the remaining states belong to the West.

U.S. land. I also use the NHGIS database to calculate the fraction of cells that is covered by land *and* is part of the U.S. in any census year. In particular, I determine whether each 2 by 2 arc minute cell was part of U.S. territory in census year t. Then I calculate the land area of each 20 by 20 minute cell as the fraction of 2 by 2 minute cells inside the 20 by 20 minute cell that were part of U.S. territory at t.

For periods between census years (1795, 1805, 1815, 1825, 1835, 1845 and 1855), I use

[51] This database is available at *census.gov/population/www/documentation/twps0027/twps0027.html*.

the fact that no significant border change took place between 1790 and 1800, between 1805 and 1815, between 1825 and 1840, and between 1855 and 1860. Therefore, I can use the 1790 (or the 1800) distribution of land in 1795, the 1810 distribution in 1805 and 1815, the 1830 (or 1840) distribution in 1825 and 1835, and the 1860 distribution in 1855. This leaves me with the task of obtaining the distribution in 1845. To accomplish this, I georeference a map showing 1845 borders to the 20 by 20 minute grid, and determine whether the centroid of each grid cell was in U.S. territory in 1845.

C Slow migration to the West

C.1 Setup

The slow settlement of the West could be another reason for the emergence of cities at trading routes. To see if such a slow migration process could quantitatively generate the disproportionate emergence of cities at trading routes, I consider a simple process in which a resident of location j moves to location i next period with probability $\pi_{ji} \in [0, 1]$. That is, I assume that the population distribution evolves according to the equation

$$\ell_{i,t+1} = \nu_{t+1} \sum_{S_t} \ell_{jt} \pi_{ji}$$

where i and j index locations, S_t denotes the set of U.S. locations in period t, and $\ell_{i,t}$ is the population (per unit of land) of location i at time t. ν_{t+1} is a number that guarantees that population levels sum to total population in period $t + 1$.

To generate slow transitions, I assume that the probability of moving from j to i is negatively related to the moving cost between these two locations:

$$\pi_{ji} = e^{-m_{ji}}$$

where m_{ji} is the cost of moving from j to i. Moving from j to i requires passing through a set of locations that connect i to j. I denote the cost of passing through a specific location k by $\xi_k > 0$. This means that the total cost of moving from j to i is

$$m_{ji} = \sum_{g_{ji}} \xi_k$$

where g_{ji} denotes the least-cost route from j to i. Note that this formulation implies that staying put is costless but moving to another location is costly. As a result, people mostly stay where they have lived in the past and occupy previously uninhabited Western regions at a slow pace. The speed of the transition is driven by the magnitude of the moving cost parameters ξ_k.

I simulate this benchmark process of slow migration to see if it is able to replicate the disproportionately frequent emergence of population clusters at water, confluences or railroads. Simulation of the process requires choosing the length of a time period, defining the set of locations for each period, choosing an initial (period 1) distribution of population across locations, and calibrating the values of moving cost parameters ξ_k. I choose the length of a period to be five years.[52] I define the set of locations in period t as the set of 20 by 20 arc minute grid cells that were part of the U.S. in period t.[53] To obtain the initial distribution of population, I assign the population of each county from the 1790 census to the grid cells it occupies, based on the share of land belonging to each cell.

For the key parameters ξ_k, I consider two calibrations. In one calibration, I make the simplest possible assumption and set ξ_k equal across locations: $\xi_k = \xi$. This implies that the cost of moving between two locations is proportional to the geographic distance between them. As the magnitude of ξ drives the speed of transition, I choose the value of ξ such that the mean center of population moves as much to the West in the model as in the data. This implies $\xi = 0.016$. Under this value of moving costs, the probability that a person moves 50 miles from her current residence next period is 45% of the probability that she does not move.

One concern about the assumption of equal moving costs is that migration may have been, similarly to trade, less costly along water routes than inland. I address this issue in the second calibration by making moving costs heterogeneous across locations. In particular, I assume that ξ_k equals ξ^W if location k is located at a waterway, and ξ^I otherwise. Since I have two structural parameters now, I need to calibrate them to two moments in the data. To find the right data moments, note that the speed of the transition now depends on the *average* of these costs. On the other hand, the *difference* between inland and water costs drives the extent to which people locate near rivers, hence the concentration of population in space. Therefore, I calibrate average costs $\frac{\xi^W + \xi^I}{2}$ to match the movement in the mean center of population by 1860, while I calibrate the difference in costs $\xi^I - \xi^W$ to match the concentration of population, measured by Theil index, in 1860. The resulting estimates are $\xi^W = 0.01575$ and $\xi^I = 0.01725$. This implies that moving 50 miles inland decreases the probability of moving (and, therefore, the number of migrants) by 6 percent more than moving 50 miles along water.[54]

Given that cities appeared in 75 grid cells in the data between 1790 and 1860, I define cities in the simulated data as the 75 grid cells with largest population in 1860. Then I calculate the fraction of simulated cities at water, confluences and railroads. Next, I control for the amenities and productivity of each location and estimate equation (1) on

[52] Using a different period length does not alter the results substantially.

[53] That is, I incorporate political border changes between 1790 and 1860 by allowing the set of inhabitable locations to change across periods.

[54] To find the least-cost routes in an efficient way, I apply the Fast Marching Algorithm for any parameter vector $\left(\xi^W, \xi^I\right)$.

Table 13: Slow migration process: simulation results

Fraction of cities at...	Data	Slow migration process Equal cost calibration	Heterogeneous cost calibration
water	98.6%	49.3%	62.7%
confluence	85.5%	33.3%	48.0%
railroad	95.7%	60.0%	58.7%

Dependent variable: newcity

	Data (1)	(2)	(3)	Equal cost calibration (4)	(5)	(6)	Heterogeneous cost calibration (7)	(8)	(9)
water	0.018**			-0.006			0.003		
	(0.002)			(0.004)			(0.003)		
confluence		0.029**			-0.004			0.008	
		(0.004)			(0.004)			(0.004)	
railroad			0.030**			0.008			0.007
			(0.004)			(0.004)			(0.004)
Prod & amenities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of observations	7641	7641	7641	7641	7641	7641	7641	7641	7641

The first column of the top table shows the fraction of 1860 U.S. cities that formed in 20 by 20 arc minute grid cells located at water, at a confluence, or at a railroad in the data (also reported in Table 1). The second and third columns show the same numbers in the slow migration model of Appendix C. "Equal cost" refers to the calibration in which moving costs are equal across locations; "heterogeneous cost" refers to the calibration in which moving costs are lower along water than inland. The bottom table presents the results of estimating equation (1) on actual data as well as on simulated data coming from the slow migration model. 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 a city formed in the cell by 1860 in the data (columns 1 to 3), or in the simulation (columns 4 to 9). 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 & 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 3 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 5%; **: significant at 1%. Source: U.S. census, ESRI Map of U.S. Major Waters, *oldrailhistory.com* and FAO GAEZ.

the simulated data.

C.2 Simulation results

Table 13 presents the results and confirms that the process of slow migration is unable to replicate the disproportionately high share of cities near trading routes. In particular, the top panel of Table 13 shows that the fraction of cities forming near a waterway, a confluence or a railroad is substantially lower under both calibrations than in the data.

The bottom panel presents the results of estimating equation (1) on the simulated data (columns 4 to 9), and contrasts them with the estimates coming from the actual data (columns 1 to 3). As columns (1) to (3) show, the probability that a city appears at water, confluences or railroads conditional on amenities and productivity is significantly higher in the data than the probability that it appears elsewhere. This is never the case in the simulated data, no matter whether moving costs are only a function of geographic distance (columns 4 to 6), or are lower along water routes than inland (columns 7 to 9).

To sum up, the slow migration process presented in this appendix is unable to replicate the disproportionate emergence of cities near good trading opportunities. To match these facts in the data, it seems essential to have a framework in which incentives to trade attract people to these locations.

D The DNR model

In this appendix, I develop an extension of the DNR model that I can use as a benchmark to which I compare the model of Section 3. Relative to Desmet et al. (2018), I make three necessary but minor modifications that make the DNR model comparable to the model of Section 3. First, I allow the exogenous supply of land to change over time, so that I can incorporate changes in U.S. political borders.[55] Second, I allow total world population to grow as population growth is a prominent feature of the data. Finally, I allow shipping costs across locations to change over time as the U.S. transportation network develops. Appendix D.1 provides a detailed description of this model.

To take the DNR model to the data, I apply a procedure that follows, as closely as possible, the calibration strategy developed in Desmet et al. (2018). This calibration strategy involves matching the initial (1790) population distribution to recover location-specific productivity levels. It also involves matching the changes in population between 1790 and the next observed year (1800) to back out the levels of moving costs across locations. Intuitively, a location toward which people move is, everything else fixed, identified as a location with low costs of entry. Importantly, the calibration leaves the evolution of the population distribution after 1800 untargeted. As a result, I can simulate the DNR model until 1860, and compare how well it can replicate the evolution of the population distribution and the locations of cities to the model of Section 3. Appendix D.2 describes the procedure of taking the DNR model of the data. Appendix D.3 provides details on the simulation, while Appendix D.4 discusses the simulation results. Appendix D.5 provides robustness.

D.1 Setup

The setup is identical to the one in Desmet, Nagy and Rossi-Hansberg (2018), with three necessary but minor modifications that make the model comparable to the model of Section 3. First, I allow the exogenous supply of land to change over time, so that I can incorporate changes in U.S. political borders that made land in the West available. Hence, I assume that the world is a subset S of a two-dimensional surface, and the supply of land at location $r \in S$ is given by $H_t(r)$, potentially changing across periods $t \in \{0, 1, ...\}$. Second, I allow total world population to grow as population growth is a prominent feature of the data. Hence, an exogenous number of $\bar{L}_t$ agents occupy the world in period t, and each of them supplies one unit of labor. Third, I allow shipping costs across locations to change over time as the transportation network develops.

[55] Desmet et al. (2021) also consider an extension of the DNR framework with time-varying land supply.

The rest of the model is the same as Desmet et al. (2018). Most importantly, and unlike in the model of Section 3, the economy in DNR is assumed to consist of one sector that produces a continuum of differentiated varieties $\omega \in [0, 1]$.

D.1.1 Consumption and location choice

Each agent derives utility from consuming the varieties produced by the single sector, from consuming amenities at her location and from her idiosyncratic taste for the location. Her utility is also affected by the costs of moving across locations in the past. As a result, the utility of agent i living at r in period t and having a history of past locations $\{r_0, ..., r_{t-1}\}$ is given by

$$u_t^i(r_0, ..., r_{t-1}, r) = a_t(r) \left[\int_0^1 c_t^\omega(r)^\rho \, d\omega \right]^{\frac{1}{\rho}} \varepsilon_t^i(r) \prod_{s=1}^{t} m(r_{s-1}, r_s)^{-1}$$

where $a_t(r)$ is the level of amenities at r, $c_t^\omega(r)$ is the agent's consumption of variety ω, ρ is a measure of substitutability across varieties (such that $1/(1-\rho)$ is the elasticity of substitution), $\varepsilon_t^i(r)$ is the agent's taste for location r, and $m(r_{s-1}, r_s)$ is the moving cost from location r_{s-1} to location r_s. Local amenities are given by $a_t(r) = \bar{a}(r) \bar{L}_t(r)^{-\lambda}$, where $\bar{a}(r)$ is the time-invariant fundamental amenity of the location, while $\bar{L}_t(r)^{-\lambda}$ is a congestion externality that makes locations with higher population density, $\bar{L}_t(r)$, less attractive.

For tractability, moving costs need to be modeled as a product of an origin-specific term and a destination-specific term, that is,

$$m(r, s) = m_1(r) m_2(s).$$

It is reasonable to assume that staying put is costless, implying $m(r, r) = 1$ for any r and hence

$$m_1(r) = m_2(r)^{-1}.$$

As a result, the spatial distribution of moving costs is fully characterized by the function $m_2(\cdot)$.

Location tastes $\varepsilon_t^i(r)$ are drawn independently across agents, locations and time from a Fréchet distribution with shape parameter $1/\Omega$. This implies that the equilibrium number of people choosing to reside at location r is given by

$$H_t(r) \bar{L}_t(r) = \frac{u_t(r)^{1/\Omega} m_2(r)^{-1/\Omega}}{\int_S u_t(s)^{1/\Omega} m_2(s)^{-1/\Omega}} \bar{L}_t \tag{57}$$

where $u_t(r)$ is the agent's utility stemming from her consumption of local amenities and varieties, thus equal to

$$u_t(r) = a_t(r)\, \frac{w_t(r) + R_t(r)/\bar{L}_t(r)}{P_t(r)} \tag{58}$$

where $w_t(r)$ is the agent's labor income, $R_t(r)$ denotes local land rents (per unit of land) redistributed to agents at r, and $P_t(r)$ denotes the CES price index of varieties at r.

D.1.2 Technology

At any location r, there is a large number of competitive firms producing each variety ω. Each of these firms has access to the production function

$$q_t^\omega(r) = \phi_t^\omega(r)^{\gamma_1}\, z_t^\omega(r)\, L_t^\omega(r)^\mu$$

where $q_t^\omega(r)$ denotes the firm's output per unit of land, $L_t^\omega(r)$ denotes labor per unit of land used by the firm in the production of ω, $z_t^\omega(r)$ is a location-, variety- and time-specific productivity shifter, and $\phi_t^\omega(r)$ is the amount of innovation conducted by the firm to increase its productivity. Innovation requires labor, such that the firm needs to hire $\nu_t \phi_t^\omega(r)^\xi$ workers to innovate $\phi_t^\omega(r)$ in period t. Idiosyncratic productivity shifters $z_t^\omega(r)$ are drawn independently across locations, varieties and time from a Fréchet distribution with cdf

$$Pr\left[z_t^\omega(r) \leq z\right] = e^{-T_t(r) z^{-\theta}}$$

such that the scale parameter of the distribution is given by $T_t(r) = \tau_t(r)\, \bar{L}_t(r)^\alpha$. $\bar{L}_t(r)^\alpha$ is an agglomeration externality that makes the location's productivity increase in population density, while $\tau_t(r)$ evolves according to the law of motion

$$\tau_t(r) = \phi_{t-1}(r)^{\theta\gamma_1} \left[\int_S \eta \tau_{t-1}(s)\, ds\right]^{1-\gamma_2} \tau_{t-1}(r)^{\gamma_2} \tag{59}$$

where the term $\left[\int_S \eta \tau_{t-1}(s)\, ds\right]^{1-\gamma_2}$ allows for global diffusion of technology as long as $\gamma_2 < 1$, while $\phi_{t-1}(r)$ stands for average innovation conducted by local firms at time $t-1$.[56] As a result, innovation at a location not only increases the location's current productivity but also shifts up the distribution of local productivity draws next period. Thus, innovation can increase not only short-run productivity but also long-run growth. At the same time, just like in the model of Section 3, firms in the DNR model are aware that their innovation spills over to their local competitors next period, driving down their future profits to zero. Hence, they choose a level of innovation that maximizes their current-period profits. In other words, firms solve a static optimization problem each period, which makes

[56]In equilibrium, all firms in a given location conduct the same level of innovation. Hence, it does not matter which moment of the distribution of innovation levels I consider; for completeness, I assume that it is the average level of innovation that matters for the evolution of productivity.

the model computationally tractable even though it features dynamics with a large number of asymmetric locations. The optimal level of innovation chosen by each firm at location r is given by

$$\phi_t(r) = \left[\frac{\gamma_1}{\nu_t(\mu\xi + \gamma_1)}\bar{L}_t(r)\right]^{1/\xi}. \tag{60}$$

Varieties are tradable across locations, but trade is subject to shipping costs. Shipping costs take the iceberg form. If q units of a variety are shipped from location r to location s, the quantity that actually arrives at s is given by $\frac{q}{\varsigma_t(r,s)}$. Thus, $\varsigma_t(r,s) \geq 1$ denotes the iceberg shipping cost between r and s.

D.1.3 Equilibrium

Competition among firms implies that the price of any variety produced at r and sold at s becomes equal to its marginal cost:

$$p_t^\omega(r,s) = \frac{mc_t(r)\,\varsigma_t(r,s)}{z_t^\omega(r)}$$

where

$$mc_t(r) = \mu^{-\mu}\left(\frac{\nu_t\xi}{\gamma_1}\right)^{1-\mu}\left[\frac{\gamma_1 R_t(r)}{w_t(r)\,\nu_t(\xi(1-\mu)-\gamma_1)}\right]^{(1-\mu)-\frac{\gamma_1}{\xi}}\frac{w_t(r)\,\varsigma_t(r,s)}{z_t^\omega(r)}. \tag{61}$$

Using similar steps as in Eaton and Kortum (2002), this relationship between the price and $mc_t(r)$ as well as the Fréchet distribution of $z_t^\omega(r)$ can be used to obtain the share of location s's spending on varieties from location r as

$$\pi_t(r,s) = \frac{T_t(r)\left[mc_t(r)\,\varsigma_t(r,s)\right]^{-\theta}}{\int_S T_t(u)\left[mc_t(u)\,\varsigma_t(u,s)\right]^{-\theta} du} \tag{62}$$

while, by the definition of the CES price index,

$$P_t(s) = \Gamma\left(1 - \frac{\rho}{(1-\rho)\theta}\right)^{-\frac{1-\rho}{\rho}}\left[\int_S T_t(r)\left[mc_t(r)\,\varsigma_t(r,s)\right]^{-\theta} dr\right]^{-\frac{1}{\theta}}. \tag{63}$$

Firms' Cobb–Douglas technology in labor and land and the local redistribution of land rents imply that land rents at location r are proportional to labor income:

$$R_t(r) = \frac{\xi - \mu\xi - \gamma_1}{\mu\xi + \gamma_1} w_t(r)\,\bar{L}_t(r) \tag{64}$$

and therefore market clearing conditions can be written in the form

$$w_t(r)\,H_t(r)\,\bar{L}_t(r) = \int_S \pi_t(r,s)\,w_t(s)\,H_t(s)\,\bar{L}_t(s)\,ds. \tag{65}$$

For a given initial spatial distribution of productivity $\tau_0(\cdot)$, equations (57) to (65) characterize the equilibrium spatial distribution of productivity, innovation, wages, land rents, trade flows and population levels in any period $t \in \{0, 1, ...\}$. Desmet et al. (2018) show that the equilibrium of the model exists and is unique if

$$\frac{\alpha}{\theta} + \frac{\gamma_1}{\xi} < \lambda + 1 - \mu + \Omega \tag{66}$$

which is a condition that, intuitively, requires that agglomeration forces are weaker than congestion forces. Following the same steps as in Desmet et al. (2018), it is straightforward to prove that condition (66) is also sufficient for the existence and uniqueness of the equilibrium in this model. Moreover, condition (66) holds under the values of structural parameters chosen in the calibration. Hence, one can be sure that the equilibrium of the DNR model found in the simulation is indeed the only equilibrium.

In the model of Desmet et al. (2018), the equilibrium growth rate of the world economy is increasing in world population. I assume that world population grows in the current model, which, everything else fixed, should then lead to accelerating growth. To avoid this, I follow footnote 11 in Desmet et al. (2018) and assume that the cost of innovation, ν_t, is linear in world population: $\nu_t = \nu \bar{L}_t$. In this case, Desmet et al. (2018) show that the growth rate is still a function of the population distribution, but not a function of $\bar{L}_t$.

D.2 Taking the DNR model to the data

This section describes how I take the DNR model to the data. Similarly to the model of Section 3, I treat each 20 by 20 arc minute grid cell of the U.S. as one location. One additional location represents Europe. I calculate the supply of land, $H_t(r)$, at each U.S. location in the same way as in the model of Section 3. I set the supply of land in Europe, $H_t(e)$, to $10,000$, as the land area of the European continent is approximately 10,000 times the land area of a grid cell. I also follow the model of Section 3 by setting the length of one period to five years.

Next, I need to choose the values of structural parameters, the distribution of shipping costs, amenities, initial productivity at time $t = 0$, and moving costs. I discuss each of these below.

Calibration of structural parameters. The scale parameter of firms' productivity draws, θ, equals the trade elasticity in the DNR model, which I set to eight, in line with how I calibrate the model of Section 3. I choose the share of land in production to equal 0.2, guided by Desmet and Rappaport (2017). This implies $\mu = 0.8$. Just like Desmet et al. (2018), I calibrate the level parameter of innovation costs, ν, to match the growth rate of the economy. In particular, I choose the value of ν for which the annual growth rate in U.S. real GDP per capita between 1800 and 1820 matches the growth rate in the data, 0.5% (Weiss, 1992). Unfortunately, the data do not provide any guidance on the calibration of the remaining structural parameters. As a result, I set them equal to their baseline calibrated values in Desmet et al. (2018). This implies setting $\alpha = 0.06$, $\xi = 125$, $\gamma_1 = 0.319$, $\gamma_2 = 0.99246$, $\Omega = 0.5$ and $\lambda = 0.32$.

Shipping costs. I calculate shipping costs across locations in the same way as in the model of Section 3. In particular, I borrow the relative costs of shipping via rail, water or wagon from Fogel (1964) and Donaldson and Hornbeck (2016). Just like in the calibration of the model of Section 3, I calibrate parameter ϕ_B, which drives the level of per-distance shipping costs between U.S. coastal locations and Europe, to match the U.S. export to GDP ratio in 1790. The level of per-distance shipping costs within the U.S. are driven by two separate parameters in the model of Section 3: parameter ϕ_F for the farm sector, and parameter ϕ_N for the non-farm sector. The DNR model features only one sector, hence only one parameter ϕ. In the baseline calibration of the DNR model, I set $\phi = \phi_F$, but the results are robust to setting $\phi = \phi_N$ instead.

Amenities and initial productivity. In Desmet et al. (2018), fundamental amenities relative to initial utility $\frac{\bar{a}(r)}{u_0(r)}$ and initial productivity levels $\tau_0(r)$ are calibrated such that the model matches the period-0 distribution of population and nominal income across locations. This procedure cannot be applied in my setting, as there are no consistent data on nominal income at the grid cell level for the pre-Civil War United States. As a result, I make the assumption that every location features the same level of amenities, which I normalize to $\bar{a}(r) = 1$. It is an extreme assumption, but it is consistent with the model of Section 3, in which I also abstract from amenity differences across locations. As the primary aim of the DNR model is that it serves as a benchmark to which my model's fit can be compared, it is in fact reassuring that the two models share this feature.

In Desmet et al. (2018), initial utility levels $u_0(r)$ are chosen to match country-level subjective wellbeing data, with the assumption that $u_0(r)$ do not vary within countries. I adopt this assumption, setting $u_0(r) = u_0$ for every U.S. location and normalizing u_0 to one. For Europe, I calibrate the value of initial utility $u_0(e)$ to match Europe's real GDP in 1790, which, by the assumption that $\bar{a}(e) = 1$, also equals $u_0(e)$.

Once $\frac{\bar{a}(r)}{u_0(r)}$ are known for every r, equations (57) to (65) can be used to back out the distribution of initial productivities $\tau_0(\cdot)$ that rationalize the initial distribution of population in 1790, $\bar{L}_0(\cdot)$. In particular, we have the following lemma.

Lemma 4. *Given* $\frac{\bar{a}(\cdot)}{u_0(\cdot)}$, $H_0(\cdot)$, $\varsigma_0(\cdot,\cdot)$ *and* $\bar{L}_0(\cdot)$, *the spatial distribution of initial produc-*

tivity levels $\tau_0(\cdot)$ *can be obtained as the solution to the equation*

$$\left[\frac{\bar{a}(r)}{u_0(r)}\right]^{-\frac{\theta(1+\theta)}{1+2\theta}} \tau_0(r)^{-\frac{\theta}{1+2\theta}} H_0(r)^{\frac{\theta}{1+2\theta}} \bar{L}_0(r)^{\lambda\theta-\frac{\theta}{1+2\theta}\left[\alpha-1+\left[\lambda+\frac{\gamma_1}{\xi}-(1-\mu)\right]\theta\right]} =$$

$$\kappa_1 \int_S \left[\frac{\bar{a}(s)}{u_0(s)}\right]^{\frac{\theta^2}{1+2\theta}} \tau_0(s)^{\frac{1+\theta}{1+2\theta}} H_0(s)^{\frac{\theta}{1+2\theta}} \varsigma_0(s)^{-\theta} \bar{L}_0(s)^{1-\lambda\theta+\frac{1+\theta}{1+2\theta}\left[\alpha-1+\left[\lambda+\frac{\gamma_1}{\xi}-(1-\mu)\right]\theta\right]} ds$$

where $\kappa_1 = \left[\frac{\mu\xi+\gamma_1}{\xi}\right]^{-\left(\mu+\frac{\gamma_1}{\xi}\right)\theta} \mu^{\mu\theta} \left[\frac{\xi\nu}{\gamma_1}\right]^{-\frac{\gamma_1\theta}{\xi}} \Gamma\left(1-\frac{\rho}{(1-\rho)\theta}\right)^{\theta\frac{1-\rho}{\rho}}$.

Proof. Combining equations (57), (58) and (60) to (65) yields the result. □

Lemma 4 provides me with an initial productivity level for every location that was part of the U.S. in 1790. I assign the productivity of the nearest location to locations that became part of the U.S. due to the political border changes after 1790.

Moving costs. I follow the procedure in Desmet et al. (2018) to back out location-specific moving costs $m_2(r)$. In particular, I choose $m_2(r)$ such that I exactly match the change in location r's population between 1790 and 1795. For U.S. grid cells, population data are available for the census years of 1790 and 1800 but not for 1795. Hence, I assume that the growth rate of population in any cell was the same between 1790 and 1795 as between 1795 and 1800. Desmet et al. (2018) show that, whenever condition (66) holds, this procedure identifies a unique set of moving costs that rationalize the changes in population. Just like with productivity, I assign the moving cost of the nearest locations to locations that became part of the U.S. after 1790.

D.3 Simulation

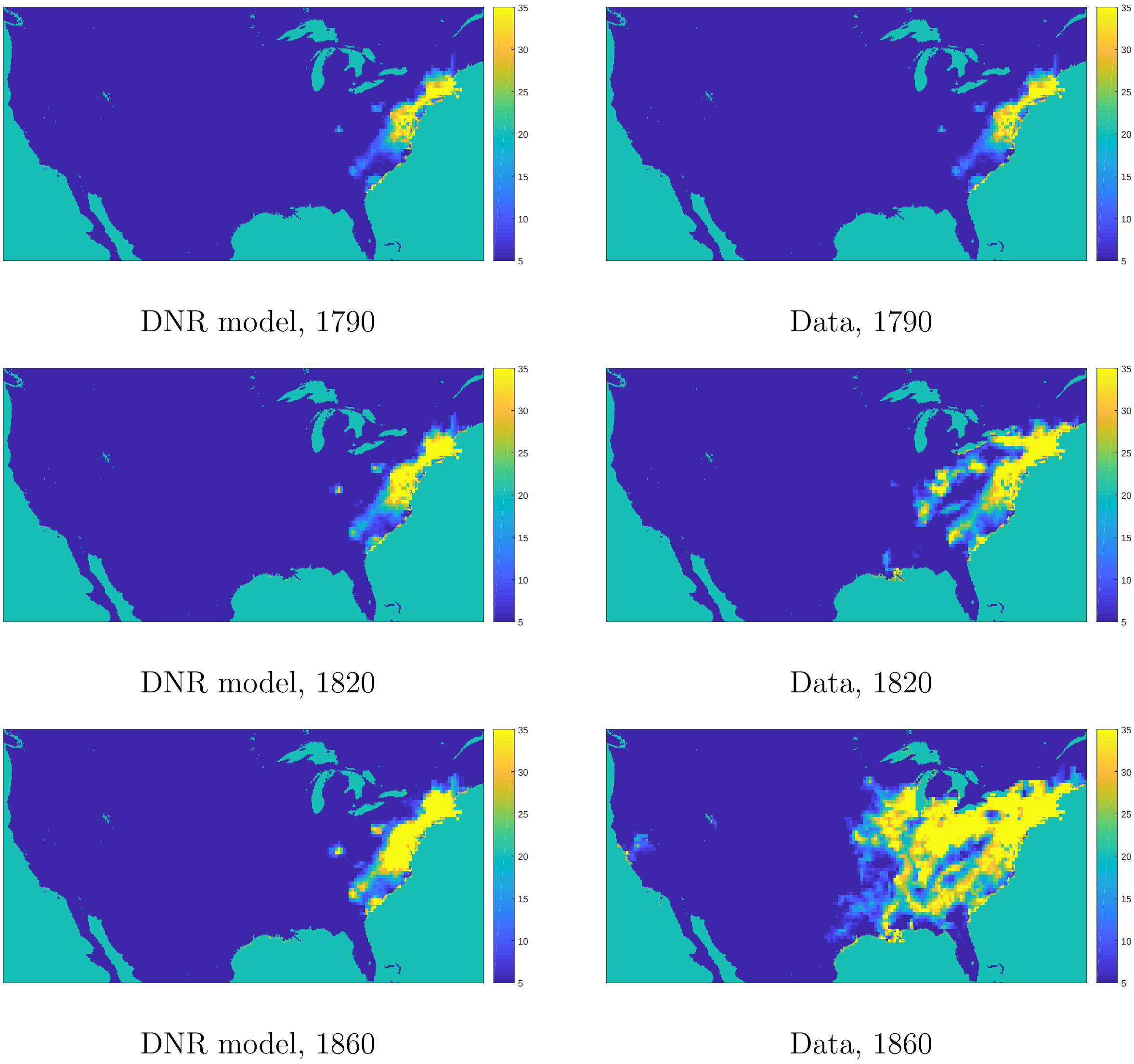
Armed with values of the model's fundamentals, I simulate the DNR model for 14 subsequent periods, corresponding to the years 1795, 1800, ..., 1860.[57] During the simulation, I change world population $\bar{L}_t$ (sum of the population of the U.S. and Europe), U.S. land $H_t(r)$ and shipping costs $\varsigma_t(r,s)$ as they change in the data. In particular, I feed the construction of railroads and canals into the model through changes in $\varsigma_t(r,s)$. To simulate the model, I use the iterative procedure proposed in Desmet et al. (2018).

D.4 Results

Figure 11 compares the evolution of the population distribution in the DNR model to the data. The figure shows that the DNR model is unable to replicate the first-order feature of the data: the substantial reallocation of U.S. population to the West. The

[57]The model does not need to be simulated for 1790, as it matches the 1790 population data exactly.

Figure 11: Population distribution: DNR model versus data



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 DNR model 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 35 for visibility. Source: U.S. census.

Figure 12: City locations by 1860: DNR model versus data

In the left map, yellow 20 by 20 arc minute grid cells are those that have a city in them in 1860 in the DNR model. 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.

intuition for this is as follows. In the DNR model, locations with the highest population density (and therefore the largest markets) offer the highest returns to innovation. As population density is initially the highest in the Northeast, these locations innovate the most, which reinforces their productivity advantage over the rest of the United States. Changing political borders and the expanding transport infrastructure in the West mitigate this effect, but are quantitatively unable to overturn it.

A key difference relative to the DNR model is that innovation is not a function of local population density in the model of Section 3. Instead, innovation depends on the size of the location's effective hinterland (Proposition 1). This feature of the model helps locations in the Midwest, with access to a large farm hinterland especially after the construction of railroads, to grow fast and attract a substantial fraction of U.S. population by 1860. This force, coming from the interaction between the farm and non-farm sectors, is absent from one-sector models such as the DNR model.

In line with this, the DNR model seems unable to rationalize the formation of cities in the Midwest. Although it is not obvious how to define the notion of a city in a one-sector model, I can select the cells with the highest population density in 1860, such that their total population equals total city population in the data. This gives me a total of 30 city cells in 1860. Figure 12 presents the location of these cells and contrasts them with the locations of 1860 cities in the data. As expected, the DNR model features a single cluster of cities in the Northeast, unlike the data and the model of Section 3.

In conclusion, the one-sector framework of Desmet et al. (2018) is unable to replicate the striking westward expansion of population and the process of city formation. The reason the model of Section 3 does substantially better seems to be that it gives a role to farm hinterlands in the growth of the non-farm sector. This interaction between the two sectors allows cities to form at Western locations with access to a large farm hinterland, which in turn can explain both the shift of U.S. economic activity to the West and the geographic patterns of city formation prior to the Civil War. In other words, the comparison of my model to the DNR framework further highlights the quantitative relevance of the hinterland hypothesis.

D.5 Robustness

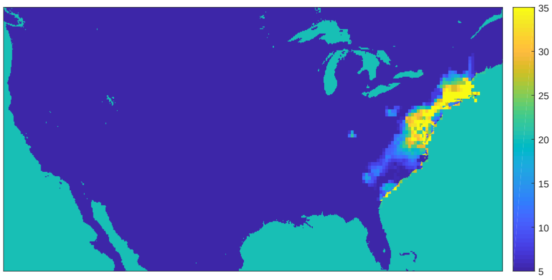
The simulation procedure of Appendix D.3 matches world population every period, but not necessarily U.S. population every period. In particular, U.S. population grows somewhat less in the model than in the data, which suggests that the frictions to moving from Europe to the U.S. might have become weaker over time. To see if my results are robust to this discrepancy between the model and the data, I simulate an alternative version of the DNR model, which I call "DNR (match total U.S. population)." In this version of the model, I let $m_2(e)$ change over time. As only the ratio of moving costs matters for the equilibrium, this procedure is equivalent to letting the cost of entering any U.S. location change to the same extent. I choose the value of $m_{2t}(e)$ in every period t such that the model matches U.S. population at t.

Figure 13 presents the evolution of the population distribution in the "DNR (match total U.S. population)" model. The figure shows that matching total U.S. population does not change the overall picture: the one-sector DNR model cannot replicate the westward movement of U.S. population. This is not surprising as the ratio of two locations' innovation rates depends on the ratio of their population density in the DNR model, not on absolute population levels (Desmet et al., 2018).

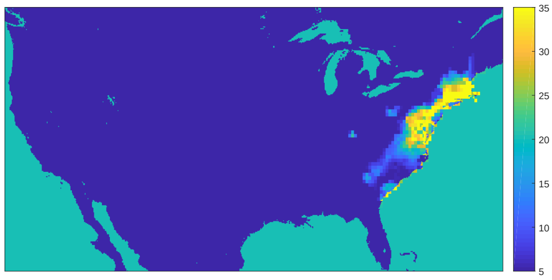
As a second robustness check, I also simulate an alternative version of the DNR model in which productivity does not grow over time. This alternative model, which I call "DNR (no productivity growth)," can be thought of as a one-sector static quantitative model of economic geography, simulated over a sequence of time periods over which total world population, U.S. land and shipping costs change as in the data.

Figure 14 shows that this model also fails to replicate the westward movement of population. This is not surprising as, unlike the model of Section 3, this alternative model also lacks the relationship between farm hinterland size and productivity growth. Figure 14 confirms that absent this relationship, cities do not start to develop and attract a substantial fraction of U.S. population to the West.

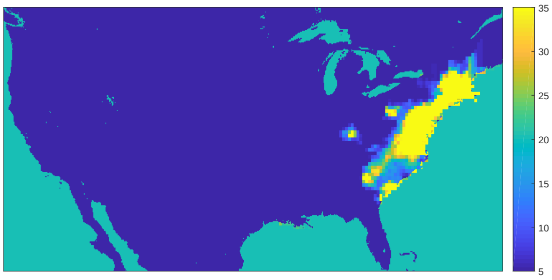
Figure 13: Population distribution: DNR (match total U.S. population) versus data



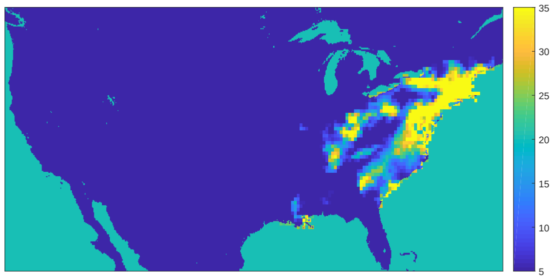
DNR model, 1790



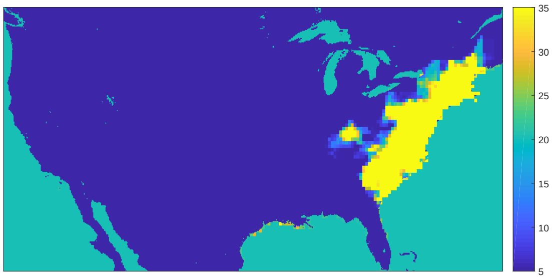
Data, 1790



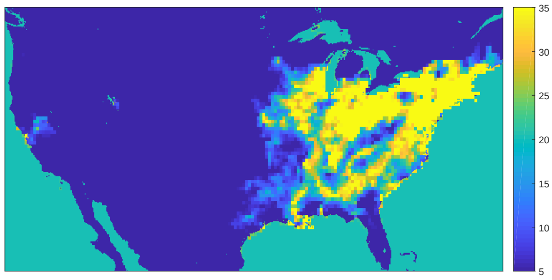
DNR model, 1820



Data, 1820



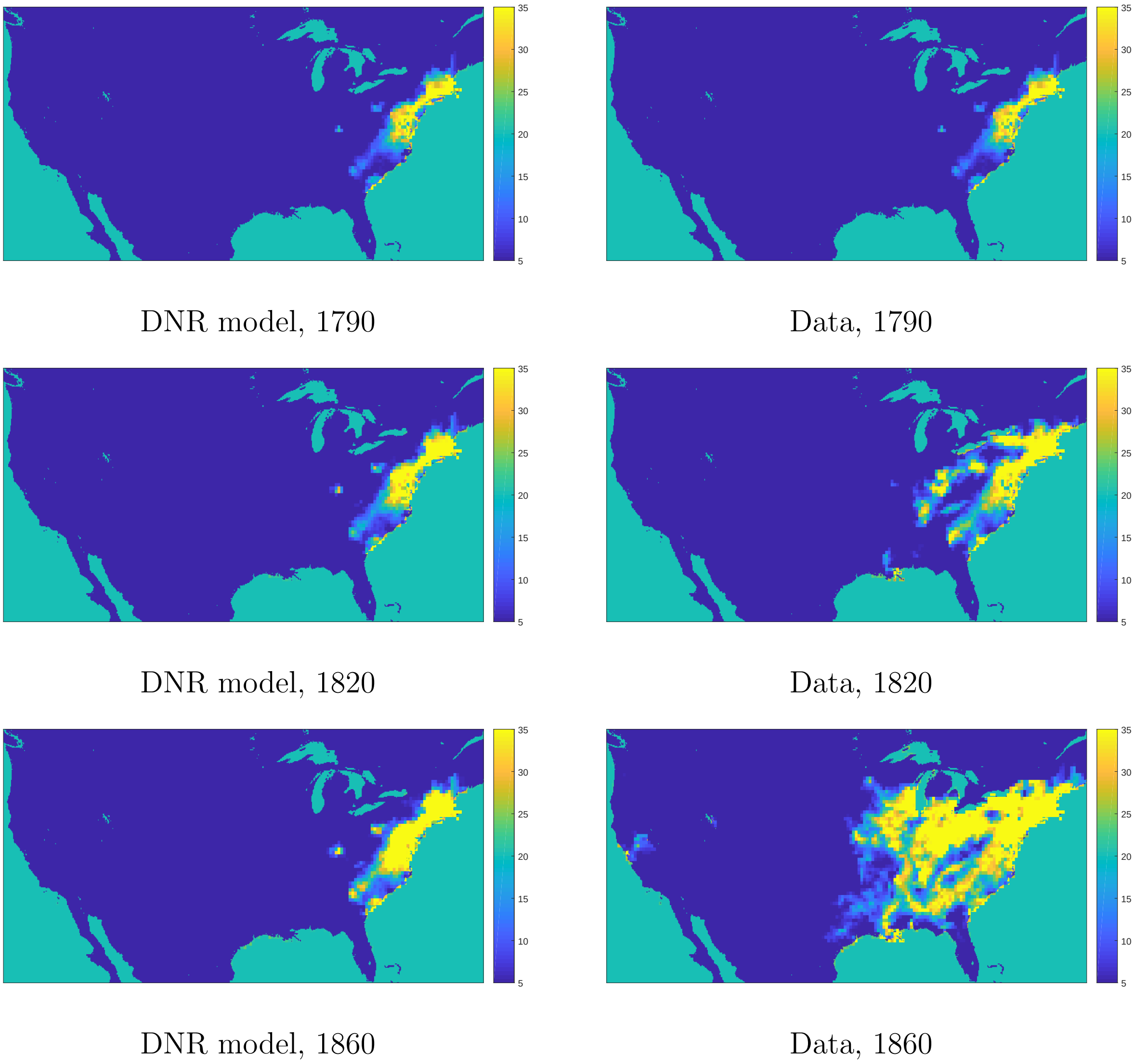
DNR model, 1860



Data, 1860

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 DNR (match total U.S. population) model 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 35 for visibility. Source: U.S. census.

Figure 14: Population distribution: DNR (no productivity growth) versus data



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 DNR (no productivity growth) model 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 35 for visibility. Source: U.S. census.

E Additional tables and figures

Table 14: Discontinuity in the population growth rate of settlements

Dependent variable: ln(growth)				
	Threshold between cities and towns			
	(1)	(2)	(3)	(4)
	10,000	6,000	8,000	15,000
ln(size)	-0.02	0.05**	0.01	-0.01
	(0.02)	(0.02)	(0.02)	(0.03)
city	0.12**	-0.05	0.05	0.10
	(0.05)	(0.04)	(0.04)	(0.06)
No. of observations	371	371	371	371
R^2	0.03	0.01	0.01	0.02

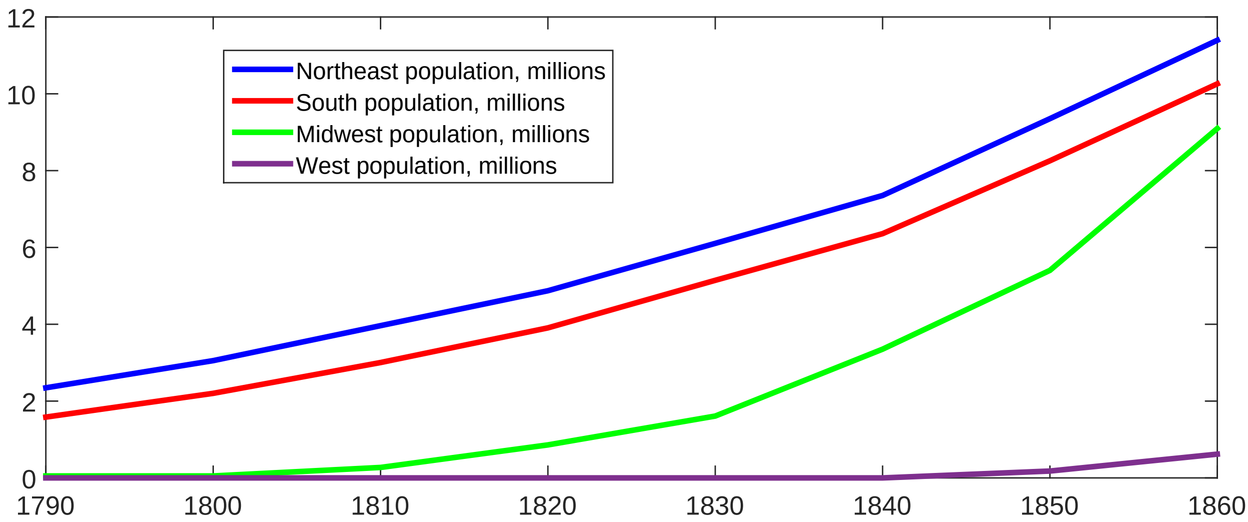
This table presents the results of regressing log settlement population growth on log settlement size and a variable indicating whether the settlement is a city. The unit of observation is a settlement (city or town) in the United States in a given census year (1790, 1800, 1810, 1820, 1830, 1840 or 1850). The dependent variable, *ln(growth)*, is the log of the settlement's population growth over the subsequent decade. Independent variable *ln(size)* is the log of the settlement's population. Independent variable *city* is a dummy variable equal to 1 if the settlement is classified as a city in the given year. A settlement is classified as a city if it exceeds a given population threshold. The columns of the table correspond to different population thresholds. For instance, column (1) classifies every settlement above 10,000 inhabitants in the given year as a city. Heteroskedasticity-robust standard errors in parentheses. *: significant at 5%; **: significant at 1%. Source: U.S. census.

Table 15: Large regions' shares in total U.S. population, 1860

Region	Model	Data
Northeast	41.7%	36.3%
South	26.4%	32.7%
Midwest	31.1%	29.0%
West	0.8%	2.0%

The first column of this table presents the share of 1860 U.S. population living in either the four large regions (Northeast, South, Midwest and West) in the baseline model simulation. The second column presents the same population shares in the data. Source: U.S. census.

Figure 15: Population of the four large U.S. regions

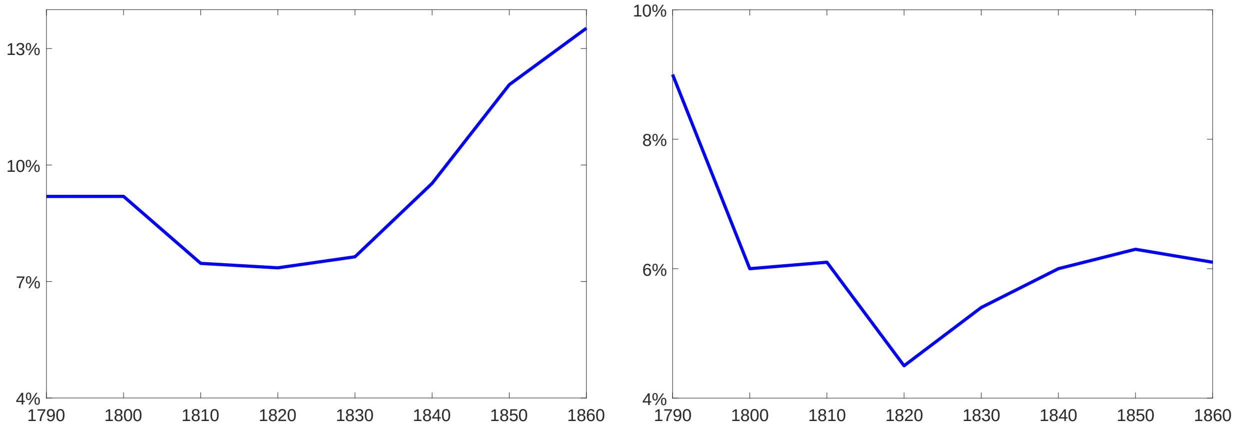


The four lines represent the population of the four large U.S. regions (Northeast, South, Midwest and West), measured in millions of inhabitants, in each census year. Source: U.S. census.

Figure 16: U.S. cities forming between 1790 and 1860

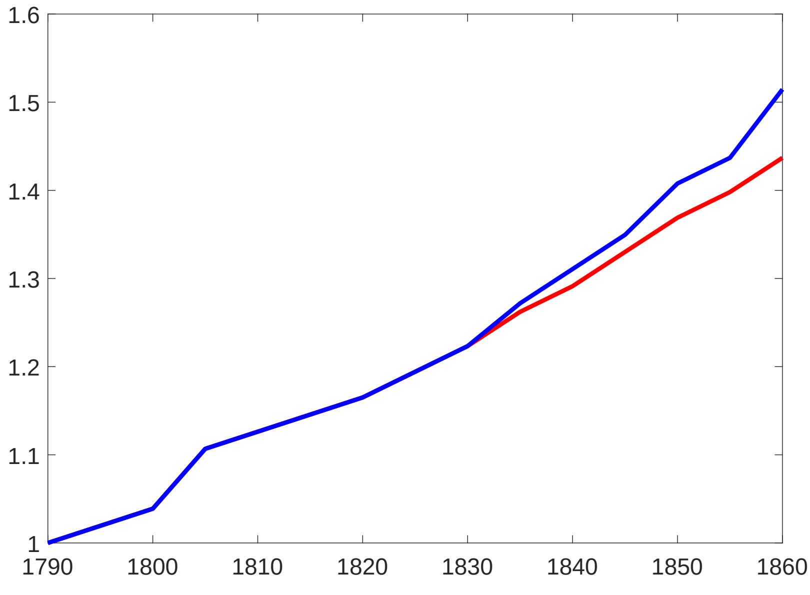
Each red dot in the map represents a city (a settlement above 10,000 inhabitants) that came into existence between 1790 and 1860 in the United States. The West has been cut for visibility. Source: U.S. census.

Figure 17: Exports to U.S. GDP: model versus data



The left graph shows the ratio of exports, measured in U.S. prices, to U.S. GDP in each census year in the baseline model simulation. The right graph shows the same object in the data. Source: Lipsey (1994).

Figure 18: U.S. real GDP per capita, baseline simulation (blue) versus no railroads (red)



The blue line corresponds to U.S. real GDP per capita relative to 1790 in the baseline model simulation. The red line corresponds to U.S. real GDP per capita relative to 1790 in the counterfactual with no railroads (Section 5.2).