City location and economic development*

Dávid Krisztián Nagy†

March 16, 2017

Abstract

I present a dynamic model of the U.S. economy with trade, labor mobility, endogenous growth and realistic geography to examine the relationship between spatial frictions, city formation and aggregate development. In the model, a subset of locations endogenously specialize in innovative industries that are subject to economies of scale. This leads to the formation and development of cities. Spatial frictions affect innovation, thus aggregate growth, by shaping the locations and sizes of cities. I take the model to historical U.S. data at a 20 by 20 arc minute spatial resolution. I show that the model can quantitatively replicate the large population reallocation toward the West and the rapid urbanization in the 19th century, as well as various moments of the location and growth of newly forming cities. I use the model to quantify how the construction of the U.S. railroad network affected city formation, aggregate output and growth. Results indicate that railroads were responsible for 27% of U.S. growth before the Civil War, increasing U.S. real GDP by 9.3% in 1860. I also show that the formation and development of cities amplified the effects of railroads on real GDP by at least 18%.

1 Introduction

Cities are engines of modern economic growth. They tend to host the fastest growing, most innovative firms and industries. At the same time, the locations and sizes of cities are shaped to a great extent by spatial frictions such as trade costs, the immobility of land and the costly diffusion of technology across space. As a result, understanding the relationship between spatial frictions and aggregate economic growth requires a spatial theory of cities. This paper proposes such a theory, and applies it to study city location

*I am grateful to Esteban Rossi-Hansberg for invaluable guidance and support. I am also grateful to Judit Badics, Klaus Desmet, Cécile Gaubert, Gene Grossman, Oleg Itskhoki, Réka Juhász, Miklós Koren, Jan de Loecker, Ildikó Magyari, Thierry Mayer, Eduardo Morales, Fernando Parro, Stephen Redding, Jacques Thisse, Felix Tintelnot, Áron Tóbiás, Sharon Traiberman, seminar participants at Princeton University, and participants of the 2014 Zurich Initiative on Computational Economics 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. Email: dnagy@crei.cat.

and economic development in 19th-century United States, a time period characterized by large population reallocations, the formation of many cities, and tremendous changes in spatial frictions. I use the theoretical framework to quantify how changing spatial frictions, in the form of railroad construction, affected city formation, the growth of the U.S. economy and welfare.

In the model I propose, cities arise due to locations' specialization in farm or non-farm activities. The farm sector uses land and labor to produce a homogeneous good under constant returns, 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 increasing returns. They also hire workers to innovate, which leads to endogenous growth in their productivity. These assumptions guarantee that most of non-farm labor and innovation concentrate at a few locations with large population: *cities*. The locations and sizes of cities, hence sectoral specialization, innovation and growth, are shaped by spatial frictions arising from trade costs, immobile land and costly technology diffusion. Since the model can tractably accommodate a large number of locations and any distribution of shipping costs, land and productivity across these locations, I can incorporate all of these frictions at a high spatial resolution.

I take the model to historical data on U.S. geographic attributes such as rivers, lakes and railroads, the distribution of farm productivity, as well as county, city and town populations. To calculate shipping costs, I combine the geographic data with freight rate estimates based on Fogel (1964). To obtain farm productivity, I use high-resolution data on crop yields along with the 1860 Census of Agriculture. To back out the initial distribution of non-farm productivity and the values of structural parameters, I use the structure of the model, moments of the population distribution in 1790, and aggregate moments of growth and urbanization between 1790 and 1820. This approach also addresses the potential bias due to endogeneity of railroad placement, as I explain in Section 4. Using the estimated shipping costs, the initial distribution of productivity and structural parameters, I solve the model forward to simulate the dynamic evolution of the U.S. economy until 1860, and compare it to the evolution seen in the data.

I find that the model can quantitatively replicate the major patterns of pre-Civil War U.S. urban history. The model predicts, in line with the data, that U.S. population reallocates towards previously unused land in the West. The share of people living in cities increases by a factor of five, but a large part of this rapid urbanization can be attributed to cities appearing near the old, high-productivity cities of the Northeast. This suggests a crucial role of *technology diffusion* in city location. The model also successfully matches the pattern of urbanization and city location outside the Northeast, where new cities appeared farther away from each other. This points to the important role of land in city location: cities must *compete for the hinterland* from which they are supplied. Therefore, they locate far apart from each other, unless technology diffusion is strong. Finally, the model

can replicate the fact that most new cities formed near favorable trading opportunities. The fraction of cities appearing at navigable waterways and confluences of waterways is disproportionately high both in the model and in the data. This shows the importance of *trade* in city location.

Next, I use the model to study the effects of changing spatial frictions, in the form of railroad construction, on growth and city formation. To this end, I simulate the model in the counterfactual scenario in which no railroads were built between 1790 and 1860. The results suggest that railroads largely influenced the growth of U.S. cities. In the absence of railroads, the size distribution of cities would have been more compressed, and real GDP would have been 9.3% lower in 1860. I find that railroads accounted for 27% of aggregate U.S. growth between 1790 and 1860. Finally, a model-based decomposition shows that city formation and development was responsible for at least 18% of the effect of railroads on real GDP. Hence, ignoring this channel would lead to underestimating the effect of transport infrastructure on output and welfare.

This paper is related to four strands of the literature. The first main contribution of the paper is that it investigates urbanization and the role of cities in fostering economic 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), Buera and Kaboski (2012a,b) and Caselli and Coleman (2001), focuses on structural change, which has been recognized as a key source of urbanization.[1] The models in these papers are also highly stylized in their spatial structure, or are completely aspatial. In contrast, the flexible geographic structure in my paper allows me to take the model to spatially disaggregated data, and to measure the effect of real-world changes in spatial frictions on city location, structural change and aggregate growth.

The fact that my paper models the economy over a rich geography relates it to the rapidly growing literature applying quantitative trade models with labor mobility to examine the spatial distribution of economic activity, such as Allen and Arkolakis (2014), Caliendo et al. (2016), Fajgelbaum and Redding (2014), Monte, Redding and Rossi-Hansberg (2016) and Redding (2016). 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 spatial frictions shape growth. My main contribution to this literature is developing a quantitative framework that allows me to study not only the spatial distribution of economic activity, but also its evolution. Using this framework, I find that accounting for the time dimension substantially changes the estimated welfare effects of spatial frictions. This

[1] See Herrendorf, Rogerson and Valentinyi (2014) for a review of the literature on structural change.

is because spatial frictions not only have a spatially heterogeneous impact on within-period welfare, but also a spatially heterogeneous impact on growth through altering the formation and development of cities.[2]

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, Nagy and Rossi-Hansberg (2016) and Michaels, Rauch and Redding (2012). The way I model the diffusion of productivity, which allows me to solve the model as a sequence of static problems, builds on Desmet and Rossi-Hansberg (2014). Relative to this literature, I emphasize the key role of cities in determining the relationship between geography and growth. Geography shapes the number, locations and sizes of cities through trade costs, the immobility of land and the spatial diffusion of technology. The distribution of cities, in turn, shapes the spatial distribution of innovation, hence aggregate growth in the economy.

Finally, my paper is related to the literature that aims at quantifying the economic impact of 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.[3] 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. I show that this new effect amplifies the effect of railroads on output by at least 18%.

The structure of the paper is as follows. Section 2 provides evidence on 19th-century U.S. urban history that motivates the choice of a structural framework with trade, geography, labor mobility and endogenous growth. Section 3 outlines the model, then Section 4 describes the steps of taking the model to the data. Section 5 presents how well the simulated model fits the data, as well as the results of the counterfactual exercise with no railroads. Section 6 concludes.

[2]This finding is in line with Caliendo, Dvorkin and Parro (2015), who argue, using a spatial model in which households make dynamic location choices subject to mobility frictions, that capturing within-country population dynamics is key to understanding the impact of trade liberalization on welfare. I reach a similar conclusion in a model in which, unlike in Caliendo et al. (2015), labor mobility is frictionless but the spatial distribution of productivity evolves endogenously.

[3]Other recent papers studying the economic impact of transport infrastructure in quantitative spatial frameworks include Fajgelbaum and Schaal (2016), Allen and Arkolakis (2016), Herrendorf, Schmitz and Teixeira (2012), Swisher (2014) and Trew (2016). Also see the review of this literature in Redding and Turner (2015).

2 Motivating evidence: City formation in 19th-century United States

This section presents the major patterns of U.S. urban history in the 19th-century, which guide the choice of the structural model. During the 19th century, the United States saw a substantial shift in its population and economic activity toward the West (Fact 1). At the same time, the share of people living in cities increased by a factor of five, and most of this urbanization can be attributed to new cities forming as opposed to pre-existing cities growing (Fact 2). The location of newly forming cities seems to be strongly connected to better trading opportunities, such as access to navigable rivers or railroads (Fact 3). Furthermore, there were substantial geographic differences in the patterns of city location: new cities appeared nearby existing ones in the Northeast, but not in the rest of the U.S. (Fact 4). Finally, city growth was approximately orthogonal to city size (Gibrat's Law) but with slight convergence; also, small towns grew slower than cities above 10,000 inhabitants (Fact 5). These patterns can only be replicated in a model with labor mobility and geography (Fact 1), structural change (Fact 2), trade and technology diffusion across locations (Facts 3 and 4), and geographically heterogeneous growth (Fact 5).

I use three datasets to document these patterns. 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*.[4] 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. The Data Appendix 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). 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 structural framework, I focus on the years 1790 to 1860 throughout the paper.[5]

[4]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 (2015).

[5]It would be intruiging to use the framework presented in this paper to learn about what would have happened to the U.S. economy without the Civil War, by simulating the model for periods after 1860. However, this would require the alternative railroad network, which would have emerged in the absence of

Fact 1: Expansion to the West

U.S. population increased by a factor of eight between 1790 and 1860. Population growth was, however, far from uniform across space.[6] Population in the West increased 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 a convergence in population sizes among three large regions of the United States: the Northeast, the South and the Midwest.[7] 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 2).[8]

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 3). Part of this increase can be the result of population growth: as total population grows, the population of each settlement is likely to rise as well, hence more and more settlements reach the 10,000-inhabitant threshold. To avoid this issue, I set a threshold that equals 10,000 in 1790, but then grows in proportion with total population. The red line in the left panel of Figure 3 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 3 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

the war, as an input. Recent advances in optimal transport infrastructure placement such as Fajgelbaum and Schaal (2016), combined with a dynamic spatial framework such as the one presented in this paper, can provide a fruitful direction for future research to answer such questions.

[6]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 borders during the period, mainly as a result of the Louisiana Purchase (1803), the annexation of Texas (1845) and the Mexican–American War (1847). 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. Therefore, I incorporate these changes when simulating the model and conducting counterfactuals.

[7]I follow Caselli and Coleman (2001) when defining the geographical boundaries of these regions. See the Data Appendix for the list of states belonging to each region.

[8]Although political border changes increased the area of the Midwest and the South over 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.

(green line in the left panel of Figure 3). That is, about 60% of the urbanization was due to cities that did not yet exist in 1790.

Finally, the right panel of Figure 3 shows that the rapid increase in 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 and the Midwest. This suggests an increase in specialization among large U.S. regions. Whereas the Northeast specialized in activities that can be carried out in cities, the other regions remained largely rural.

Fact 3: Importance of trading routes

Where did new cities form? Figure 4 shows the locations of the 94 cities appearing between 1790 and 1860. A striking regularity is that almost all new cities are close to places with good trading opportunities such as rivers (especially the Mississippi, the Ohio River and the Hudson), the Great Lakes, or the sea. To confirm this, I discretize the territory of the U.S. into 20 by 20 arc minute grid cells,[9] and classify each cell depending on whether it is located near a large body of water (i.e., next to a cell that includes a navigable river or canal, a lake, or the sea), near a confluence of large bodies of water, or near an early-built railroad. I consider early-built railroads to mitigate the bias coming from the fact that cities might have caused railroad building in their surroundings, not vice versa. Still, the emergence of cities at railroads and canals should not be interpreted as a strict causal relationship, but rather as 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 early-built railroads.

The formation of most cities at trading routes does not necessarily imply that trade attracted people to these locations. In what follows, I address two alternative explanations and show that neither of them are able to explain the disproportionate emergence of cities near trading routes.

Better amenities, or higher agricultural productivity. It is possible that locations with good trading opportunities were also better in their natural amenities, or had higher agricultural productivity, and such advantages, not trade itself, attracted cities to these locations. To control for natural amenities and agricultural productivity, I collect

[9]This means that each cell is approximately 20 by 20 miles large.

data from the FAO GAEZ database,[10] and consider specifications of the form

$$newcity_i = \beta_0 + \beta_1 tradeopp_i + \beta_2 amen_i + \beta_3 prod_i + \epsilon_i \quad (1)$$

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 (between 1840 and 1860 in the case of railroads), and zero otherwise. $amen_i$ and $prod_i$ are amenity and productivity controls. I estimate specification (1) as a linear probability model.[11]

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 large body of water, or at a confluence receive a city with a 2 to 3 percent higher probability than other locations. If a location is at a railroad built before 1840, the chance that a city forms at the location after 1840 is more than 6 percent higher. The inclusion of amenity and productivity variables hardly affects the significance or the magnitude of these estimates.

Slow migration to the West. The gradual settlement of the West could be another reason for the emergence of cities at trading routes. Occupying the West was a slow transition process, and this transition might have played out in a way that more settlers came together at locations with good trading opportunities. 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 consider a process in which a resident of location j moves to location i next period with probability $\pi_{ji} \in [0, 1]$ as a benchmark. 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 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 distance between j and i:

$$\pi_{ji} = e^{-\xi |j-i|}$$

where $|j - i|$ denotes the distance (in miles) between locations j and i. This assumption

[10] The natural amenity variables are based on measures of temperature and precipitation, while the productivity variables are based on yields of the major crops grown in 19th-century U.S. The Data Appendix describes these variables in detail.

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

guarantees that people mostly stay close to where they have lived in the past. Thus, they occupy previously uninhabited Western regions at a slow pace. The speed of the transition is driven by the parameter ξ.

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 value of parameter ξ. I choose the length of a period to be five years.[12] 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 the beginning of the period.[13] 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. Finally, as parameter ξ drives the speed of transition, I calibrate its value such that the mean center of population moves as much to the West in the model as in the data. This implies $\xi = 0.016$. Hence, the probability that a person moves 50 miles from her current residence next period is $e^{-0.016\times 50} = 45\%$ of the probability that she does not move.

Given that cities appeared in 75 grid cells in the data during the period, 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. I also control for the amenities and productivity of each location by estimating equation (1) on simulated data. Table 3 presents the results. As can be seen from the table, the process of slow migration is unable to replicate the disproportionately high share of cities near trading routes. The probability that a city appears at a large body of water, a confluence, or a railroad is not significantly higher than the probability that it appears elsewhere. Although the fraction of cities at railroads is higher than in the data (60%), the effect of railroads on the emergence of cities becomes insignificant once I control for amenities and productivity.[14]

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

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

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

[14]One concern about these findings can be that migration may have been, similar to trade, less costly along water routes than inland. I address this issue by generalizing the transition process to one in which moving costs are lower along water, calibrating the difference between water and inland costs to match the concentration of population. The results, available from the author upon request, suggest that even this generalized transition process stops short at capturing the disproportionate emergence of cities near trading routes.

Fact 4: Clustering of cities 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 4 suggests that cities tended to form close to other cities in the Northeast, but this was less true in the rest of the United States. 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 4. These results suggest that forces attracting cities to each other, such as technology diffusion, were more active in the Northeast, whereas forces of dispersion, such as competition between cities for land, dominated in the rest of the country.[15]

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). The separate treatment of these two types of settlements is motivated by Desmet and Rappaport (2015), who find substantially different growth patterns between small and large U.S. counties, and the threshold that they find between these two groups lies at about 10,000 inhabitants.

The right panel of Figure 5 plots the log decennial growth rate of cities against their log size in the beginning of the decade. The left panel plots the same for towns.

Cities and towns exhibit convergence in their size since larger settlements grow slower than smaller ones, as can be seen in Figure 5. 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 (2015) show that population growth deviates from Gibrat's Law and exhibits convergence among small counties, in line with my findings.

Figure 5 also shows that city growth rates tend to be above town growth rates. The

[15]One might 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.

difference between the average growth rate of cities and towns is substantial: it is about 1 percentage points per year. To confirm this discontinuity in growth rates between settlements above and below 10,000 inhabitants, I estimate the specification

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

where s indexes settlements, $size_{st}$ denotes the population of settlement s in census year t, $growth_{st}$ denotes the population growth rate of settlement s between t and $t+1$, and $city_{st}$ is a dummy variable that equals one if settlement s had more than 10,000 inhabitants in year t. Column (1) of Table 5 presents the results. The fact that the estimated coefficient on $city_{st}$ is significantly different from zero confirms that the growth rate is indeed higher in settlements above 10,000 inhabitants. Columns (2) to (4) consider thresholds different from 10,000, and find no significant break in the growth rate at these alternative thresholds. The finding that cities grow faster than towns is in line with the different growth patterns of small versus large counties documented by Desmet and Rappaport (2015). 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).

The evidence presented in this section suggests that a theory addressing the relationship between geography, city formation and growth in 19th-century U.S. needs to incorporate various elements – trade and labor mobility, structural change, and growth that is heterogeneous across locations – to capture the important patterns in the data. The next section aims at developing such a theory.

3 A dynamic spatial model of cities

Motivated by the analysis of Section 2, this section presents the structural framework I use to study the forces shaping city formation and growth in a world with spatial frictions. The economy is populated by a mass of consumers who choose where to live and where to trade goods. 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 evolution of the locations and sizes of cities.

I first outline the setup. Next, I define the equilibrium of the economy, show how the recursive structure of the model allows me to solve the equilibrium quickly on the computer, and discuss how growth and geography are related in the equilibrium of the model.

3.1 Setup

3.1.1 Geography, goods and factor endowments

The world is a compact subset S of a two-dimensional surface. Elements of S are called *locations*, and 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 varieties, n_t, is determined endogenously in equilibrium.

Factor endowments are as follows. In period $t \in \{1, 2, ..., \}$, an exogenously given mass of $\bar{L}_t$ consumers live in the world. Each consumer owns one unit of labor which she supplies inelastically. Consumers working in the farm sector also receive an equal share of the land rents paid at their residential location.[16] The supply of land at location r is denoted by $H_t(r)$. $H_t(\cdot)$ is an exogenous measurable function.

To allow for the possibility of international trade in the model, I assume that the world consists of a finite number of countries $c \in \{1, ..., C\}$, such that each location r belongs to one country $c(r)$.[17] Consumers can move freely within countries, but not across countries. Thus, countries' population levels $\bar{L}_{ct}$ are given exogenously.[18]

3.1.2 Consumers

Consumers 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 consumer living at location r and trading at location s obtains the following utility at time t:

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

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. $\zeta = \nu^{-\nu}(1-\nu)^{-(1-\nu)}$ is a constant that simplifies the subsequent formulas algebraically.

Besides choosing the quantity of goods consumed, consumers also pick a location where they live and work, and a location where they trade goods. For simplicity, I assume that

[16] This assumption follows Redding and Sturm (2008) and Desmet et al. (2016). Assuming that farmers actually own (and, therefore, can buy and sell) the land at their location would imply that they need to take into account 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.

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

[18] Note that country populations can change over time. This allows me to read their evolution off the data in every period, thus incorporating birth, death and immigration. Modeling consumers' fertility and immigration decisions is outside the scope of this paper. See Desmet et al. (2016) for a spatial growth model with endogenous immigration.

consumption also happens at the trading place.[19] Finally, consumers also choose a sector in which they work in each time period.[20] In what follows, I describe farmers' and non-farm workers' choice variables and constraints.

3.1.3 Farmers

Farmers produce the farm good at home, ship it to a trading location, and exchange it there for consumption goods. As a consequence, they 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 the budget constraint.

The production function that a farmer producing at r faces is Cobb–Douglas in labor and land:

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

where $q_t^F(r)$ is the quantity of the good produced, $A_t^F(r)$ 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 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

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

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

$$\varsigma_t(r,u) \leq \varsigma_t(r,s)\,\varsigma_t(s,u)$$

for any r, s and $u \in S$. I normalize $\varsigma_t(s,s)$ 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.

[19] This assumption allows me to abstract from shipping costs incurred between home and the trading place. The Theory Appendix 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.

[20] Although it is possible to introduce frictions to 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 early-19th century U.S. (David, 2005).

Finally, the farmer faces the budget constraint

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

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, $R_t(r)$ is the land rent at r, and $y_t(r)$ is the farmer's income coming from receiving an equal share of local land rents.

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 income net of land rents equals her production revenues, and the indirect utility of a farmer living at r and trading at s is

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

where $L_t^F(r)$ is the farm population of r, and $P_t(s)$ is the CES price index of non-farm goods at s, that is,

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

3.1.4 Non-farm workers

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 r chooses her trading place s and her consumption of each good to maximize her utility subject to the budget constraint

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

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

Non-farm workers are subject to commuting costs between their residence r and the trading location s. I normalize the commuting cost to zero for $r = s$, but assume that it is large enough for $s \neq r$ such that non-farm workers do not have an incentive to trade at another location than r. This assumption is motivated by the prohibitively high costs of medium- and long-distance commuting in the 19th century.

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

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

3.1.5 Non-farm technology

Non-farm varieties are produced by firms operating under monopolistic competition with free entry, which 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 increasing returns, which leads to agglomeration. 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,

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

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, $\hat{A}_t^N(s,i)$ 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, $\hat{A}_t^N(s)$, allows firms to hire more workers for production. Higher efficiency, on the other hand, decreases demand for the farm good, lowering demand for farm workers at s and in its surroundings. As a consequence, the ratio of non-farm to farm population goes up, and location s becomes more urbanized. The extent of urbanization depends on the parameter β.[21] This specification relates the paper to the strand of the literature originating from the seminal paper by Ngai and Pissarides (2007), in which structural change is induced by demand complementarities and differential productivity growth across sectors.

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 $A_t^N(s)$ in the beginning of the period and its period-t innovation $\ell_t^I(s)^{1-\mu}$. That is,

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

where $\ell_t^I(s)$ denotes the number of workers hired to innovate.[22] Workers can freely switch between the two tasks, innovation and production. As I show later in Section 3.1.7, a

[21] 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).

[22] Note that the exponents on innovation labor and the farm good sum to one. This assumption helps keep the model tractable by guaranteeing constant returns to scale *after* the fixed cost has been paid.

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.

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 $\tau_t(s,u)$. $\tau_t(\cdot,\cdot)$ is an exogenously given measurable function that is symmetric, that is, $\tau_t(s,u) = \tau_t(u,s)$. 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 trade in agricultural products (both within and across countries) as a non-farm activity, which corresponds to the way farm and non-farm activities are measured in the data.

3.1.6 Evolution of productivity

To incorporate the heterogeneity in growth across space and city size groups 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 country-specific exogenous growth shifters $f_{c(s)}$ and 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. In particular, I assume that the non-farm productivity of s evolves according to the equation

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

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).

I also assume that non-farm technology can diffuse across space 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 $\tilde{A}_{t+1}^N(s)$ in period $t+1$, but can also borrow technology $e^{-\delta|u-s|}\tilde{A}_{t+1}^N(u)$ from another location $u \in S$. $|u-s|$ denotes the great-circle distance between u and s, and $\delta > 0$ is the spatial decay in technology diffusion. The technology that firms at location s *actually use* in period $t+1$

is the best of all these available technologies:

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

Figure 6 shows the timing of events that follows from the above assumptions. Firms at location s start with productivity $A_t^N(s)$ 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, $\hat{A}_t^N(s) = A_t^N(s) \ell_t^I(s)^{1-\mu}$. At the same time, consumers choose their residential location, trading place, production and consumption quantities, and markets clear. After all these, productivity growth shifters are realized, and increase productivity to $\tilde{A}_{t+1}^N(s)$, given by (4). Finally, technology diffuses across space, which leads to a productivity level $A_{t+1}^N(s)$ 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, $A_t^F(r) = A^F(r)$, where $A^F(\cdot)$ is an exogenous measurable function. This assumption is motivated by evidence showing that, although there were innovations on farms in the first half of the nineteenth century, they only led to significant increases in farm TFP in the second half of the century (Towne and Rasmussen, 1960). 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 Towne and Rasmussen's estimates, 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.

3.1.7 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 3.1.5 and 3.1.6. Solving this dynamic optimization problem simplifies to solving 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 transport 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

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

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

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

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

and

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

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.

3.2 Equilibrium and solving the model

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

(Lemma 1). Besides helping solve the model forward in time, this relationship also allows me to draw qualitative conclusions about how spatial frictions shape growth in the model through affecting the distribution of population.

Definition. *Given parameters α, β, γ, δ, ϵ, λ, μ, ν, f, geography S, countries' population levels $\bar{L}_{ct}$ and growth shifters f_{ct}, as well as functions $H_t, A^F, A_1^N : S \to \mathbb{R}$ and $\tau_t, \varsigma_t : S^2 \to \mathbb{R}$, an* ***equilibrium*** *of the economy is a set of functions $L_t^F, L_t^N, p_t^F, p_t^N, q_t^N, h_t, \ell_t^P, \ell_t^I, x_t^F, \bar{c}_t$, $n_t, P_t, \sigma_t, w_t, R_t, A_t^N : S \to \mathbb{R}$, 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, and their utility is equalized within countries. That is,*

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

   for any $r, u \in S_c$, where $\sigma_t(r)$ denotes the trading place chosen by the residents of r,[23] and $U_t^F(\cdot,\cdot)$ is given by (2).

2. *Non-farm workers maximize utility, and their utility is equalized within countries. That is,*

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

   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. Hence,*

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

   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*

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

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

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

[23] I only consider equilibria in which $\sigma_t(\cdot)$ is a measurable function.

[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.

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,*

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

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:

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

7. *National labor markets clear. That is,*

$$\bar{L}_{ct} = \int_{S_c} \left[L_t^F(r) + L_t^N(r) \right] dr \tag{19}$$

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

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

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

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

Note that all equilibrium conditions except for (4) and (5) establish relationships between endogenous variables *within* a period, while (4) and (5) drive the evolution of productivity *between* consecutive periods. Hence, knowing the productivity distribution in period t, equations (2), (3) and (6) to (20) can be used to find the period-t distribution of population and innovation. Once these are known, one can use equations (4) and (5) to update productivities to their levels in period $t+1$. In fact, as presented in the following lemma, knowing the population distribution alone is sufficient since innovation levels only depend on this distribution.

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

$$\ell_t^I(s)^{1-\mu} = \rho\left[\left(1+\frac{L_t^N(s)}{\int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr}\right)^{-1} - (1-\nu)\right]^{1-\mu} \tag{21}$$

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

Proof. See Appendix A.2.

□

Appendix A.1 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 $A_1^N(\cdot)$, one can calculate the population distribution in period 1. Then one can use equations (21), (4) and (5) to update productivity levels at each location, and obtain $A_2^N(\cdot)$. $A_2^N(\cdot)$, in turn, can be used to calculate the population distribution in period 2, which allows one to update productivities again, and so on.[25]

By Lemma 1, firms at non-farm locations with a large farm hinterland, $\int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr$, relative to their own size, $L_t^N(s)$, innovate more. This has three implications. First, innovation exhibits convergence: everything else fixed, large non-farm locations innovate less as they have a relatively low share of farmers in their trading population. This allows the model to replicate the fact that growth exhibits convergence both among cities and among towns (Fact 5 of Section 2). Second, the degree of convergence is driven by the share of innovation labor in firms' expenditures, $1-\mu$. If this share is small, the relationship between innovation and size approximates Gibrat's Law, that is, innovation being independent of size. Finally, changes in spatial frictions that allow the location to expand its hinterland boost innovation. As a result, expanding the railroad network can potentially have large growth effects if it allows non-farm locations to capture a larger hinterland, and hence to innovate more.

To see if the model can quantitatively capture the spatial reallocation of population and the locations of cities, I take the model to the data and simulate it over the period of investigation. Section 4 describes the details of this procedure, while Section 5 presents the simulation results.

[25]For any given population distribution, equations (21), (4) 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.

4 Empirical strategy

This section describes the empirical strategy I use 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 calculate shipping costs and 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, and the values of structural parameters such that the simulated model matches two sets of moments in the data: the concentration of population and the non-farm employment share in 1790, along with the sizes of the five pre-existing cities; and moments of aggregate growth and urbanization between 1790 and 1820. The advantage of this approach is twofold. First, 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 construction and economic outcomes is irrelevant. The second advantage of my calibration approach is that it leaves the evolution of individual locations' population untargeted, both before and after 1820. Hence, I can assess the model's fit by looking at how well the model predicts the evolution of the population distribution in Section 5.1.

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., as of today, 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.[26] $WR_t(r)$ 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 which 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

[26]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.

at *oldrailhistory.com*.[27]

4.2 Incorporating the international dimension

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 as the U.S. was an open economy throughout the 19th century. Economic historians estimate that exports constituted ten to fifteen percent 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 the highest quality data are available for Europe during the period.

4.3 Calculating shipping costs

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

Motivated by this evidence, I assume that farmers' cost of passing through a cell takes the following form. The cost of passing through cell r in period t is (1) ς^W if $WR_t(r) = 1$, that is, the cell has a large body of water in it; (2) $\varsigma^R > \varsigma^W$ 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) $\varsigma^I > \varsigma^R$ if $WR(r) = RR_t(r) = 0$, that is, the cell has neither water nor railroad access. Once the cost of passing 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 ς^W, ς^R and ς^I on the freight rate estimates of Donaldson and Hornbeck (2016) – who, in turn, borrow the estimates from Fogel (1964) –, but with two modifications. First, Donaldson and Hornbeck find that transshipment, that is, transferring goods between different transportation modes, is very costly: on average, changing from one mode to another costs the same as shipping the good 100 miles along water. As the Fast Marching Algorithm cannot incorporate transshipment costs directly, I increase the costs

[27] I incorporate railroads and navigable canals starting from the year of their construction. Details of this procedure are provided in the Data Appendix.

of water and rail shipment by 50% to reflect these additional costs.[28] Second, Donaldson and Hornbeck's baseline value of inland shipping costs is as high as 23.1 cents per ton-mile; simulating the model under this value, however, yields a fraction of population living near rivers that is far higher than in the data, and a low overall correlation between cells' predicted and actual population levels. This suggests using Donaldson and Hornbeck's lower estimate of inland costs, 14 cents per ton-mile, a value for which they find their results to be robust.[29]

Donaldson and Hornbeck's estimates are for agricultural goods, but my model also features non-farm shipping costs $\tau_t(\cdot,\cdot)$. To obtain non-farm costs, I assume that they were related to farm shipping costs according to

$$\tau_t(s,u) = \varsigma_t(s,u)^{\phi}$$

hence the parameter $\phi > 0$ drives the scale of non-farm to farm costs. In what follows, I treat ϕ as a structural parameter, and calibrate it to match moments of the 1790 population distribution in Section 4.5.

4.4 Calculating the distribution of farm productivity

I use high-resolution data on agricultural yields to calculate the spatial distribution of farm productivity $A^F(\cdot)$. 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).[30] Although the yields are only available for the period 1961 to 1990, I apply the filters "low input level" and "rain-fed" to be as close as possible to 19th-century conditions.

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

[28] Moreover, as Bleakley and Lin (2012) argue, transshipment was often necessary even if the mode of transportation remained water, due to the geomorphological features of many U.S. rivers. Similarly, rail track gauge was not standardized until 1886, which made transshipment necessary between lines with different track gauge. Donaldson and Hornbeck's cost estimates do not include these additional costs.

[29] 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 locations at the sea, I measure their great-circle distance to the port of London in miles, and multiply it by the per-mile water shipping cost ς^W. 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). For inland locations in the U.S., I minimize the sum of the shipping cost to locations along the sea and the shipping cost between those locations and Europe.

[30] Costinot et al. (2016) is another recent paper that uses the FAO GAEZ dataset to obtain the spatial distribution of agricultural productivity.

the county's value of farm output.[31] 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 GAEZ to assign the productivity of their main crop to grid cells.

4.5 Calibration

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

Structural parameters borrowed from the literature. 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 spatial decay of productivity diffusion δ – have been estimated and qualified well in the literature. Therefore, I borrow these estimates. In particular, I follow Caselli and Coleman (2001) and set the labor share in farming to $\alpha = 0.71$.[32] To choose the non-farm share in consumption, note that my model regards any trade across cities or towns as non-farm good trade. Therefore, the non-farm consumption share should be interpreted as the total consumption share of manufacturing, services and interregionally traded agricultural products. 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 $\epsilon = 9$.[33]

Finally, to obtain the value of the productivity decay parameter δ, I rely on evidence provided by Comin, Dmitriev and Rossi-Hansberg (2013) on the spatial diffusion of technology. Comin et al. 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

[31] Data are not available for a few counties in the West; to these counties, I assign the main crop of their closest neighbors.

[32] 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.

[33] 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 cities or towns, not between these locations and the 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.

a conservative estimate of $\delta = 0.006$, which implies a spatial decay of 45% in productivity over 100 kilometers.

Non-farm productivity in Europe. To choose the initial non-farm productivity of Europe $A_1^N(e)$ and the European growth shifters f_{et},[34] note that, by equation (4), this is the same problem as choosing the non-farm productivity of Europe every period, $A_t^N(e)$. Also note that, combining equations (8), (11), (12) and (17), it is possible to express the non-farm productivity of Europe as

$$A_t^N(e) = \epsilon^{\frac{1}{\beta-1}} \tilde{A}^F(e)^{-\mu} L_t^F(e)^{(1-\alpha)\mu} \left[\frac{\mu(\epsilon-1)}{\nu \frac{L_t^F(e)}{L_t^N(e)} - (1-\nu)} - \mu - \epsilon(1-\mu) \right]^{\frac{1}{1-\beta}} \tag{22}$$

where $\tilde{A}^F(e) = A^F(e) H(e)^{1-\alpha}$ 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 estimates of European farm and non-farm populations.[35] Plugging these estimates into equation (22), I can back out the non-farm productivity of Europe in every period t as a function of land-adjusted farm productivity $\tilde{A}^F(e)$ and structural parameters.

Initial non-farm productivity in the U.S. My goal is to calibrate the initial distribution of non-farm productivity to the sizes of pre-existing U.S. cities. I index the five cities already existing in 1790 by $k \in 1, ..., 5$. Then I assume that non-farm productivity was initially distributed such that $A_1^N(s) = A_k$ if city k was located in cell s, and $A_1^N(s) = \bar{A}$ if no initial city was located in s. As a result, the values of $\bar{A}$ and $\{A_k\}_{k=1}^5$ fully characterize the distribution of non-farm productivity in the first period.

Calibration. We are left with choosing the values of non-farm productivity parameters $\bar{A}$ and $\{A_k\}_{k=1}^5$, six structural parameters – the ratio of non-farm to farm shipping costs ϕ, the fixed cost in non-farming f, the share of the farm input in non-farming μ, the elasticity of substitution between labor and the farm good in non-farming β, as well as γ and λ that drive dynamic externalities –, and Europe's land-adjusted farm productivity

[34] I normalize the U.S. growth shifter to zero in each period.

[35] 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.

$\tilde{A}^F(e)$. I choose the values of these parameters such that simulated moments of the initial population distribution as well as aggregate growth and urbanization until 1820 equal the corresponding moments in the data. In what follows, I describe the targeted moments corresponding to each parameter. Although the identification of each parameter depends on the values of other parameters, it is still intuitive to think of the identification as if each parameter were chosen to match one moment in the data.

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 0.46% per year between 1800 and 1820 (Weiss, 1992). This procedure pins down $f = 0.63$.

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 the same size groups of non-farm locations differ by the same number in the model. This implies $\gamma = 0.010$. Finally, Lemma 1 establishes a decreasing relationship between the parameter μ and the convergence of city sizes. To obtain the same degree of convergence between 1790 and 1820, that is, the same coefficient of log size on log growth in the model as in the data, I need to set $\mu = 0.75$.

To find the elasticity of substitution β, recall that it plays a key role in generating urbanization, as discussed in Section 3.1.5. The lower the value of β, the higher the degree of complementarity between productivity and non-farm labor, hence the more productivity growth induces urbanization. Thus, I choose the value of β such that urbanization, measured by the fraction of population living at non-farm locations above 10,000 inhabitants, increases by the same factor in the model between 1790 and 1820 as in the data. This suggests setting $\beta = 0.37$.

Also note that an increase in $\bar{A}$ increases the share of non-farm workers in U.S. population, as it makes the non-farm sector more productive uniformly across space. Setting $\bar{A} = 0.004$ leads to an 1800 non-farm population share of 26%, which equals the share of non-farm employment estimated by Weiss (1992) for this year. To recover the productivity of each initial city k, I choose the combination of A_k that correctly predicts the population of cells in which these initial cities were located.[36]

As a higher value of Europe's land-adjusted farm productivity $\tilde{A}^F(e)$ implies a higher

[36]In principle, there could exist different combinations of A_k that rationalize the data. However, simulations suggest that, conditional on the values of structural parameters, the set of initial city productivities is uniquely identified.

real GDP in Europe, I choose the value of this parameter to match Europe's real GDP in the first period.

To identify the value of parameter ϕ that drives the scale of non-farm to farm shipping costs, note that lower non-farm costs imply a larger role of trade in the economy, hence a higher concentration of economic activity near trading routes. Therefore, I calibrate the value of ϕ such that my model replicates the spatial concentration of population in 1790, measured by the Theil index. This procedure leads to an estimate of $\phi = 0.24$. The fact that ϕ is lower than one is in line with the observation that transporting manufacturing goods was substantially cheaper at the time than transporting agricultural products (Herrendorf et al., 2012).

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 Baseline simulation

To assess the quantitative performance of the structural framework, this section studies the evolution of economic activity predicted by the model, and compares it to the one seen in the data. Since the model is simulated over 20 by 20 arc minute grid cells, comparing simulation results to the evolution of actual population requires assigning census data on county populations to these cells. Therefore, I distribute the population of each county across the grid cells it occupies based on the share of land belonging to each cell. In other words, I assume that population density was uniform within each county. Although it is unlikely that the distribution was exactly uniform, this assumption must lead to little bias since the most densely populated counties were small, therefore usually fully contained in a single cell.

Table 6 shows that the correlation between cells' population levels predicted by the model and those in the data is high, and the same is true for correlation between population per unit of land. Even though I only match the overall concentration of population, non-farm employment and the sizes of the pre-existing five cities in the first period, the correlation coefficient is already above 0.2 in 1790. As the dynamics of the model start playing out, correlation increases further to values between 0.6 and 0.65.[37] Looking at

[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. That is, the slow migration process does poorly in replicating the population movements seen in the data.

Figure 7, 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 actual cities such as Atlanta, Detroit and St. Louis in the model.

Besides featuring a high correlation with the data, the model can also replicate the qualitative features of the evolution of U.S. population and economic activity. In line with the data, the model predicts that economic activity expanded substantially to the West. By 1860, the population share of the Midwest increases to 29%, while the share of the Northeast is about 37%, and the share of the South is about 32% both in the model and in the data (Table 7).

To assess performance of the model in replicating urbanization as well as the locations and sizes of cities, I need to define the notion of a city in the model. Analogously with the 10,000-inhabitant threshold that I use in the data, I define a city in the model as a grid cell with non-farm population above 10,000 inhabitants.[38] This definition does not necessarily coincide with the administrative definition of cities, hence the number of cities or the levels of urbanization figures might differ between the model and the data. Yet, the model does a good job at predicting the number of cities forming, even by region. In 1860, the Northeast, the South and the Midwest have 43, 22 and 22 grid cells classified as cities in the model, respectively. The corresponding numbers in the data are 41, 14 and 17.

Figure 8 presents the evolution of urbanization in the model. In the figure, I normalize the U.S. urbanization rate in 1790 to one. Comparing the left panel of this figure to the left panel of Figure 3, one can see that the population share of cities started to increase rapidly after 1820 both in the model and in the data, although the model implies an evolution of urbanization that is somewhat too slow in the 1840s and 1850s. The right panel of Figure 8 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. This result indicates that *technology diffusion* is an important force of city formation: high-productivity cities in the Northeast spread their technology to their surroundings, which attracted most newly forming cities to these locations.

Just like in the data, the fraction of cities forming near trading routes is disproportionately high in the model (Table 8). Although the model somewhat underpredicts the share of new cities near confluences (69.0%, as opposed to 87.1% in the data), it almost exactly matches the fraction of cities appearing at water (95.4% in the model, 98.6% in the data) and those appearing near railroads (37.9% in the model, 40.0% in the data). Thus, the model does much better than a process of random assignment of cities to locations, or the slow migration process presented in Section 2. These results indicate that *trade* plays an important role in city formation. Cities formed at trading routes to benefit from these

[38]Note that this also coincides with the threshold above which dynamic externalities are present at the location.

locations' better access to other locations.

The model also predicts, in line with the data, that the pattern of city location was different in the Northeast than in the rest of the United States. As shown in Table 9, the model can replicate the fact that average distance from other cities and average distance from the closest neighbor were smaller in the Northeast, although the difference between the Northeast and the rest of the country is less striking than in the data. This result indicates that *competition for land* is important in city formation: non-Northeast cities appeared relatively far from each other to avoid tough competition from their neighbors, whereas tough competition was at least partly counterbalanced by spillovers from pre-existing high-productivity cities in the Northeast.

Finally, the model is successful at predicting the history of the largest U.S. cities. In 1860, New York, Philadelphia, Boston and Baltimore are the four largest cities both in the model and in the data, although the model stops short at replicating their ranking: whereas New York is the largest in the data, Boston and Philadelphia take over New York and occupy the first and second place in the model. However, Glaeser (2011) argues that the leading position of New York over Boston and Philadelphia heavily depended on its deep and protected harbor that could accommodate big clipper ships. Since these types of ships only appeared in the first half of the nineteenth century, they could not provide this advantage to New York in 1790. As this force is outside my model, the model's ranking of cities can be seen as an indication that New York could not have become the largest city in its absence.

5.2 The effect of railroads on city formation and welfare

How much railroads contributed to American economic growth and welfare 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 this 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 consequence, 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's work, I also account for the impact of railroads on U.S. output and welfare through railroads changing the dynamic evolution of cities.

To assess the aggregate impact of railroads on city formation, growth and welfare, I simulate the model under the assumption that no railroads were built in my period of

investigation. The absence of railroads directly increases shipping costs in 2.7% of all grid cells in 1840, in 4.2% of the cells in 1850, and in 10.9% of the cells in 1860. As I will argue, however, the absence of railroads affects a much larger share of U.S. locations indirectly through its effects on trade and competition among cities.

The results indicate that the U.S. economy would look very different in the absence of railroads. Figure 9 plots the 1860 distribution of population in this counterfactual scenario. As can be seen from the figure, the lack of railroads makes most of the population shift toward navigable rivers, whereas regions far from water remain low-density. Many of the large cities remain substantially smaller without railroads. The grid cell in which Chicago is located loses 15% of its population relative to the baseline, while Buffalo loses as much as 68%, and Boston loses 72% of its population in 1860. The number of cities forming, on the other hand, is larger without railroads, with new cities appearing along rivers in the Midwest and in the surroundings of Boston. Since many of the largest cities have become smaller, locations in their vicinities experience less tough competition, and can benefit from technology spillovers to grow larger. As a result, the absence of railroads makes the size distribution of cities more compressed.

The substantial impact that railroads have on the development of cities translates into a large effect on real GDP and consumers' welfare. Figure 10 shows the evolution of U.S. real GDP per capita, which equals consumers' per-period utility in the model. The absence of railroads decreases the growth rate of the U.S. economy between 1790 and 1860 by 27%, from 0.59% to 0.45% per year. As a result, real GDP is 9.3% smaller 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 9.3% loss in consumers' welfare in the steady state.

A substantial fraction of these gains arises due to railroads changing the spatial distribution of growth and, hence, the formation and development of cities. To show this, I simulate a version of the model in which productivity growth is uniform across space, both with and without railroads. In this alternative model, the absence of railroads leads to a 8.0% loss in real GDP in 1860, suggesting that differences in growth rates across space, arising from differences in innovation and dynamic externalities in cities, amplify the effect of railroads on output by a factor of 18%. It is, however, important to note that cities still form and develop in this alternative model, 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 18% number should be viewed as a lower bound on the contribution of cities to the effect of spatial frictions on output. To sum up, city development seems to be an influential factor driving the impact of transport infrastructure on economic growth, output and welfare.

6 Conclusion

This paper proposes a quantitative model of spatial growth to study how spatial frictions affect economic development. The flexible geographic structure of the model allows me to take the model to historical U.S. data at a high spatial resolution, and to measure how the distribution of population, city formation, urbanization and aggregate growth were driven by the availability of good transport infrastructure, in the form of railroads. The results suggest that railroads had a large effect on the locations and sizes of cities, U.S. output and growth. Moreover, city development substantially amplifies the effect relative to the case of uniform productivity growth across space, or relative to the effect found in static models of trade and geography.

To the best of my knowledge, mine is the first paper pointing to city formation and development as a factor that we cannot abstact from if we want to quantify how spatial frictions shape aggregate economic growth. Incorporating this new channel allows for better measurement of the impact of spatial frictions. As I have shown in Section 5.2, it even has the potential to contribute to the century-old debate of how much railroads mattered for 19th-century U.S. economic development.

Yet, studying the impact of 19th-century U.S. railroads is not the only possible application of the quantitative framework developed in this paper. Using the model to examine the emerging city structure and its relationship with growth in today's developing economies can be one fruitful direction for further research. Countries such as China, India or Brazil have seen massive improvements in their transport infrastructure recently (Faber 2014, Ghani et al. 2016), as well as large population reallocations, rapid urbanization and the formation of many new cities. How did these transport improvements affect the locations and sizes of cities? How did city development contribute to the gains from transport improvements? What would be the effects of proposed infrastructure projects, or the further economic integration of these countries with foreign markets? If applied to these countries, the setup proposed in this paper has the potential to answer these, as well as other similar 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. (2016): The welfare effects of transportation infrastructure improvements. Mimeo.

- Arkolakis, C., Costinot, A. and Rodríguez-Clare, A. (2012): New trade models, same old gains? *American Economic Review*, vol. 102(1), 94–130.
- 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.
- 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. (2015): Trade and labor market dynamics. *NBER Working Paper* 21149.
- Caliendo, L., Parro, F., Rossi-Hansberg, E., and Sarte, P. (2016): The impact of regional and sectoral productivity changes on the U.S. economy. Mimeo.
- 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.
- David, P. (2005): Real income and economic welfare growth in the early republic. *Stanford University Working Paper*.
- Desmet, K., Nagy, D., and Rossi-Hansberg, E. (2016): The geography of development. *Journal of Political Economy*, forthcoming.

- Desmet, K. and Rappaport, J. (2015): The settlement of the United States, 1800-2000: The long transition to Gibrat's Law. *Journal of Urban Economics*, forthcoming.
- 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.
- 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.
- 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. (2014): External integration, structural transformation and economic development: Evidence from Argentina 1870-1914. *NBER Working Paper* 20217.
- Fajgelbaum, P. and Schaal, E. (2016): Optimal transport networks in spatial equilibrium. Mimeo.
- Fogel, R. (1964): Railroads and American economic growth: Essays in econometric history. The Johns Hopkins Press.
- 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.
- 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.
- 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.
- Lebergott, S. (1996): Consumer expenditures: New measures and old motives. Princeton University Press.
- 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.
- 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. (2016): Commuting, migration and local employment elasticities. Mimeo.
- Nagy, D. (2013): Border effects and urban structure. Mimeo.
- Ngai, L. and Pissarides, C. (2007): Structural change in a multisector model of growth. *American Economic Review*, vol. 97, 429–443.
- Redding, S. (2016): Goods trade, factor mobility and welfare. *Journal of International Economics*, vol. 101, 148–167.
- 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.
- 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. (2016): Endogenous infrastructure development and spatial takeoff. *School of Economics and Finance Discussion Paper* 1601.
- Weiss, T. (1992): 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.

Tables and figures

Table 1: Fraction of cities forming at trading routes

	Fraction of cells	Fraction of land	Fraction of cities
water	39.1%	37.0%	98.6%
confluence	23.1%	20.7%	87.1%
early railroad	5.4%	5.2%	40.0%†

Cell is classified as "water" if part the cell, or part of a cell next to it, is a navigable river or lake, or the sea; classified as "confluence" if the cell, or a cell next to it, is surrounded by at least 3 cells with water; and classified as "early railroad" if the cell, or a cell next to it, was part of the rail network in 1840. †: out of cities appearing after 1840.

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		0.033**			0.031**	
		(0.004)			(0.005)	
early railroad			0.066**			0.061**
			(0.013)			(0.012)
Prod & amenities	No	No	No	Yes	Yes	Yes
No. of observations	7641	7641	7641	7641	7641	7641

Heteroskedasticity-robust standard errors in parentheses.
*: significant at 5%; **: significant at 1%.

Table 3: Slow migration process: simulation results

	Fraction of cells	Fraction of land	Fraction of cities
water	39.1%	37.0%	49.3%
confluence	23.1%	20.7%	33.3%
early railroad	5.4%	5.2%	60.0%

Dependent variable: newcity (simulated)						
	(1)	(2)	(3)	(4)	(5)	(6)
water	0.004			-0.006		
	(0.002)			(0.004)		
confluence		0.006			-0.004	
		(0.003)			(0.004)	
early railroad			0.019**			0.008
			(0.004)			(0.004)
Prod & amenities	No	No	No	Yes	Yes	Yes
No. of observations	7641	7641	7641	7641	7641	7641

Heteroskedasticity-robust standard errors in parentheses.
*: significant at 5%; **: significant at 1%.
Cell is classified as "water" if part the cell, or part of a cell next to it, is a navigable river or lake, or the sea; classified as "confluence" if the cell, or a cell next to it, is surrounded by at least 3 cells with water; and classified as "early railroad" if the cell, or a cell next to it, was part of the rail network in 1840.

Table 4: Clustering of cities

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

Table 5: Discontinuity in the 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

Heteroskedasticity-robust standard errors in parentheses.
*: significant at 5%; **: significant at 1%.

Table 6: Correlation of population between the model and the data

Year	Correlation of population (levels)	Correlation of population (per unit of land)
1790	0.214	0.280
1800	0.390	0.464
1810	0.431	0.501
1820	0.474	0.538
1830	0.560	0.605
1840	0.629	0.618
1850	0.667	0.632
1860	0.641	0.587

Table 7: Large regions' shares in total population, 1860

Region	Model	Data
Northeast	38.3%	36.3%
South	32.4%	32.7%
Midwest	28.8%	29.0%
West	0.4%	2.0%

Table 8: Fraction of cities forming at trading routes in the model

	Fraction of cells	Fraction of land	Fraction of cities
water	39.1%	37.0%	95.4%
confluence	23.1%	20.7%	69.0%
early railroad	5.4%	5.2%	37.9%†

Cell is classified as "water" if part the cell, or part of a cell next to it, is a navigable river or lake, or the sea; classified as "confluence" if the cell, or a cell next to it, is surrounded by at least 3 cells with water; and classified as "early railroad" if the cell, or a cell next to it, was part of the rail network in 1840. †: out of cities appearing after 1840.

Table 9: Clustering of cities in the model

	Northeast	Rest of U.S.
Average distance from cities in the region (miles)	346.4	359.3
Average distance from closest city in region (miles)	44.0	62.4

Figure 1: Population per square mile and mean center of population (red)

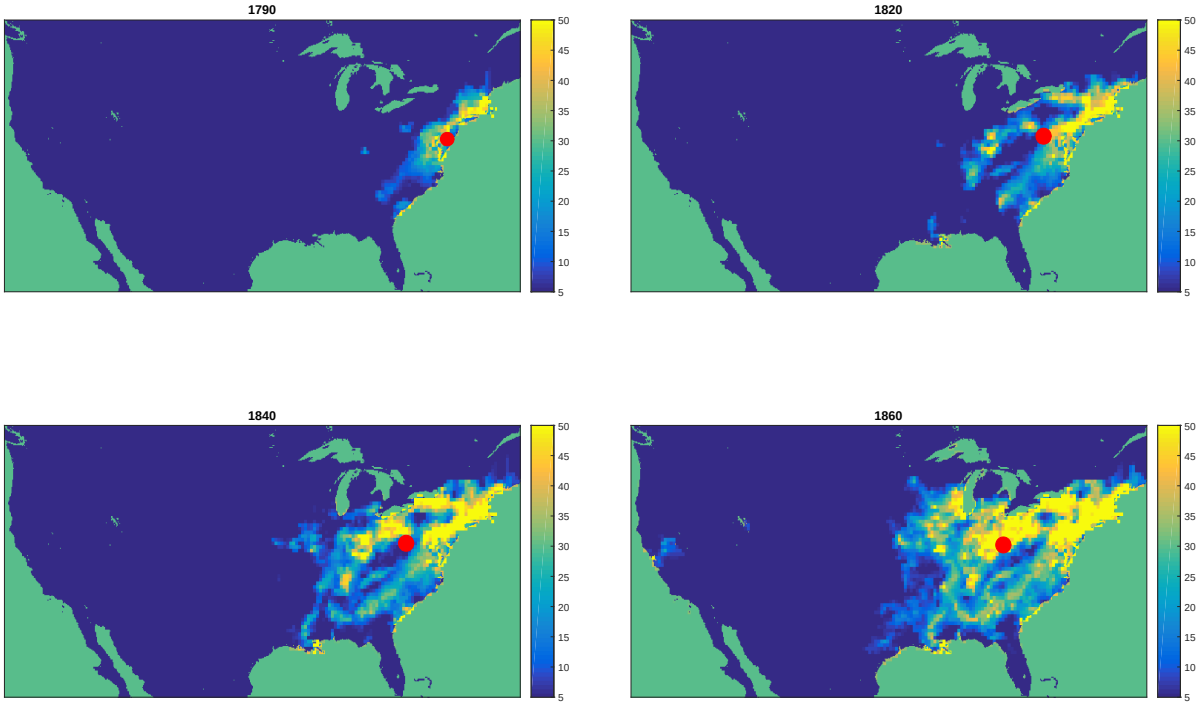


Figure 2: Population of the four large U.S. regions

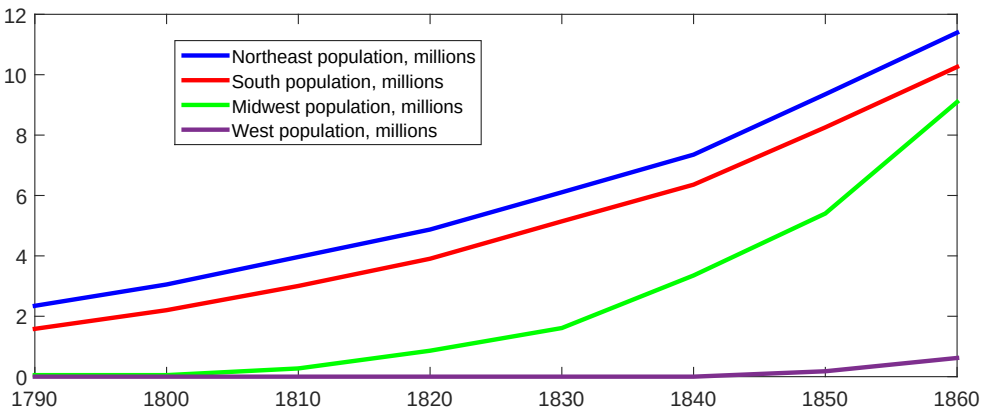


Figure 3: Urbanization in the U.S.

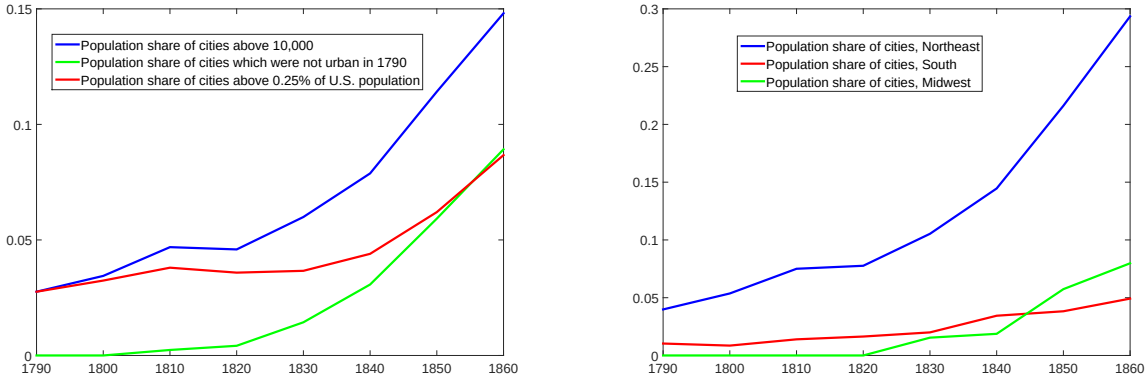


Figure 4: U.S. cities forming between 1790 and 1860

Figure 5: City growth versus size among towns (left) and cities (right)

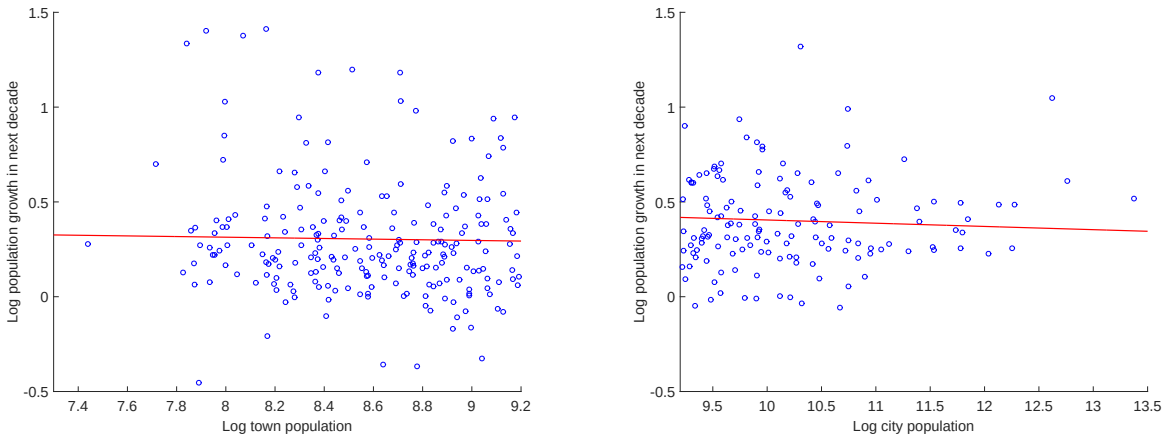


Figure 6: Timing of events in the model

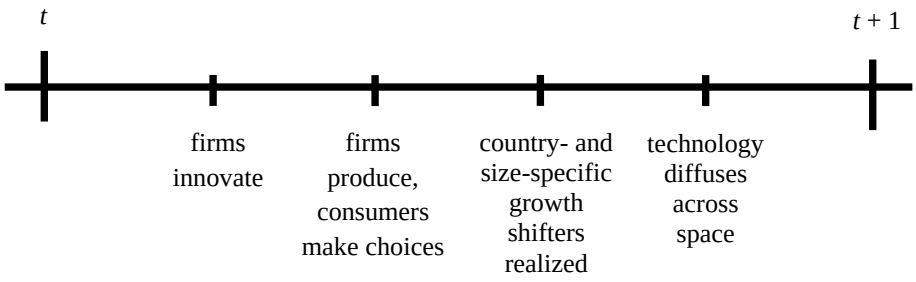


Figure 7: Population distribution in 1860: model versus data

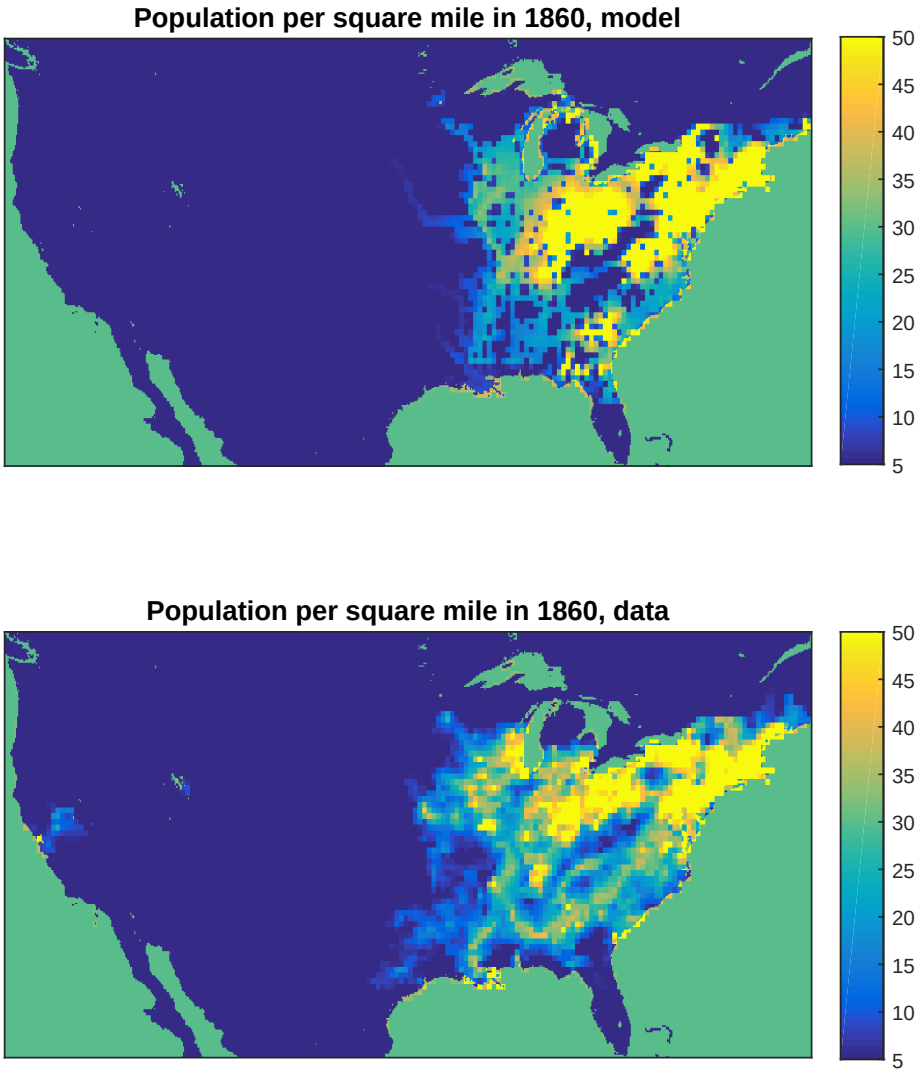


Figure 8: Urbanization in the model

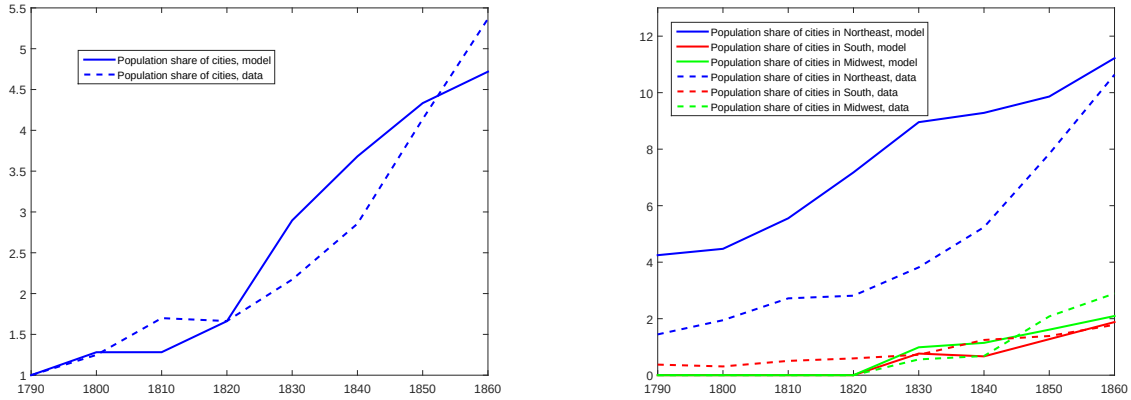
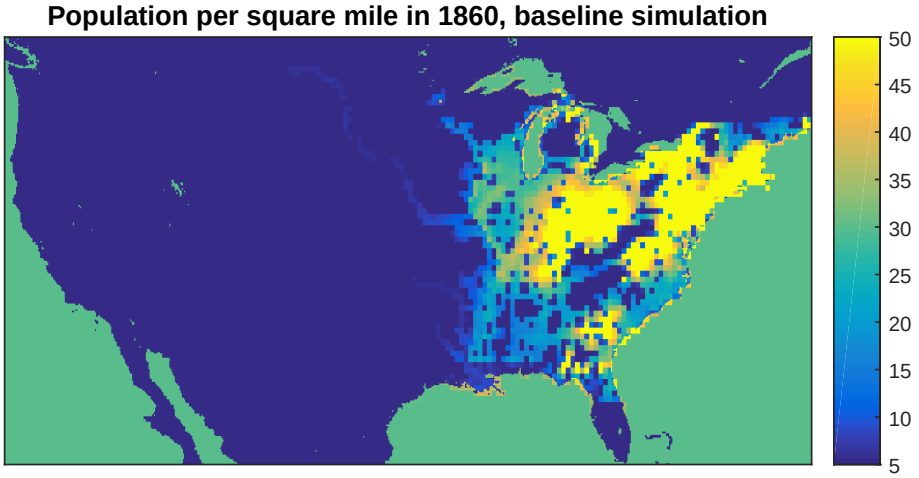


Figure 9: Population distribution in 1860: baseline versus no railroads



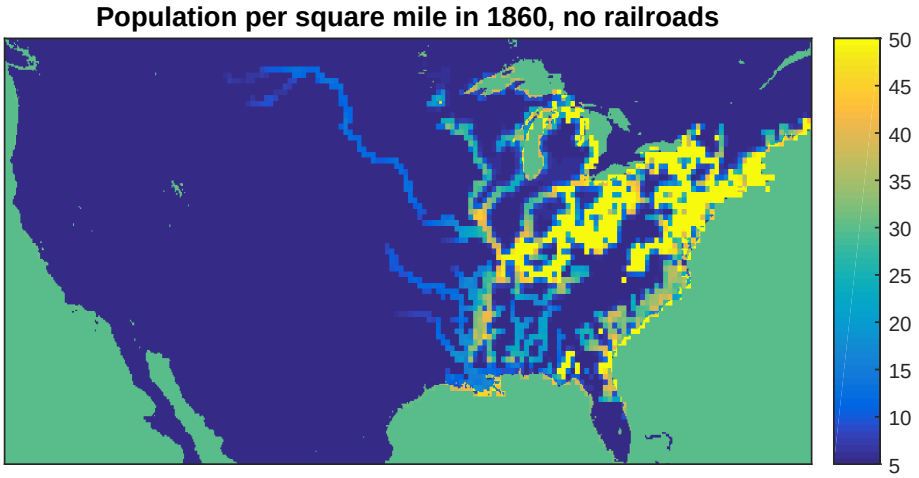
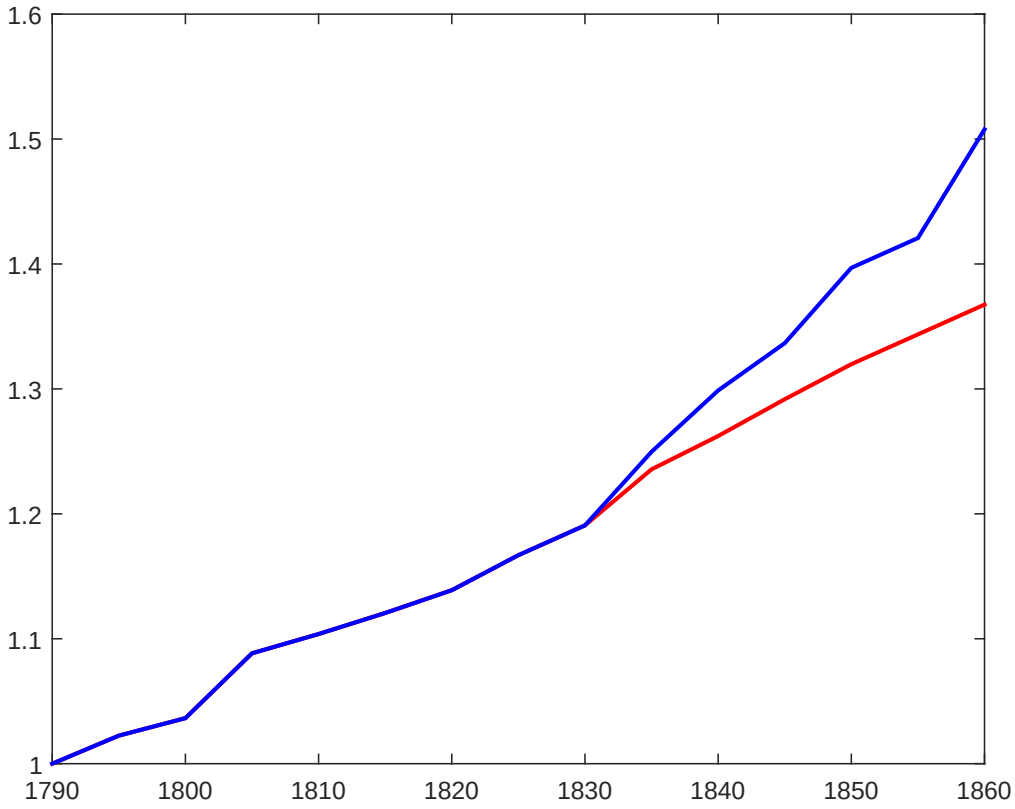


Figure 10: U.S. real GDP per capita, baseline simulation (blue) versus no railroads (red)



A Theory appendix

This appendix supplements the theoretical framework presented in Section 3 in three respects. First, Section A.1 shows how the population distribution in period t can be obtained by solving a system of three equations, and describes an algorithm to solve these equations. Next, Section A.2 provides the proofs of Lemmas 1 and 2. Finally, Section A.3 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 3.1.

A.1 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) and (6) to (20) –, 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 (15), the utility of a farmer living at a trading place $s \in S_c$ equals the utility of a non-farmer (3),

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

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

$$p_t^F(s) = A^F(s)^{-1} \left[\frac{L_t^F(s)}{H_t(s)}\right]^{1-\alpha} w_t(s). \tag{23}$$

Plugging this back into (2), one obtains

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

where I used (15) 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

$$\sigma_t(r) = argmax_{s \in S_c} \varsigma_t(r,s)^{-1} A^F(s)^{-1} H(s)^{-(1-\alpha)} L_t^F(s)^{1-\alpha}. \tag{24}$$

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 (13), 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

$$L_t^F(r) = \varsigma_t(r, \sigma_t(r))^{-\frac{1}{1-\alpha}} \frac{H_t(r)}{H_t(\sigma_t(r))} \left[\frac{A^F(r)}{A^F(\sigma_t(r))}\right]^{\frac{1}{1-\alpha}} L_t^F(\sigma_t(r)) \tag{25}$$

We are only left with finding the distribution of farmers across trading places since equations (24) and (25) 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 (16), the non-farm market clearing condition (18), 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:*

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

$$\int_S U_{c(u),t}^{-\frac{(\epsilon-1)^2}{2\epsilon-1}} \frac{A^F(u)^{\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}} \left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}} \bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{(2\epsilon-1)}}}{(1-\nu)[\epsilon(1-\mu)+\mu] + (\epsilon-1)\mu\left[1-\nu\bar{c}_t(u)^{\beta-1}\right]} \left[\int_{\sigma_t^{-1}(u)} L_t^F(r)\,dr\right] \tau_t(u,s)^{1-\epsilon}\,du \tag{26}$$

where $L_t^F(r)$ as a function of $L_t^F(u)$ is given by (25), $\sigma_t^{-1}(u)$ is given by (24), $\bar{c}_t(s)$ is given by (8), and $\kappa = \epsilon^{1-\epsilon}(\epsilon-1)^{\epsilon-2}\mu^{-1}f^{-1}$ is a constant.

Proof. See Appendix A.2.

□

Next, non-farm population at trading places can be obtained using the distribution of farm population and farm market clearing (17). Combining (17) with (11) and (12) yields

$$L_t^N(s) = \frac{\nu}{(1-\nu)+\mu(\epsilon-1)\left[\mu+\epsilon(1-\mu)+\frac{\epsilon}{\bar{c}_t(s)^{1-\beta}-1}\right]^{-1}} \int_{\sigma_t^{-1}(s)} L_t^F(r)\,dr \tag{27}$$

as the non-farm population of location s. Note that, by (27), 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.

Finally, countries' utility levels U_{ct} can be obtained by imposing national labor market conditions (19). Hence, we have the following lemma.

Lemma 3. *Given the values of parameters, geography S, countries' population levels $\bar{L}_{ct}$ and growth shifters f_{ct}, functions $H_t(\cdot)$, $A^F(\cdot)$, $\tau_t(\cdot,\cdot)$ and $\varsigma_t(\cdot,\cdot)$ and the distribution of non-farm productivity $A_t^N(\cdot)$, the system of three equations (19), (26) and (27) 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 (19), (26) and (27). In the two-country world considered in this paper, I solve this system by an iteration algorithm, similar to the one applied in Nagy (2013). The algorithm consists of the following steps.

1. Calculate the non-farm productivity of Europe, using equation (22).
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 (26). 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 (24) and (25) 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 (27) for non-farm populations $L_t^N(s)$.
8. Check if national labor market clearing conditions (19) 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.2 Proofs

Proof of Lemma 1

The firm's optimal innovation decision (10) implies

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

from which, using (8), we obtain

$$\ell_t^I(s)^{1-\mu} = (1-\mu)^{1-\mu}(\epsilon-1)^{1-\mu} f^{1-\mu}\left[1-\bar{c}_t(s)^{\beta-1}\right]^{1-\mu}. \tag{28}$$

Equation (27) 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 (28) and rearranging yields the result.

Proof of Lemma 2

Equation (16) provides the price index at any trading place s. Combining it with equations (6) and (12) yields

$$P_t(s)^{1-\epsilon} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1}\int_S \frac{\bar{c}_t(u)^{1-\epsilon}}{(\epsilon-1)\left(\mu\bar{c}_t(u)^{\beta-1}+1-\mu\right)+1} w_t(u)^{1-\epsilon} L_t^N(u)\,\tau_t(u,s)^{1-\epsilon}\,du. \tag{29}$$

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

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

where I used equations (15) and (23). Plugging this into the left-hand side of equation (29) implies

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

$$\tilde{\kappa}\int_S \frac{\bar{c}_t(u)^{1-\epsilon}}{(\epsilon-1)\left(\mu\bar{c}_t(u)^{\beta-1}+1-\mu\right)+1} w_t(u)^{1-\epsilon} L_t^N(u)\,\tau_t(u,s)^{1-\epsilon}\,du \tag{30}$$

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

Equation (18) provides the market clearing condition for each non-farm good at any trading place s. Combining it with equations (6) and (7) yields

$$\bar{c}_t(s)^{\epsilon-1} w_t(s)^{\epsilon} = \nu \epsilon^{-\epsilon} (\epsilon-1)^{\epsilon-1} f^{-1} \int_S P_t(u)^{\epsilon-1} I_t(u) \tau_t(s,u)^{1-\epsilon} du \tag{31}$$

where

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

is the sum of farmers' and non-farm workers' income at trading place u. Equations (2) and (13) allow me to rewrite income as

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

from which, by equation (23), 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 (17) 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 (31) can be written as

$$\bar{c}_t(s)^{\epsilon-1} w_t(s)^{\epsilon} = \tilde{\kappa} \int_S U_{c(u),t}^{-\frac{\epsilon-1}{\nu}} \frac{A^F(u)^{(\epsilon-1)\frac{1-\nu}{\nu}} \left[\frac{L_t^F(u)}{H_t(u)} \right]^{-(1-\alpha)(\epsilon-1)\frac{1-\nu}{\nu}}}{(\epsilon-1)\left(\mu \bar{c}_t(u)^{\beta-1} + 1 - \mu\right) + 1} w_t(u)^{\epsilon} L_t^N(u) \tau_t(s,u)^{1-\epsilon} du \tag{32}$$

where $\tilde{\kappa} = \left(\frac{\epsilon-1}{\epsilon}\right)^{\epsilon-1} f^{-1}$, and I used equations (3), (15) and (23) to substitute for the price index on the right-hand side.

In what follows, I show that equations (30) and (32) reduce to a single equation. To see this, note first that, since non-farm shipping costs are symmetric, $\tau_t(s,u) = \tau_t(u,s)$.

Second, guess that wages at location s take the form

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

where ι_1, ι_2, ι_3 and ι_4 are constants. Plugging the guess into (30) and (32), 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

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

and equations (30) and (32) reduce to equation (26).

A.3 A model with home consumption

This section presents a version of the model of Section 3.1 in which consumers consume goods at their residential location, not at the trading place. For the goods that consumers 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 good they produce, the assumption leads to savings in shipping costs as consumers do not need to ship the fraction of the good that they consume to the trading place. Note that non-farm workers are not affected by these changes as they always live where they trade. In what follows, I describe the farmer's problem, as well as the set of equilibrium conditions that change relative to Section 3.1.

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

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

and

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

where the definitions of all variables are the same as in Section 3.1. 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 $\tau_t(s,r)$ 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

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

Combining this with equations (3), (15) and (23) implies that the trading place chosen by farmers living at location r is

$$\sigma_t(r) = argmax_{s\in S_c}\left[\frac{\tau_t(s,r)^{\nu}}{\varsigma_t(r,s)^{1-\nu}}\right]^{-1}\varsigma_t(r,s)^{-1}\,A^F(s)^{-1}\,H(s)^{-(1-\alpha)}\,L_t^F(s)^{1-\alpha} \tag{24'}$$

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

$$L_t^F(r) = \left[\frac{\tau_t(\sigma_t(r),r)^{\nu}}{\varsigma_t(r,\sigma_t(r))^{1-\nu}}\right]^{-\frac{1}{1-\alpha}}\varsigma_t(r,\sigma_t(r))^{-\frac{1}{1-\alpha}}\,\frac{H_t(r)}{H_t(\sigma_t(r))}\left[\frac{A^F(r)}{A^F(\sigma_t(r))}\right]^{\frac{1}{1-\alpha}}L_t^F(\sigma_t(r)). \tag{25'}$$

As a result, equations (26) and (27) become

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

$$\int_S \bar{U}_{c(u),t}^{-\frac{(\epsilon-1)^2}{2\epsilon-1}}\frac{A^F(u)^{\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\left[\frac{L_t^F(u)}{H_t(u)}\right]^{-(1-\alpha)\frac{1-\nu}{\nu}\frac{(\epsilon-1)^2}{2\epsilon-1}}\bar{c}_t(u)^{-\frac{\epsilon(\epsilon-1)}{(2\epsilon-1)}}}{(1-\nu)\left[\epsilon(1-\mu)+\mu\right]+(\epsilon-1)\,\mu\left[1-\nu\bar{c}_t(u)^{\beta-1}\right]}\cdot$$

$$\left[\int_{\sigma_t^{-1}(u)}\frac{\tau_t(u,r)^{\nu}}{\varsigma_t(r,u)^{1-\nu}}L_t^F(r)\,dr\right]\tau_t(u,s)^{1-\epsilon}\,du \tag{26'}$$

and

$$L_t^N(s) = \frac{\nu}{(1-\nu)+\mu(\epsilon-1)\left[\mu+\epsilon(1-\mu)+\frac{\epsilon}{\bar{c}_t(s)^{1-\beta}-1}\right]^{-1}}\int_{\sigma_t^{-1}(s)}\frac{\tau_t(s,r)^{\nu}}{\varsigma_t(r,s)^{1-\nu}}L_t^F(r)\,dr, \tag{27'}$$

respectively.

A comparison of equations (24'), (25'), (26') and (27') to their counterparts (24), (25), (26) and (27) shows how additional shipping costs $\tau_t(s,r)^{\nu}$ and shipping cost savings $\varsigma_t(r,s)^{1-\nu}$ alter the equilibrium relative to Section 3.1. If these shipping cost changes exactly counterbalance each other, that is, $\frac{\tau_t(s,r)^{\nu}}{\varsigma_t(r,s)^{1-\nu}} = 1$, then the equilibrium population, productivity and utility levels of the two models become identical. This is stated formally in the next proposition.

Proposition 1 (Isomorphism with model of Section 3.1)**.** *Assume* $\frac{\tau_t(s,r)^{\nu}}{\varsigma_t(r,s)^{1-\nu}} = 1$ *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 3.1.

In the baseline calibration of the model of Section 3.1, we have $\nu = 0.75$ and $\tau_t(s,r) = \varsigma_t(r,s)^{\phi}$. We obtain the isomorphism whenever $\frac{\varsigma_t(r,s)^{0.75\phi}}{\varsigma_t(r,s)^{0.25}} = 1$, that is, $0.75\phi - 0.25 = 0$, or $\phi = 1/3$. The value of ϕ used in the calibration is indeed close to this value. Therefore, Proposition 1 implies that the difference between the two models is small, and changing the assumption about where consumption happens in space is unlikely to alter the results substantially.

B Data appendix

This appendix describes the datasets used to document the major patterns of 19th-century U.S. urban history, to take the model to the data, and to evaluate the model's fit. I use geographical 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 Geographical Information Software (GIS), and combine it with other sources of geographical data to determine whether any given cell is at a major body of water 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.

Major bodies of water. 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 body of water. 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.[39] I georeference these maps to the 20 by 20

[39] Although the first railroads started to be built in the late 1820s, there only existed a small number of

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.[40]

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. 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.[41] 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.[42] 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.[43] I use the county 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

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.

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

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

[42]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.

[43]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.

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,[44] 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 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.

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

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

Source: FAO GAEZ database. Filters "low-input level" and "rain-fed" have been applied for each crop.

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	269	111.7	61.0
Number of days with mean temperature above 10 °C	23.4	365	201.9	58.3
Annual precipitation (mm)	44.8	2607.8	694.7	416.8

Source: FAO GAEZ database. All data are for 1960, except annual temperature range which is for the period between 1961 and 1990.