NBER WORKING PAPER SERIES

PERSISTENCE AND PATH DEPENDENCE IN THE SPATIAL ECONOMY

Treb Allen
Dave Donaldson

Working Paper 28059
http://www.nber.org/papers/w28059

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2020, Revised November 2022

We are grateful to our discussants, Klaus Desmet, Jonathan Eaton, Je rey Lin, David Nagy, and David Weinstein, as well as to Rodrigo Ad˜ao, Kristian Behrens, Lorenzo Caliendo, Arnaud Costinot, Don Davis, Jonathan Dingel, Gilles Duranton, Cecile Gaubert, Rick Hornbeck, Xiangliang Li, Vincent Lohmann, Robert Margo, Nathan Nunn, Michael Peters, James Rauch, Steve Redding, Stuart Rosenthal, John Sturm, Ivan Werning and many seminar participants for comments that improved this paper, as well as to Nicholas Crafts and Alexander Klein for sharing data. Richard Dionne, Saptarshi Majumdar and Yunus Tuncbilek provided excellent research assistance. This material is based upon work supported by the National Science Foundation under grant SES-1658838. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Persistence and Path Dependence in the Spatial Economy
Treb Allen and Dave Donaldson
NBER Working Paper No. 28059
November 2020, Revised November 2022
JEL No. C33,C62,F1,R11,R13,R23

ABSTRACT

How much of the spatial distribution of economic activity today is determined by history rather than by geographic fundamentals? And if history matters for spatial allocations, does it also matter for overall e ciency? This paper develops a forward-looking dynamic framework for the theoretical and empirical study of such questions. We derive conditions on the strength of agglomeration externalities under which equilibria are unique and yet temporary historical shocks can have particularly persistent, or even permanent (i.e. path-dependent) consequences. When estimated using U.S. data from 1800-2000, this model displays multi-century persistence from small and temporary shocks as well as path dependence (with large aggregate welfare e ects) throughout much of our estimated parameter range.

Treb Allen
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
treb@dartmouth.edu

Dave Donaldson
Department of Economics, E52-554
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
ddonald@mit.edu

1 Introduction

Economic activity today is staggeringly concentrated. For example, more than $1/6^{th}$ of value-added in the United States is produced in just three cities that occupy less than $1/160^{th}$ of its land area. Perhaps even more remarkable are the historical accidents that may have determined the location of these three cities—one was a Dutch trading post, one a *pueblo* for 22 adult and 22 children settlers designated by a Spanish governor to honor the angels, and one a river mouth known to Algonquin residents for its wild garlic (or, *chicago-ua*).

There is no shortage of examples in which the quirks of history appear to influence the current location of economic activity through either *persistence*—the long-lived dependence of current outcomes on temporary events—or *path dependence*—where temporary events can permanently shape long-run outcomes; see Nunn (2014) and Voth (2021). But how widespread should we expect these phenomena to be in the spatial economies around us? Going further, "does history matter only when it matters little?"—in Rauch's (1993) phrase—because it serves merely to reshuffle the current location of economic activity without much affecting aggregate efficiency?

In this paper we develop a new framework designed to shed light on these questions and apply it to data from the United States between 1800 and 2000. Our model features agglomeration externalities, forward-looking agents, and many heterogeneous locations that interact through costly trade and migration. We derive conditions under which such an environment can feature unique dynamics that nevertheless display substantial persistence and even path dependence. Finally, our simulations, based on our estimated parameter values, display exactly such phenomena for the U.S. spatial economy: small historical shocks leave a sizable trace for several centuries and much of our estimated parameter range implies that such shocks can cause large and permanent differences in long-run aggregate welfare.

To arrive at this conclusion, we begin in Section 2 by describing a dynamic model of economic geography that combines essential features from two generations of work in the field. An earlier tradition—pioneered by Krugman (1991), Matsuyama (1991), Fujita et al. (1999), and Ottaviano (2001)—combined agglomeration externalities with infinitely-lived and forward-looking agents who inhabit a small number of symmetric locations. More recent work—such as that by Desmet et al. (2018), Caliendo et al. (2019), and Kleinman et al. (2021)—has pioneered the study of more empirically realistic settings with many locations that have arbitrarily heterogeneous characteristics such as trade costs, migration costs, productivities and amenities. But to date it has done so without the combination of agglomeration externalities and forward-looking agents that is necessary to embrace both sides of the "history" (i.e. path dependence) versus "expectations" (i.e. the potential for self-fulfilling equilibria based on forward-looking behavior) trade-off that was central to the earlier tradition.

In this model environment, we obtain three new theoretical results about dynamic economic geography models. Our first result characterizes a condition for (bounded) equilibria to be unique, regardless of the underlying path of geographic fundamentals, as is important for the quantitative questions that we pose here. Indeed, we develop this result by first providing a new statement about uniqueness in general economic systems that feature an arbitrary set of variables (such as populations, wages, prices, etc.) that interact nonlinearly in both forward- and backward-looking manners across heterogeneous entities (such as locations). Our second result highlights how temporary shocks may be particularly persistent—that is, feature a slow rate of convergence to a steady-state—when an economy gets close to the parameter threshold at which uniqueness is not guaranteed. Finally, our third result characterizes necessary (and "globally" sufficient) conditions for the economy to feature multiple steady-states, which then creates the potential for path-dependent impacts of a temporary shock that could push an economy toward a permanently different outcome.

The conditions in these results hinge on the strength of agglomeration forces relative to dispersion forces and the extent to which agents discount the future. However, we draw a new distinction between contemporaneous agglomeration spillovers, which operate within the same period, and historical spillovers, which operate with a lag. In particular, our condition for uniqueness depends on the two types of spillovers differently—and, indeed, in a region of the parameter space that is empirically relevant to our application, this condition depends only on the contemporaneous version of spillovers—whereas it is the sum of contemporaneous and historical spillovers that matters for the existence of multiple steady-states. As a result, there exists a plausible parameter range in which transition paths are known to be unique and yet still have the potential to generate the rich phenomena of path dependence.

Section 3 turns to our empirical application, which draws on long-run spatial data available for the United States from 1800-2000. We estimate both contemporaneous spillover elasticities, which have been the focus of recent work, and historical spillovers, which have received far less attention. Our estimating equations take the familiar form of a multi-location labor supply and demand system—as in the canonical Rosen-Roback tradition (Rosen 1979, Roback 1982, Glaeser 2008) but augmented to allow for interactions across locations due to costly trade and migration. Doing so requires estimates of historical bilateral migration and trade costs, which we obtain by using migration data from individual-count Census records and trade flow data that we have digitized from historical records on intranational commodity shipments and a novel non-linear least squares approach to estimation.

Despite this added empirical flexibility, parameter identification—even with an underlying potential for multiplicity—is still assured via familiar exclusion restrictions of the sort discussed by Roback (1982), expressed as time-varying versions of these restrictions in our case. For the locational labor supply equation, which is identified from demand-side variation, we use shifters of agricultural productivity due to the changing importance of certain crops over time and the advent of higher intensity cultivation methods. And for the locational labor demand equation we use shifters of the relevance of temperature extremes over time, which plausibly have changed the amenity value of certain locations, and hence labor supply, due to the development of technologies such as air conditioning. Our estimates imply modest productivity spillovers, but an important role for positive historical spillovers on amenities—which are, as we show, consistent with models that feature durable locational investments, for example in housing. These values are within the parameter region that corresponds to uniqueness, slow persistence, and the potential for path dependence.

Based on these parameter estimates we turn in Section 4 to a simulation exercise that is designed to illustrate the role that historical shocks can play in a spatial economy. Amidst the so-called "Technological Revolution" at the dawn of the 20th Century (c.f. Landes 2003) it seems plausible that innovations such as electrification and the automobile had differential impacts across space for reasons that could be partially attributed to chance. For example, Henry Ford was born on a farm near Detroit, and Thomas Edison chose the 1901 Pan-American Exposition to demonstrate mass illumination via his new AC power, earning the host city of Buffalo its nickname, the "City of Light". Inspired by such anecdotes of happenstance, our counterfactual exercise asks what would have happened to the trajectories of two similar locations if their 1900 productivity fundamentals were randomly swapped, while holding all other exogenous characteristics constant both before and after 1900. In practice, we pair locations on the basis of their closest match in terms of 1900 population—for example, Buffalo (with a population of 436,000 in 1900) is paired with Cincinnati (412,000). In order to derive general lessons from such counterfactual swaps, we conduct one hundred simulations in which every location has an equal chance of either drawing its factual 1900 productivity or its counterfactual swap partner's 1900 productivity.

Even these relatively modest counterfactual swap histories turn out to have dramatic consequences. For example, across our simulations the median location has an elasticity (when estimated using technology shocks as an instrumental variable) of 0.37 between its population in 2000 and its population in 1900—so that a 10% drop in population due to an unfavorable but one-off productivity shock leaves the location about 4% smaller even a century later. And while trade and migration opportunities mean that the present discounted value of residing in a location is less affected by local historical shocks, we find that this welfare persistence elasticity is still 0.09 for the median location.

Simulating the economies forward into the future—undoubtedly a heroic exercise, but one that illustrates the workings of a model like ours—we find that the long arm of history can reach very far into the future. When letting our simulations run forwards within the range of historical spillover elasticities consistent with the 95% confidence interval of our estimates we find that our historical swaps leave permanent impacts throughout much (but not all) of this range. This indicates that our model U.S. spatial economy is perched on the cusp of a bifurcation between a plausible region of the parameter space in which temporary, local shocks will and will not leave a path-dependent trace on the spatial distribution of economic activity. Perhaps surprisingly, this bifurcation is also consequential in terms of long-run efficiency—many of our alternative random swap configurations result in permanent aggregate welfare levels that differ by a factor of almost two. Our answer to Rauch's (1993) question is therefore that when history matters, it seems to matter a great deal.

Related literature

These findings shed new light on a number of strands of related work. First, we are inspired by an empirical literature that documents examples of spatial persistence and path dependence, or lack thereof, in the aftermath of historical events in a vast array of settings. Seminal work by Davis & Weinstein (2002, 2008) and Bleakley & Lin (2012, 2015) is emblematic of such lessons since Bleakley & Lin (2012, 2015) demonstrate relatively persistent (multi-century) impacts of long-obsolescent shipping technologies in the U.S. whereas Davis & Weinstein (2002, 2008) find that World War II bombing left only a relatively transitory (multi-decade) spatial trace in Japan. Wider examples from the U.S. alone include enduring impacts of slavery (Nunn 2008), political boundaries (Dippel 2014), flooding (Hornbeck & Naidu 2014), mining activity (Glaeser et al. 2015), fire damage (Hornbeck & Keniston 2017), frontier exposure (Bazzi et al. 2020), immigration (Sequeira et al. 2020), and war destruction (Feigenbaum et al. 2022)—among many other factors (see, e.g., Kim & Margo 2014).[1]

Our findings clarify the conditions under which one could expect spatial persistence and path dependence to arise, which may rationalize the heterogeneous effects seen in prior work. More generally, much of the above literature is primarily interested in the hypothesis that

[1]Further afield, Dell (2010) documents persistent negative effects of forced labor institutions in Peru, Redding et al. (2011) uncover evidence for persistence in the location of airline hubs amidst the division and reunification of Germany, Jedwab & Moradi (2016) find persistent impacts of colonial railroads throughout most of Africa, Hanlon (2017) illustrates a long-lived spatial imprint resulting from the interruption of supplies to Britain's cotton textile industry cities during the U.S. Civil War, Henderson et al. (2018) describe how the differing extent to which physical geography attributes matter today for early and late developing countries is consistent with long persistence, Michaels & Rauch (2018) highlight the differing extents of persistence of Roman towns in England and France, and Dell & Olken (2020) document the enduring industrial development around sites of colonial investment in Indonesia.

historical shocks leave persistent traces on the location of economic activity because fundamentals themselves are persistent (e.g. that historical institutions affect modern economies because they affect modern institutions, which affect modern productivity). So it is vulnerable to the critique—discussed in Nunn (2014) and Voth (2021)—that one may expect *any* temporary shock to fundamentals to cause a persistent geographic impact due to the logic of agglomeration and endogenous spatial lock-in. Our results can be used to assess such concerns by benchmarking the amount of spatial persistence and path dependence one may expect even in the absence of persistent fundamentals.

Second, on the theory side, our goal has been to build a bridge between two prominent strands of dynamic spatial modeling that have flourished in the last thirty years. An earlier wave built the foundations of economists' understanding of path-dependent geographic settings by combining agglomeration forces with forward-looking behavior.[2] However, it typically did so via deliberately small-scale and simplified models that were designed to maximize qualitative insights.

A more recent tradition has instead pioneered the study of models that are sufficiently flexible as to admit calibration to high-dimensional empirical settings with realistic geographies.[3] However, the ability to incorporate both forward-looking agents and local economies of density has so far lagged behind. Unsurprisingly, therefore, this literature has not yet focused on the study of path dependence that animates our paper. Our new theoretical results are designed to make progress towards this goal. For example, we build on Desmet et al. (2018) by adding forward-looking mechanisms, general migration frictions, and (as concerns the multiplicity of steady-states) an understanding of global necessity and not just sufficiency.[4] We build on Caliendo et al. (2019)'s model of forward-looking migration behavior by deriving conditions for uniqueness, persistence, and the multiplicity of steady-states, and do so under the more general environment in which (both static and dynamic) agglomeration externalities are present. And we provide a complementary result about persistence to that in Kleinman et al. (2021) by studying the full non-linear properties of our model (rather than a linearized version of it). By providing a tractable dynamic framework able

[2]Canonical examples of this approach—albeit often with a focus on the isomorphic problem of agglomeration across sectors rather than space—include Krugman (1991), Matsuyama (1991), Fukao & Benabou (1993), Rauch (1993), Fujita et al. (1999), Puga (1999), Herrendorf et al. (2000), Baldwin (2001), Ottaviano (2001), Ottaviano et al. (2002), Robert-Nicoud (2005), and Baldwin et al. (2011).

[3]Dynamic examples include Artuç et al. (2010), Desmet & Rossi-Hansberg (2014), Desmet et al. (2018), Caliendo et al. (2019), Nagy (forthcoming), Kleinman et al. (2021). Static frameworks featuring realistic geographies include Roback (1982), Glaeser (2008), Allen & Arkolakis (2014), and Ahlfeldt et al. (2015); see Redding & Rossi-Hansberg (2017) for a review.

[4]Indeed, apart from a difference in the assumed depreciation schedule, a special case of our model with myopic agents, restricted migration costs and no historical amenity spillovers is formally isomorphic to the framework of Desmet et al. (2018).

to incorporate rich geographic heterogeneity and characterizing its properties, we hope to facilitate the study of the role of history in shaping the evolution of the spatial economy.[5]

Due to the limits of data in our historical context, our analysis omits some features that have proved important in this prior work. These include the spatial diffusion of knowledge in Desmet et al. (2018), the input-output structure of Caliendo et al. (2019), and the presence of local landlords who accumulate capital in Kleinman et al. (2021). However, our new result about uniqueness in general dynamic spatial models (Theorem 1) may prove useful for applications that feature these extensions. Doing so offers the prospect of quantifying additional mechanisms that have featured in the earlier tradition of economic geography modeling—see, for example, Baldwin & Forslid (2000) on innovation, Krugman & Venables (1995) on input-output loops, and Baldwin (1999) on capital, as well as the synthesis in Baldwin et al. (2011).

2 A model of spatial persistence and path dependence

In this section we develop a dynamic economic geography framework that is amenable to the empirical study of geographic path dependence. A large set of regions possess arbitrary, time-varying fundamentals in terms of productivity and amenities. They interact via costly trade in goods and costly but forward-looking migration. Crucially, production and locational amenities both potentially involve contemporary and historical spillovers—the forces behind both long persistence and path dependence.

2.1 Setup

There are $i \in \{1, ..., N\} \equiv \mathcal{N}$ locations and time is discrete, infinite, and indexed by $t \in \{0, 1, ...\} \equiv \mathcal{T}$. The world is inhabited by many forward-looking dynastic families, where individuals in a family live for two periods. In the first period ("childhood"), a child is born into each family, living where her parent lives and consuming what her parent consumes. At the beginning of the second period ("adulthood"), this former child (now an adult) realizes her own idiosyncratic locational preferences and chooses where to live, taking into account not only her own benefit of living in a location but also the expected benefit of all future generations of her family. Once she has made her location choice, she supplies a unit of labor inelastically to produce, she consumes, and she gives birth to a child.

[5] Recent work applying our framework includes the study of how trading patterns evolved in Brazil (Pellegrina et al. 2021), where urbanization occurred in colonial Latin America (Ellingsen 2021), and how South Korea industrialized (Choi & Shim 2021).

Let L_{it} denote the number of workers (adults) residing in location i at time t, where the total number of workers $\sum_{i=1}^{N} L_{it} = \bar{L}$, is normalized to a constant in each period t.[6]

2.1.1 Production

Each location i is capable of producing a unique good—the Armington (1969) assumption. A continuum of firms (indexed by ω) in location i produce this homogeneous good under perfectly competitive conditions with the constant returns-to-scale production function $q_{it}(\omega) = A_{it} l_{it}(\omega)$, where labor $l_{it}(\omega)$ is the only production input, and hence $\int l_{it}(\omega) d\omega = L_{it}$. The productivity level for the location is given by

$$A_{it} = \bar{A}_{it} L_{it}^{\alpha_1} L_{it-1}^{\alpha_2}, \tag{1}$$

where $\bar{A}_{it}$ is an exogenous (but unrestricted) component of this location's productivity in year t.[7] Importantly, the two additional components of a location's productivity in equation (1) depend on the number of workers in that location both in the current period, L_{it}, and in the previous period, L_{it-1}. We assume that firms take these aggregate labor quantities as given. Hence the parameter α_1 governs the strength of any potential (positive or negative) contemporaneous agglomeration spillovers working through the size of local production. This is a simple way of capturing Marshallian externalities, external economies of scale, knowledge transfers, thick market effects in output or input markets, and the like, and is standard in many approaches to modeling spatial economies (Redding & Rossi-Hansberg 2017), albeit typically in static models that would combine the effects of L_{it} and L_{it-1}.

The parameter α_2, on the other hand, governs the strength of potential *historical* agglomeration spillovers.[8] This allows for the possibility that two cities with equal fundamentals $\bar{A}_{it}$ and sizes L_{it} today might feature different productivity levels A_{it} today because they had differing sizes L_{it-1} in the past. There are many potential reasons that one might expect $\alpha_2 > 0$, and we describe two such sets of microfoundations briefly here (with complete derivations in Appendix B.1).

Consider first the potential persistence of local knowledge. In particular, we present a model based on Deneckere & Judd (1992), where firms can incur a fixed cost to develop a new variety, for which they earn monopolistic profits for a single period. In the subsequent period, the blueprint for the product becomes common knowledge so that the variety is produced

[6] Our model economy exhibits a form of scale-invariance that means that, for the purposes of our analysis here, the total number of workers in any time period is irrelevant for the distribution of economic activity.

[7] In the initial period $t = 0$, we set $A_{i0} = \bar{A}_{i0} L_{i0}^{\alpha_1}$, as there is no preceding $t = -1$ period.

[8] We consider historical spillovers that take place with a lag of one time period but the tools developed in this paper could be applied to a richer sequence of potential spillovers.

under perfect competition, and we assume the product becomes obsolete two periods after its creation. As in Krugman (1980), the equilibrium number of new varieties will be proportional to the contemporaneous local population. Given consumers' love of variety, new varieties act isomorphically to an increase in the productivity of the single Armington product, resulting in the precise form of equation (1) with $\alpha_1 \equiv \frac{\chi}{\rho-1}$ and $\alpha_2 \equiv \frac{1-\chi}{\rho-1}$, where χ is the expenditure share on all new varieties and $\rho > 1$ is the elasticity of substitution across individual varieties.

Second, consider the potential for durable investments in local productivity. In particular, we present a model based on Desmet & Rossi-Hansberg (2014), in which firms hire workers both to produce and to innovate, and where innovation increases each firm's own productivity contemporaneously and increases all firms' productivity levels in the subsequent period. If firms earn zero profits in equilibrium due to competitive bidding over a fixed factor (e.g. land), then, as in Desmet & Rossi-Hansberg (2014), the dynamic problem of the firm simplifies to a sequence of static profit-maximizing problems. With Cobb-Douglas production functions, equilibrium productivity can be written as in equation (1) with $\alpha_1 \equiv \frac{\gamma_1}{\xi} - (1-\mu)$, and $\alpha_2 \equiv \widetilde{\delta}\frac{\gamma_1}{\xi}$, where γ_1 governs the decreasing returns of innovation in productivity, ξ governs the decreasing returns of labor in innovation, $\widetilde{\delta}$ is the depreciation of investment, and μ is the share of labor in the production function.

Of course, there are surely many sets of microfoundations that could generate the productivity spillover features assumed in equation (1). In what follows, we characterize the properties of the model and estimate the strength of the spillovers without taking a stand on the particular source of these effects.

2.1.2 Consumption

An adult and her child consume with the same preferences, with a constant (but irrelevant) fraction allocated to the child. They have constant elasticity of substitution (CES) preferences, with elasticity $\sigma > 1$, across the differentiated goods that each location can produce. Letting w_{it} denote the equilibrium nominal wage, and letting P_{it} be the price index (solved for below), the deterministic component of welfare in a period t—that is, welfare up to an idiosyncratic shock that we introduce below—of any adult residing in location i at time t is given by

$$W_{it} \equiv u_{it}\frac{w_{it}}{P_{it}}, \tag{2}$$

where the component u_{it} refers to a location-specific amenity value that is given by

$$u_{it} = \bar{u}_{it} L_{it}^{\beta_1} L_{it-1}^{\beta_2}. \tag{3}$$

The term $\bar{u}_{it}$ allows for flexible exogenous amenity offerings in any location and time period.[9] Endogenous amenities work analogously to the production externality terms introduced above, with the parameters β_1 and β_2 here capturing the potential for the presence of other adults in a location to directly affect (either positively or negatively, depending on the sign of β_1 and β_2) the utility of any given resident. We similarly assume that consumers take these terms as given when making decisions.

As is well understood, a natural source of a negative value for β_1 in a model such as this one is the possibility of local congestion forces that are not directly modeled here; for example, if non-tradable goods (such as housing and land) are in fixed supply locally and are demanded with fixed expenditure shares then $-\beta_1$ would equal the share of expenditure spent on such goods. Such effects would work contemporaneously, so they would govern β_1.

Similarly to the case of productivity effects governed by α_2, the parameter β_2 stands in for phenomena through which the historical population L_{it-1} affects the utility of residents in year t directly (that is, other than through productivity, wages, prices, or current population levels). Again it seems potentially important to allow for such effects given the likelihood that previous generations of residents may leave a durable impact, positive or negative, on their former locations of residence. Positive impacts could include the construction of infrastructure (e.g. housing, parks, or sewers), and negative impacts could include environmental damage or resource depletion.

As with productivity, we emphasize that there may be other theoretical rationales for the amenity spillovers assumed in equation (3). In terms of what follows, there is no need to emphasize any one particular microfoundation. But it is again helpful to see an example. To that end, consider (with details in Appendix B.2), a model where agents consume both a tradable good and local housing, and each unit of land is owned by a real estate developer who bids for the rights to develop the land and then chooses the amount of housing to construct. To build housing, the developer combines local labor and the (depreciated) housing stock from the previous period. We assume the bidding process ensures developers earn zero profits, so as in Desmet & Rossi-Hansberg (2014) the dynamic problem of how much housing to construct simplifies into a series of static profit maximizing decisions. In equilibrium, the higher the contemporaneous population, the lower the utility of local residents (as the residents each consume less housing), whereas the higher the population in the previous period, the higher the utility of local residents (as more workers in the previous period results in a greater housing stock today). In particular, if production and utility functions are Cobb-Douglas (with μ the share of old housing in production and $1 - \lambda$ the share of housing in expenditure) this model will be isomorphic to equation (3), with $\beta_1 = -\mu\frac{1-\lambda}{\lambda} < 0$

[9] As with the case of productivity A_{i0}, we set $u_{i0} = \bar{u}_{i0} L_{i0}^{\beta_1}$ in the initial period $t = 0$.

and $\beta_2 = \rho\mu\frac{1-\lambda}{\lambda} > 0$, where ρ is the depreciation rate of the housing stock.

2.1.3 Trade

Bilateral trade from location i to location j incurs an exogenous iceberg trade cost, $\tau_{ijt} \geq 1$ (where $\tau_{ijt} = 1$ corresponds to frictionless trade). Given this, bilateral trade flow expenditures X_{ijt} take on the well-known gravity form given by

$$X_{ijt} = \tau_{ijt}^{1-\sigma} \left(\frac{w_{it}}{A_{it}}\right)^{1-\sigma} P_{jt}^{\sigma-1} w_{jt} L_{jt}, \tag{4}$$

where

$$P_{it} \equiv \left(\sum_{k=1}^{N} \left(\tau_{kit}\frac{w_{kt}}{A_{kt}}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}} \tag{5}$$

is the CES price index referred to above.

For the empirical analysis below, it is convenient to write equation (4) as:

$$X_{ijt} = \tau_{ijt}^{-\theta} \times \frac{\left(Y_{it}/Y^W\right)}{\mathcal{P}_{it}^{1-\sigma}} \times \frac{Y_{jt}}{P_{jt}^{1-\sigma}}, \tag{6}$$

where

$$\mathcal{P}_{it} \equiv \left(\frac{w_{it}}{A_{it}}\right)^{-1} \left(\frac{Y_{it}}{Y^W}\right)^{\frac{1}{1-\sigma}}, \tag{7}$$

and $Y_{it} \equiv w_{it}L_{it}$, and Y^W is total world income (which we normalize to one in what follows). In the terminology of the gravity trade literature (see e.g. Anderson & Van Wincoop 2003), (the inverse of) $\mathcal{P}_{it}$ captures the *outward trade market access* of location i and (the inverse of) P_{jt} captures the *inward trade market access* of location j.

2.1.4 Migration

We now turn to the decision of agents regarding how to migrate between different locations. This has three ingredients. First, similarly to the case of costly trade introduced above, we assume that individuals migrating from i to j in period t incur additive migration costs $\tilde{\mu}_{ijt} \geq 0$; second, we also allow for idiosyncratic unobserved heterogeneity in how each child will value living in each location j in adulthood, by assuming that each child has idiosyncratic extreme value (Gumbel) distributed preferences over potential destinations with shape parameter $\theta \geq 0$ (and location parameters normalized to one without loss); third, we assume that individuals discount the welfare of future generations of family members with a discount rate $\delta \geq 0$.

To study such a setting, we first characterize the value an adult derives from residing in location i at time t, which we refer to as $\tilde{V}_{it}$. This value $\tilde{V}_{it}$ has two parts: first, as described above, this adult enjoys her own consumption and amenity value of residing in the location; second, she also values the expected welfare of all future generations of her family, taking into account the fact that those descendants will be making their own optimal migration decisions, yielding the following equation for any adult located in i at time t:

$$\tilde{V}_{it} = \log W_{it} + \delta \mathbb{E}_t \left[\max_j \left\{ \tilde{V}_{j,t+1} - \tilde{\mu}_{ij,t+1} + \varepsilon_{ijt+1} \right\} \right].$$

Here, the expectation is taken over the possible realizations of this adult's child's idiosyncratic preferences $\{\varepsilon_{ijt+1}\}$ that are unknown at the time the adult makes her own migration decision (but which will be realized by the child when she makes her own migration decision next period) and $\delta \geq 0$ captures the discount that this parent applies to her child's welfare relative to her own. Given the assumed extreme value distribution, this expectation has a convenient analytical expression, allowing us to write:

$$V_{it} = W_{it} \Pi_{i,t+1}^{\delta}, \tag{8}$$

where we refer to $V_{it} \equiv \exp\left(\tilde{V}_{it}\right)$ as the *present discounted value* (PDV) of a family whose living adult member at time t resides in location i, $\mu_{ik,t+1} \equiv \exp\left(\tilde{\mu}_{ik,t+1}\right) \geq 1$ is the migration cost and $\Pi_{it} \equiv \left(\sum_{k=1}^{N} \left(V_{kt}/\mu_{ikt}\right)^{\theta}\right)^{\frac{1}{\theta}}$ summarizes the appeal of migration options for those who are born in period $t-1$ in location i. Equation (8) characterizes the present discounted value V_{it} an adult receives from residing in a location i in time t, accounting for both her own period payoff and her dynastic considerations of future generations of her family.

We now consider the migration decision of a child. Recall from the discussion of timing above that L_{it-1} adults reside in location i at time $t-1$, and they have one child each. Those children choose at the beginning of period t—as they pass into adulthood—where they want to live as adults, accounting for the (deterministic) value they receive from that location, the migration costs they incur, and their own idiosyncratic preferences. Letting the vector of such idiosyncratic taste differences (one for each location) be denoted by $\vec{\varepsilon}$, the actual period payoff of a child who receives the draw $\vec{\varepsilon}$ while living in location i at time $t-1$ and who chooses to move to location j as an adult is:

$$\tilde{V}_{ijt}\left(\vec{\varepsilon}\right) \equiv \tilde{V}_{jt} - \tilde{\mu}_{ij,t} + \varepsilon_{ijt}, \tag{9}$$

Hence, any new adult chooses her location as follows:

$$\max_j \tilde{V}_{ijt}\left(\vec{\varepsilon}\right) = \max_j \left(\tilde{V}_{jt} - \tilde{\mu}_{ij,t} + \varepsilon_{ijt}\right).$$

Given the assumed extreme value distribution, the number of children in location i at time $t-1$ who choose to move to location j at time t, L_{ijt}, is then given by:

$$L_{ijt} = \mu_{ijt}^{-\theta} \Pi_{it}^{-\theta} L_{it-1} V_{jt}^{\theta}. \tag{10}$$

Equation (10) says that there will be greater migration toward destination locations j with higher dynamic value V_{jt} and low bilateral migration costs μ_{ijt}, and coming from origin locations i that either have a lot of residents L_{it-1} or poor outside options Π_{it}.

Finally, for the empirical analysis below, it is convenient to write equation (10) as:

$$L_{ijt} = \mu_{ijt}^{-\theta} \times \frac{L_{it-1}}{\Pi_{it}^{\theta}} \times \frac{L_{jt}/\bar{L}}{\Lambda_{jt}^{-\theta}}, \tag{11}$$

where

$$\Lambda_{it} \equiv V_{it} \left(\frac{L_{it}}{\bar{L}}\right)^{-\frac{1}{\theta}}. \tag{12}$$

As with the flow of goods described above, Π_{it} captures the *outward migration market access* from i and the (inverse of) Λ_{jt} captures the *inward migration market access* to j.

2.2 Dynamic equilibrium

An equilibrium in this dynamic economy is a sequence of values of (finite) prices and (strictly positive) allocations such that goods and factor markets clear in all periods.[10] More formally, for any strictly positive initial population vector $\{L_{i0}\}$ and geography vector $\left\{\bar{A}_{it}, \bar{u}_{it}, \tau_{ijt}, \mu_{ijt}\right\}$, an equilibrium is a vector of endogenous variables $\{L_{it}, w_{it}, W_{it}, \Pi_{it}, V_{it}\}$ such that, for all locations i and time periods t, we have:

1. *Total sales are equal to payments to labor:* That is, a location's income is equal to the value of all locations' purchases from it, or $w_{it}L_{it} = \sum_j X_{ijt}$. Using equation (4) this

[10]Throughout, we confine attention to equilibria where all locations are inhabited, as *(i)* these are the empirically relevant types of equilibria at our geographic scale of analysis; and *(ii)* in the presence of productivity and/or amenity spillovers, from equations (1) and (3), an uninhabited location will (trivially) remain uninhabited forever.

can be written as

$$w_{it}^{\sigma} L_{it}^{1-\alpha_1(\sigma-1)} = \sum_j K_{ijt} L_{jt}^{\beta_1(\sigma-1)} W_{jt}^{1-\sigma} w_{jt}^{\sigma} L_{jt}, \tag{13}$$

with $K_{ijt} \equiv \left(\frac{\tau_{ijt}}{\bar{A}_{it} L_{it-1}^{\alpha_2} \bar{u}_{jt} L_{jt-1}^{\beta_2}} \right)^{1-\sigma}$ defined as a collection of terms that are either exogenous, or predetermined from the perspective of period t.

2. *Trade is balanced:* That is, a location's income is fully spent on goods from all locations, or $w_{it} L_{it} = \sum_j X_{jit}$. Using equation (4) this can be written as

$$w_{it}^{1-\sigma} L_{it}^{\beta_1(1-\sigma)} W_{it}^{\sigma-1} = \sum_j K_{jit} L_{jt}^{\alpha_1(\sigma-1)} w_{jt}^{1-\sigma}. \tag{14}$$

3. *A location's population is equal to the population arriving in that location:* That is, $L_{it} = \sum_j L_{jit}$. From equation (10) this implies

$$L_{it} V_{it}^{-\theta} = \sum_j \mu_{jit}^{-\theta} \Pi_{jt}^{-\theta} L_{jt-1}. \tag{15}$$

4. *A location's population in the previous period is equal to the number of people exiting that location:* That is, $L_{it-1} = \sum_j L_{ijt}$. From equation (10) this can be written as

$$L_{it-1} = \sum_j \mu_{ijt}^{-\theta} \Pi_{it}^{-\theta} L_{it-1} V_{jt}^{\theta},$$

which can then be written more compactly as

$$\Pi_{it}^{\theta} \equiv \sum_j \mu_{ijt}^{-\theta} V_{jt}^{\theta}. \tag{16}$$

5. *Agents are forward-looking*: That is, the payoffs of residing in a location depend both on the period payoffs and the present discounted value of future generations. From equation (8) this can be written as:

$$V_{it} = W_{it} \Pi_{i,t+1}^{\delta}. \tag{17}$$

Summarizing, the dynamic equilibrium can be represented as the system of $5 \times N$ equations (in equations 13-17) in $5 \times N$ unknowns, $\{L_{it}, w_{it}, W_{it}, \Pi_{it}, V_{it}\}$ for all countably infinite time periods $t \in \{1, ..., \}$.

This system of equations (13)-(17) is a special case of the following more general dynamic system of nonlinear equations:

$$x_{i,h,t} = \sum_{j=1}^{N} K_{ij,h,t} \prod_{h'=1}^{H} \left(x_{j,h',t}\right)^{\varepsilon^{\mathtt{j,t}}_{h,h'}} \left(x_{j,h',t+1}\right)^{\varepsilon^{\mathtt{j,t+1}}_{h,h'}} \left(x_{i,h',t+1}\right)^{\varepsilon^{\mathtt{i,t+1}}_{h,h'}} \left(x_{j,h',t-1}\right)^{\varepsilon^{\mathtt{j,t-1}}_{h,h'}} \left(x_{i,h',t-1}\right)^{\varepsilon^{\mathtt{i,t-1}}_{h,h'}}, \tag{18}$$

In this system, the values of $\{x_{i,h,t}\}_{i,h,t} \in \mathbb{R}^{N\times H}_{++} \times ....$ are unknown, whereas those of $\{K_{ij,h,t}\}_{ij,h,t} \in \mathcal{K} \subseteq \mathbb{R}^{N^2\times H}_{+} \times ....$ and $\left\{\varepsilon^{\mathtt{j,t}}_{h,h'}, \varepsilon^{\mathtt{j,t+1}}_{h,h'}, \varepsilon^{\mathtt{i,t+1}}_{h,h'}, \varepsilon^{\mathtt{j,t-1}}_{h,h'}, \varepsilon^{\mathtt{i,t-1}}_{h,h'}\right\}_{h,h'} \in \mathbb{R}^{5(H\times H)}$ are finite and given. Similarly, for some subset $\tilde{\mathcal{H}} \subseteq \mathcal{H}$, the initial conditions $\{x_{i,h,t}\}_{i\in\mathcal{N},h\in\tilde{\mathcal{H}},t=0} \in \mathbb{R}^{N\times\tilde{H}}_{++}$ are finite and given, where $\tilde{H}$ is the dimension of $\tilde{\mathcal{H}}$. One can interpret equation (18) as describing a dynamic model with N "locations" (indexed by i and j) and H different "types" of co-determined endogenous outcomes (indexed by h). Economic interactions (potentially) take place within the same location but across adjacent time periods (in which case the $H \times H$ matrix of cross-type elasticities is given by $\mathbf{E}^{\mathtt{i,t-1}} \equiv \left\{\varepsilon^{\mathtt{i,t-1}}_{h,h'}\right\}_{h,h'}$ for interactions between t and $t-1$ and by $\mathbf{E}^{\mathtt{i,t+1}}$ for those between t and $t+1$), within the same time period but across different locations (denoted by the elasticities $\mathbf{E}^{\mathtt{j,t}}$), and across both different locations and adjacent time periods (with elasticities denoted by $\mathbf{E}^{\mathtt{j,t-1}}$ and $\mathbf{E}^{\mathtt{j,t+1}}$). In each case, the elasticities in the $\mathbf{E}$ matrices govern the strength of dynamic interactions but are themselves time-invariant.[11]

The analysis of such a system can prove challenging given both the large state space and the forward-looking dynamic behavior. Indeed, we believe the following result is the first to offer a characterization of equilibrium properties of such a forward-looking non-linear general equilibrium quantitative spatial model.

We focus on *bounded equilibria*, for which the solution $x_{i,h,t}$ to equation (18) has the property that there exists a set of finite strictly positive scalars $\{m_{h,t}, M_{h,t}\}_{h,t}$ such that for all $h \in \mathcal{H}$ and $t \in \mathcal{T}$ we have $0 < m_{h,t} \leq x_{i,h,t} \leq M_{h,t} < \infty$ for all $i \in \mathcal{N}$. The following Theorem establishes sufficient conditions for the uniqueness of a bounded equilibrium that satisfies equation (18):

Theorem 1. *Consider the inhomogeneous linear second-order difference equation,*

$$\left(\left|\mathbf{E}^{\mathtt{i,t-1}}\right| + \left|\mathbf{E}^{\mathtt{j,t-1}}\right|\right)\boldsymbol{\mu}_{t-1} - \left(\mathbf{I} - \left|\mathbf{E}^{\mathtt{j,t}}\right|\right)\boldsymbol{\mu}_{t} + \left(\left|\mathbf{E}^{\mathtt{j,t+1}}\right| + \left|\mathbf{E}^{\mathtt{i,t+1}}\right|\right)\boldsymbol{\mu}_{t+1} = \boldsymbol{b}_t, \tag{19}$$

[11] The notation used here for the $5H^2$ elasticities $\left\{\varepsilon^{\mathtt{j,t}}_{h,h'}, \varepsilon^{\mathtt{j,t+1}}_{h,h'}, \varepsilon^{\mathtt{i,t+1}}_{h,h'}, \varepsilon^{\mathtt{j,t-1}}_{h,h'}, \varepsilon^{\mathtt{i,t-1}}_{h,h'}\right\}_{h,h'}$ is such that the superscripts denote the nature of the interaction across locations (where "$\mathtt{i}$" denotes "within-location", and "$\mathtt{j}$" denotes "cross-location") and time (where "$\mathtt{t-1}$" denotes "with the previous period", "$\mathtt{t}$" denotes "within the same period", and "$\mathtt{t+1}$" denotes "with the next period"), and the subscripts denote the elasticity of variable type h with respect to type h'.

where: the absolute value operator $|\cdot|$ is taken element-wise; $\mathbf{I}$ *denotes the $H \times H$ identity matrix; the $H \times H$ matrices* $\mathbf{E}^{\mathtt{i,t-1}}$, $\mathbf{E}^{\mathtt{j,t-1}}$, $\mathbf{E}^{\mathtt{j,t}}$, $\mathbf{E}^{\mathtt{j,t+1}}$, *and* $\mathbf{E}^{\mathtt{i,t+1}}$ *are given and correspond to the values defined in equation (18); the sequence* $\boldsymbol{b}_t$ *is given for all $t \in \mathcal{T}$; the initial conditions $\mu_{\tilde{h},0} = 0$ for all $\tilde{h} \in \tilde{\mathcal{H}}$; and* $\boldsymbol{\mu}_t$ *is unknown for all $t > 0$ and for all $\tilde{h} \notin \tilde{\mathcal{H}}$ at $t = 0$. Then there is at most one bounded equilibrium solution to equation (18) if the following two conditions hold:*

(a) In the case where $\boldsymbol{b}_t = \mathbf{0}$ *for all $t \in \mathcal{T}$, the unique solution to (19) is* $\boldsymbol{\mu}_t = \mathbf{0}$ *for all $t \in \mathcal{T}$.*

(b) In the case where $\boldsymbol{b}_t \geq \mathbf{0}$ *(and where at least one element of the inequality is strict) for all $t \in \mathcal{T}$, there exists no solution to (19) of the form* $\boldsymbol{\mu}_t \geq \mathbf{0}$ *for all $t \in \mathcal{T}$.*

Proof. See Online Appendix A.1. □

This result provides a sufficient condition (concerning both the values of elasticities $\mathbf{E}$ and the set of variable types $\tilde{h} \in \tilde{\mathcal{H}}$ whose initial conditions are given) that ensures the uniqueness of a bounded equilibrium in any model that can be written in the form of equation (18). Importantly, this condition is sufficient irrespective of the values taken by the sequence of exogenous fundamentals $\{K_{ij,h,t}\}_{ij,h,t}$ and by the initial conditions $\{x_{i,h,t}\}_{h \in \tilde{\mathcal{H}}, i \in \mathcal{N}, t=0}$. Appendix Remark 1 describes a procedure that allows one to check whether conditions (a) and (b) of Theorem 1 are satisfied in any given application.[12]

The proof of this result is based on following insight. Letting $\boldsymbol{\mu}_t$ denote the log ratio of the maximum and minimum (taken across all locations $i \in \mathcal{N}$ for a given $h \in \mathcal{H}$ and $t \in \mathcal{T}$) between any two candidate solutions to the $N \times H \times \mathcal{T}$ *nonlinear* dynamic system (18), we first show that $\boldsymbol{\mu}_t$ is bounded above by the solution to the $H \times \mathcal{T}$ *linear* dynamic system in equation (19) (and from below by zero, by construction). Because such a linear system has been previously characterized (see, e.g., Theorem 8.3 of Gohberg et al. 2005), we can provide conditions under which its only (weakly) positive solution is $\boldsymbol{\mu}_t = \mathbf{0}$. Under such conditions the upper and lower bounds coincide, and the two candidate solutions must be equal.

As mentioned above, our model's dynamic system (13)-(17) is an example of one that can be cast in the form in equation (18), allowing the application of Theorem 1. Before doing so, we simplify the system from one with $H = 5$ types of endogenous variables to one with $H = 3$. A first simplification follows from using equation (17) to substitute for W_{it}. A second follows if we assume that trade costs τ_{ijt} are symmetric, as they will be in our empirical application below, in which case (13) and (14) can be combined into a single

[12] The procedure is based on the properties of the corresponding quadratic eigenvalue problem; see Gohberg et al. (2005) and Tisseur & Meerbergen (2001). Sims (2002) applies closely related techniques to characterize the properties of linear rational expectations models. The distinctive feature of Theorem 1 is its use of such techniques to provide sufficient conditions for uniqueness in non-linear forward-looking economies.

equation.[13] Together, this reduces the equilibrium to a set of $3 \times N$ equations for $3 \times N$ unknowns $\{L_{it}, V_{it}, \Pi_{it}\}$ in each time period $t \in \{1, ...\}$. Then a straightforward change of variables such that $[\ln x_{i,1,t}, \ln x_{i,2,t}, \ln x_{i,3,t}]^T = \mathbf{\Gamma} [\ln L_{it}, \ln W_{it}, \ln \Pi_{it}]^T$, where $\mathbf{\Gamma}$ is a 3-by-3 matrix depending on $\{\alpha_1, \beta_1, \sigma, \theta\}$,[14] allows us to apply Theorem 1 to yield the following Corollary:

Corollary 1. *Suppose that the matrices of elasticities* $\mathbf{E}^{\mathtt{j,t}}$, $\mathbf{E}^{\mathtt{j,t+1}}$, $\mathbf{E}^{\mathtt{i,t+1}}$, $\mathbf{E}^{\mathtt{j,t-1}}$, *and* $\mathbf{E}^{\mathtt{i,t-1}}$ *described in Theorem 1 are as follows:*

$$\mathbf{E}^{\mathtt{j,t}} = \begin{pmatrix} \tilde{\sigma}\left(1+\alpha_1\sigma+\beta_1\left(\sigma-1\right)\right) & \left(1-\sigma\right)\tilde{\sigma} & 0 \\ 0 & \theta & 0 \\ 0 & 0 & -\theta \end{pmatrix} \mathbf{\Gamma}^{-1},$$

$$\mathbf{E}^{\mathtt{j,t+1}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \delta\theta \\ 0 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1}, \ \mathbf{E}^{\mathtt{i,t+1}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \delta\theta \end{pmatrix} \mathbf{\Gamma}^{-1},$$

$$\mathbf{E}^{\mathtt{j,t-1}} = \begin{pmatrix} \tilde{\sigma}\left(\alpha_2\sigma+\beta_2\left(\sigma-1\right)\right) & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1}, \ \mathbf{E}^{\mathtt{i,t-1}} = \begin{pmatrix} \tilde{\sigma}\left(\alpha_2\left(\sigma-1\right)+\beta_2\sigma\right) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1},$$

where $\tilde{\sigma} \equiv (\sigma-1)/(2\sigma-1)$. *If these matrices satisfy conditions (a) and (b) of Theorem 1, then for any initial population* $\{L_{i0}\}$ *and geography* $\{\bar{A}_{it} > 0, \bar{u}_{it} > 0, \tau_{ijt} = \tau_{jit}, \mu_{ijt} > 0\}$, *there exists at most one bounded equilibrium in the model described by equations (13)-(17).*

Proof. See Online Appendix A.2. □

As described above, conditions (a) and (b) of Theorem 1 are straightforward to verify in any application. Panel (a) of Figure 1 illustrates the results of doing so for the case of our model (and hence as an illustration of Corollary 1) across a range of values for the contemporaneous spillover elasticities α_1 and β_1, while setting the values of σ, θ, δ, α_2, and β_2 to those that we use in our empirical calculations below. Despite the added complexity of the forward-looking behavior, it is reassuring to note that the standard economic intuition continues to hold: the sufficient condition for uniqueness will be satisfied whenever α_1 and β_1 are sufficiently small. Indeed, at the values used here it turns out that uniqueness is assured

[13] This follows from the fact that when trade costs are symmetric outward and inward goods market access $\mathcal{P}_{it}$ and P_{it} are equal up to scale; see Anderson & Van Wincoop (2003) and Allen & Arkolakis (2014).

[14] In particular, $\mathbf{\Gamma} \equiv \begin{pmatrix} \left(\frac{\sigma-1}{2\sigma-1}\left(1-\alpha_1\left(\sigma-1\right)-\beta_1\sigma\right)\right) & \frac{\sigma-1}{2\sigma-1}\sigma & 0 \\ 0 & 0 & \theta \\ 1 & -\theta & 0 \end{pmatrix}$.

by the simple condition that $\alpha_1+\beta_1 < \frac{1}{\theta}$—that is, the sum of contemporaneous agglomeration forces must simply be greater than dispersion forces. However, the sufficient condition for uniqueness can fail (and hence the possibility of expectations-based multiplicity may arise) if the strength of forward-looking behavior is particularly strong. Similarly, while in principle it is possible for the strength of historical spillovers α_2 and β_2 to affect the sufficient conditions for uniqueness, such a scenario arises only under stronger forward-looking behavior. Finally, we note that in the special case where there is no forward-looking behavior (i.e. $\delta = 0$), the sufficient condition in Corollary 1 is satisfied whenever the spectral radius of $|\mathbf{E}^{\mathtt{j,t}}|$ is less than one, which corresponds to the condition that is familiar from static spatial equilibrium models (see Allen et al. 2021). We denote this condition by $\rho\left(|\mathbf{E}^{\mathtt{j,t}}|\right) < 1$, where $\rho(\cdot)$ denotes the spectral radius operator, in what follows.

To provide some intuition for the dynamic system, algebraic manipulations of equations (13)-(17) when trade costs are symmetric imply that the equilibrium distribution of population in any location and time period can be written as

$$\begin{aligned}\gamma \ln L_{it} = C_t + \sigma \ln \bar{u}_{it} + (\sigma - 1) \ln \bar{A}_{it} - (2\sigma - 1) \ln P_{it} - \sigma \ln \Lambda_{it} \\ + \sigma\delta \ln \Pi_{i,t+1} + \left(\alpha_2 (\sigma - 1) + \beta_2 \sigma\right) \ln L_{i,t-1},\end{aligned} \tag{20}$$

where $\gamma \equiv 1 + \frac{\sigma}{\theta} - \left(\alpha_1 (\sigma - 1) + \beta_1 \sigma\right)$ and C_t is a constant that ensures that $\sum_{i=1}^{N} L_{it} = \overline{L}$. Equation (20) has four implications. First, as long as $\gamma > 0$ (which corresponds to the case of our empirical estimates below), a greater density of residents can be found in any location with high productivity $\bar{A}_{it}$, high amenities $\bar{u}_{it}$, high inward migration access (low Λ_{it}), high access to imported goods (low P_{it}), higher present discounted value for future generations (high $\Pi_{i,t+1}$), and—if $\alpha_2 (\sigma - 1) + \beta_2 \sigma > 0$, so that historical spillovers are positive—with greater population density in the previous period. Second, the elasticities of the population to these characteristics are governed by the strength of γ^{-1}, where greater contemporaneous spillover elasticities α_1 and β_1 result in larger population responses. Third, history—i.e. the distribution of the population in the previous period—only affects the current population through the inward market access terms (Λ_{it} and P_{it}) and through the direct impact on productivities and amenities from the historical spillover elasticities α_2 and β_2. Fourth, the future—i.e. future productivities, amenities, and distributions of population—only affects the contemporaneous population through the next period's outward migration access ($\Pi_{i,t+1}$). While the first two determinants of population density in equation (20), $\bar{A}_{it}$ and $\bar{u}_{it}$, are exogenous in our model, the latter four determinants, Λ_{it}, P_{it}, $\Pi_{i,t+1}$,and $L_{i,t-1}$ are endogenous and are determined simultaneously through interactions with the endogenous features in all other locations. It is the self-reinforcing potential of these interactions, both over time and across space, that leads to the potentially rich dynamics that we explore below.

2.3 Persistence and path dependence

We now turn to a characterization of the dynamic properties of the model, namely the persistence of shocks to the economy and the possibility of multiple steady-states (i.e. the potential for path dependence).

Persistence

Consider first the question of persistence: how long does a temporary shock to the economy take to dissipate? Here (and only here) we consider the special case of our model where $\delta = 0$, i.e. where agents do not care about the welfare of future generations. Even in the absence of forward-looking behavior, it turns out that there is the possibility of extreme persistence in the economy. We begin by defining $\chi_{x,t} \equiv \frac{\max_i x_{i,t}/x_{i,t-1}}{\min_i x_{i,t}/x_{i,t-1}}$ to be the ratio of the maximum to minimum change (from $t-1$ to t) in any variable $x_{i,t}$ across all locations. Note that $\chi_{x,t} \geq 1$ and is equal to one if and only if $x_{i,t} \propto x_{i,t-1}$ for all i, i.e. the economy is on a balanced growth path (or, in our case where aggregate population is fixed, a steady-state). As such, it provides a convenient economy-wide measure of how far $x_{i,t}$ is from a steady-state. We can then define the economy-wide persistence of variable $x_{i,t}$ as the effect of $\chi_{x,t-1}$ on $\chi_{x,t}$—that is, how much deviations from the steady-state in period $t-1$ affect deviations from the steady-state in period t. The following proposition bounds the persistence of all endogenous outcomes in the model in this manner:

Proposition 1. *Consider any initial population* $\{L_{i0}\}$ *and time-invariant geography* $\{\bar{A}_i > 0, \bar{u}_i > 0, \tau_{ij} = \tau_{ji}, \mu_{ij} > 0\}$. *Suppose that* $\delta = 0$ *and* $\rho\left(\left|\mathbf{E}^{\mathsf{j,t}}\right|\right) < 1$, *where* $\mathbf{E}^{\mathsf{j,t}}$ *is defined in Corollary 1, so that the dynamic equilibrium is unique. Then the following relationship holds:*

$$\begin{pmatrix} \ln \chi_{L,t} \\ \ln \chi_{V,t} \\ \ln \chi_{\Pi,t} \end{pmatrix} \leq \left|\mathbf{\Gamma}^{-1}\right| \left(\mathbf{I} - \mathbf{E}^{\mathsf{j,t}}\right)^{-1} \mathbf{G} \left|\mathbf{\Gamma}\right| \begin{pmatrix} \ln \chi_{L,t-1} \\ \ln \chi_{V,t-1} \\ \ln \chi_{\Pi,t-1} \end{pmatrix}, \tag{21}$$

where $\mathbf{G}$ *is a 3-by-3 matrix whose first two rows are strictly positive (with values that depend on the parameters* $\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma$ *and* θ*, as fully defined in Section A.3) and whose second row consists entirely of zeroes.*

Proof. See Section A.3. □

Proposition 1 provides an upper bound on how much the endogenous variables L_{it}, V_{it} and Π_{it} change from period $t-1$ to period t that depends on how much they changed from period

$t-2$ to period $t-1$, holding constant the underlying geography. The proposition states that the closer the spillover parameters are to the boundary at which uniqueness can no longer be guaranteed, the less stringent the upper bound places on the rate of persistence. To see this, note that as the spectral radius $\rho\left(\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)$ approaches one from below, the largest eigenvalue of $\left(\mathbf{I}-\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)^{-1}$—and hence also the largest eigenvalue of $\left|\mathbf{\Gamma}^{-1}\right|\left(\mathbf{I}-\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)^{-1}\mathbf{G}\left|\mathbf{\Gamma}\right|$, given the properties of $\mathbf{G}$ stated in the Proposition—approaches infinity.[15]

Path dependence

So far we have described the dynamic transition paths of this spatial economy. We now discuss the steady-state(s) to which these paths may converge. Intuitively, if local agglomeration economies are strong enough then there could be multiple allocations at which the economy would be in steady-state. Agents who come to reside in a location could find it optimal, on average, to stay there; and yet the same could simultaneously be true for another location, thanks to the reinforcing logic of local positive spillovers.

To evaluate this possibility we consider a version of the above economy but for which the potentially time-varying fundamentals $\left\{\bar{A}_{it}\right\}$ and $\left\{\bar{u}_{it}\right\}$, as well as the trade $\left\{\tau_{ijt}\right\}$ and migration $\left\{\mu_{ijt}\right\}$ costs, are constant over time at the values $\left\{\bar{A}_i, \bar{u}_i, \tau_{ij}, \mu_{ij}\right\}$. The steady-states of our economy will therefore be a set of time-invariant endogenous variables that we denote by $\left\{L_i, w_i, W_i, \Pi_i, V_i\right\}$.[16] The following result provides a sufficient condition for existence and uniqueness of the steady-state of this economy (for arbitrary geographies with symmetric trade and migration costs).[17] It also shows how this is a maximal domain sufficient condition—the weakest condition one could impose whose result would be true for any geographic fundamentals.

Proposition 2. *For any time-invariant geography* $\left\{\bar{A}_i > 0, \bar{u}_i > 0, \tau_{ij} = \tau_{ji}, \mu_{ij} = \mu_{ji}\right\}$, *there exists a unique steady-state equilibrium if:*

$$\rho\left(\mathbf{B}\right) < 1,$$

[15] This can be seen from a simple eigen-decomposition $\left(\mathbf{I}-\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)^{-1} = \mathbf{V}'\Lambda\mathbf{V}$ where Λ is a diagonal matrix whose elements are the eigenvalues (including $\frac{1}{1-\rho\left(\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)}$, which approaches infinity as $\rho\left(\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)$ approaches one from below) and $\mathbf{V}$ is a 3×3 matrix of the associated eigenvectors. Note that because $\mathbf{E}^{\mathtt{j,t}}$ is strictly positive and hence $\rho\left(\mathbf{E}^{\mathtt{j,t}}\right) > 0$, the largest eigenvalue of $\left(\mathbf{I}-\left|\mathbf{E}^{\mathtt{j,t}}\right|\right)^{-1}$ always exceeds unity, which indicates that long-lived persistence can never be ruled out.

[16] Note that while population levels at each location L_i are constant in steady-state, and hence net migration flows are zero, gross migration flows are still positive in a steady-state equilibrium due to the churn induced by the idiosyncratic locational preferences in equation (9).

[17] As with the case of trade costs, imposing the symmetry of migration costs both matches our empirical application and reduces the dimensionality of the system of non-linear equations governing the steady state distribution of economic activity, permitting a tighter characterization of its properties.

where

$$\mathbf{B} \equiv \begin{pmatrix} \left| \frac{1-\frac{\sigma}{\theta}-\beta_{ss}+\alpha_{ss}\sigma+\beta_{ss}\sigma+\frac{1}{\theta}}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma} \right| & \left| \frac{(1+\delta)(\alpha_{ss}+1)\left(\frac{\sigma-1}{\theta}\right)}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma} \right| \\ \left| \frac{(2\sigma-1)/(\sigma-1)}{\left(\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma\right)} \right| & \left| \frac{1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma-\delta\frac{\sigma}{\theta}}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma} \right| \end{pmatrix}$$

and $\alpha_{ss} \equiv \alpha_1 + \alpha_2$ *and* $\beta_{ss} \equiv \beta_1 + \beta_2$.

Moreover, if $\rho(\mathbf{B}) > 1$*, then there exist many geographies for which there are multiple steady-states at each geography.*[18]

Proof. See Section A.4. □

In the absence of forward-looking behavior (i.e. at $\delta = 0$), the condition for uniqueness of the steady-state in Proposition 2 is identical to that for uniqueness of transition paths in Corollary 1, with one modification: the condition on steady-states depends on the size of total (that is, contemporaneous plus historical) spillovers $\alpha_{ss} \equiv \alpha_1 + \alpha_2$ and $\beta_{ss} \equiv \beta_1 + \beta_2$ rather than just the contemporaneous spillovers α_1 and β_1. With forward-looking behavior (and at the values for δ, σ, and θ that we use below), as illustrated in Panel (b) of Figure 1, the region of $(\alpha_{ss}, \beta_{ss})$ defined by $\rho(\mathbf{B}) < 1$ in Proposition 2 is very similar to that of (α_1, β_1) defined by Corollary 1. As a result, the basic intuition remains the same: as long as the total strength of the combined productivity and amenity (contemporaneous plus historical) spillovers is sufficiently small, there will be a unique steady-state. The second part of Proposition 2 demonstrates that the sufficient condition for uniqueness of steady-states is indeed necessary for certain geographies. Indeed, the proof of this proposition provides (for any given value of $\mathbf{B}$ such that $\rho(\mathbf{B}) > 1$) a continuum of example geographies under which multiple steady-states arise.

Associated with each steady-state is a *basin of attraction*: a set of values of the initial population distribution $\{L_{i0}\}$ for which the economy will converge to the steady-state in question. When there are multiple steady-states, and hence multiple basins of attraction, the eventual steady-state equilibrium of the economy will generically depend on its initial population distribution. Such a situation offers the potential for *path dependence*: where historical events that determine $\{L_{i0}\}$ can have permanent effects on the economy's outcomes since they select the basin of attraction in which populations are distributed at time 0, and hence the eventual steady-state that is reached. Since the dynamic equilibria described in equations (13)-(17) feature a historical dependence on the state variable $\{L_{it}\}$ with only one lag, this means that from the perspective of any date t the "history" of the system (all exogenous and endogenous outcomes in the past) is fully characterized by $\{L_{i,t-1}\}$. Hence, observing the phenomenon that some event had a path-dependent impact hinges on whether

[18] If $\rho(\mathbf{B}) = 1$, there exists at most one steady-state.

the event moved $\{L_{i,t-1}\}$ across the boundary from one basin of attraction to another. We explore this feature in our counterfactual simulations in Section 4.

Combining Corollary 1 and Proposition 2, we see that the historical spillover parameters α_2 and β_2 play an important role in the study of path-dependent economies. Corollary 1 shows that when the contemporaneous spillover parameters α_1 and β_1 are low then the (bounded) dynamic equilibrium will be unique (as long as agents are not too forward-looking). However, Proposition 2 states that when $\alpha_1 + \alpha_2$ and $\beta_1 + \beta_2$ are high then steady-states are likely to be multiple. In this range of parameters (that is, with relatively low α_1 and β_1, relatively high α_2 and β_2, and δ not too large) path dependence can occur and yet be straightforward to study since the complications (for estimation, computation, and interpretation of counterfactuals) of equilibrium indeterminacy do not arise.

2.4 A three-location example

To see the implications of Propositions 1 and 2 more concretely, consider a simple economy with three locations. Suppose that these locations have identical and time-invariant fundamentals $\left\{\bar{A}_{it}, \bar{u}_{it}, \tau_{ijt}, \mu_{ijt}\right\}$ and symmetric trade and migration costs across locations; further, we use values for σ and θ that are in the empirically relevant range, and set aside amenity spillovers (i.e. $\beta_1 = \beta_2 = 0$) and forward-looking considerations (i.e. $\delta = 0$).[19] We now consider three alternative versions of this example economy under alternative values of contemporaneous and historical productivity spillovers, α_1 and α_2, though always within the range for which the dynamic equilibrium is known (from Corollary 1) to be unique.

Figure 2 illustrates phase diagrams on the two-dimensional space of L_{it} shares in each of these three examples. Blue rays indicate one period of movement (so a ray's length shows speed of adjustment) in the direction towards each red dot and yellow stars denote steady-states. Panel (a) begins with the case where $\alpha_1 = -0.2$ and $\alpha_2 = 0$. Because the contemporaneous productivity spillover exhibits a strong congestion force ($\alpha_1 < 0$), we are far from the barrier of non-uniqueness, so Proposition 1 suggests the economy will exhibit low persistence. Accordingly, the blue arrows are long, showing that the economy converges quickly to the (unique) steady state where all identical locations are equally populated.

In panel (b) we keep $\alpha_2 = 0$ but increase contemporaneous spillovers to $\alpha_1 = 0$. This increase means that the economy has moved closer to the barrier of non-uniqueness, so from Proposition 1, the economy may exhibit stronger persistence. Indeed, this is exactly what occurs, as the shorter blue arrows indicate that the move from any initial population distribution toward the (unique) steady state occurs more sluggishly.

[19] In particular, throughout all of the examples in Figure 2 we set: $\bar{A}_i = \bar{u}_i = 1$; $\mu_{ij} = 1.75$, $\tau_{ij} = 1.6$ for all $i \neq j$ and $\mu_{ii} = \tau_{ii} = 1$; and $\sigma = 9$ and $\theta = 4$.

Finally, in panel (c) we consider the case where $\alpha_1 = 0$ and $\alpha_2 = 0.2$, such that there are now strong historical productivity agglomeration forces. Proposition 2 now implies that the economy may feature multiple steady states and this is exactly what occurs. Indeed, there are three stable steady states (and four other unstable steady states), each characterized by substantial (but incomplete) concentration of the population in a single location. Intuitively, an initial infinitesimally greater concentration in one location leads to a greater productivity in that location in the next period through the historical agglomeration force, leading to yet more concentration in that location, eventually leading to near complete concentration in that location in the long-run. This illustrates how arbitrarily small initial differences in the distribution of population (or arbitrarily small temporary shocks that may give rise to such differences) can have large and permanent affects on the spatial distribution of economic activity.

3 The U.S. spatial economy, 1800-2000

We now describe a procedure for mapping the above model into observable features of the U.S. economy throughout the past two centuries. The goal is to estimate the model's parameters in this context. Armed with such estimates we turn in Section 4 to a set of counterfactual exercises designed to measure persistence and path dependence in the U.S. spatial economy.

3.1 Data

Our quantification requires data on population L_{it} and per-capita nominal incomes w_{it}. We therefore build a dataset drawing on Manson et al. (2017) that tracks these two variables for subnational regions i of the coterminous U.S. for as long a history as possible.

Starting with L_{it}, we obtain this series from decennial Census records of county-level population (by age group) from 1800 onward. To distinguish between children and adults in the model, we consider persons aged 25-74 as adults and work with 50-year steps (1800, 1850, 1900, 1950 and 2000) in order to avoid overlaps of these cohorts. Turning to w_{it}, for the years 1850-1950 we proxy for the relative amount of total income in any location, $w_{it}L_{it}$, by the estimated value of county-level agricultural and manufacturing output; for 2000 we use the per-capita income reported in the Census.[20] As a result, we have proxies for L_{it} and w_{it} from 1850-2000 and for L_{it} in 1800 as well; this allows estimation to proceed from 1850 onwards.[21]

[20] In practice, manufacturing output is not available for 1950 so we use the 1940 value of agricultural and manufacturing output. Per-capita income is not readily available prior to 1980.

[21] Appendix Figures C.1 and C.2 present maps of L_{it} and w_{it} in all years.

To account for county border changes over the years we work with the set of (the largest possible) sub-county regions that can be mapped uniquely to every county in our five years of data.[22] In the end, our sample consists of 4,975 such sub-county regions, which we refer to as "locations" (indexed by i) from now on.

Three other data sources play an auxiliary role in our model estimation: (i) migration flows; (ii) trade flows; and (iii) potential shifters of trade costs, productivities and amenities. We describe these further below.

3.2 Identification and estimation

We now describe a three-step procedure that estimates the unknown parameters of the model in Section 2. In a nutshell, the third step involves estimating a system of locational labor supply and demand equations that represent an augmented version of the spatial equilibrium model due to Rosen and Roback (Rosen 1979, Roback 1982); the first and second steps simply prepare the ingredients necessary to proceed in this tradition. We do this using available data on intranational trade X_{ijt} and migration L_{ijt} over our period of study.

3.2.1 Step #1: Estimating migration and trade costs

The goal of the first step is to determine the level of the migration and trade cost terms, raised to their respective elasticity exponents, that enter the equilibrium system of equations (13)-(16). We define these objects as $M_{ijt} \equiv \mu_{ijt}^{-\theta}$ and $T_{ijt} \equiv \tau_{ijt}^{1-\sigma}$.[23]

Consider first the estimation of migration costs. We begin by positing that migration costs depend on a proxy for passenger travel time, denoted $time_{ijt}$, through the relationship $\ln M_{ijt} = \kappa_t^{\mu} \times time_{ijt}$. We then proceed to estimate $time_{ijt}$ (between all location pairs ij in all years t) by assembling a geographic database that describes the network of navigable waterways (including canals), railroads, and roads of different types available in that year.[24] Given these networks, along with the observed topography, we use estimates of historical mode-specific travel speeds to determine the likely time it would take to traverse each square kilometer grid cell in the continental U.S. in any given year; the resulting speed maps are

[22] For example, suppose that county "A" in 1900 splits into "A1" and "A2" by 1950, and then "A2" splits into "A2(i)" and "A2(ii)" by 2000. The resulting sub-county regions that we track throughout would be "A1", "A2(i)" and "A2(ii)". We then apportion the county-level information into each of the sub-country regions on the basis of land area shares (and cluster all regression standard errors at the county-year level).

[23] Only bilateral-specific elements of such terms matter in this system because origin- or destination-specific components would be redundant conditional on the unrestricted values of $\overline{A}_{it}$ and $\overline{u}_{it}$. We therefore normalize any origin-time and destination-time components of T_{ijt} and M_{ijt} to one.

[24] The canal and railroad geographic data are based on shapefiles prepared by Atack (2015) and Atack (2016), respectively. We proxy for the 1950 road network with the 1959 road network geographic data from Jaworski & Kitchens (2019). The remainder of the geographic data derive from Allen & Arkolakis (2014).

reported in Appendix Figure C.3.[25] We then calculate the travel time along the fastest route between each pair of sub-county regions for each year using the Fast Marching Method (FMM) pioneered by Tsitsiklis (1995) and Sethian (1996) and applied to the spatial literature in Allen & Arkolakis (2014). This generates the set of travel times $\{time_{ijt}\}$.

We next estimate κ_t^μ using a non-linear least squares (NLLS) procedure that aims to match the model-predicted bilateral migration flows (from origin sub-county to destination sub-county) to observed bilateral migration flows (from origin state to destination county) in the data.[26] This procedure works as follows. Given any candidate $\tilde{\kappa}_t^\mu$, we construct its associated candidate migration costs $\tilde{M}_{ijt}$ from the estimated travel times. Given observed data on the population in each location in period t and $t-1$, $L_{i,t}$ and $L_{i,t-1}$, we then invert the model (applying Proposition 3 below) to recover the inward and outward migration market access terms and then use these terms and equation (11) to construct the unique set of bilateral sub-county to sub-county migration flows consistent with the observed location populations and candidate bilateral migration costs. We then aggregate these flows to the origin state to destination county level and calculate the sum of squared differences between observed (log) migration flows and these predicted (log) migration flows. Our estimated $\hat{\kappa}_t^\mu$ is the candidate $\tilde{\kappa}_t^\mu$ that minimizes these squared differences.[27]

We now turn to the estimation of trade costs. Our procedure is similar to that for migration costs, except that we now posit the relationship $\ln T_{ijt} = \kappa_t^\tau \times freight_{ijt}$, where $freight_{ijt}$ denotes a proxy for the user cost of freight shipping. We again estimate such costs for traversing each square kilometer in the U.S. in each year, as displayed in Appendix Figure C.4, and then apply the FMM to obtain $\{freight_{ijt}\}$, the least-cost route freight shipping cost for all location pairs and years.[28]

[25] In particular, we take the speed of travel by water and rail from Gordon (2016), who estimates speeds of 4 miles per hour by water (p.186)—which we hold constant across all years, 23.2 miles per hour by train in 1850, 33.7 miles per hour in 1900, and 49.8 miles per hour in 1950 and 2000 (averaging over the relevant routes from Table 5-1). For travel by road, we follow Jaworski & Kitchens (2019), who assume a speed of 25 miles per hour on unpaved roads, 45 miles per hour on paved state highways and minor arterial roads, 55 miles per hour on U.S. highways and principal arterial roads, and 70 miles per hour on interstate highways. For grid cells without water, rail, or roads, we calculate the speed of travel using Naismith's rule of 12 minutes per kilometer with an additional 10 minutes for each 100 meters of slope.

[26] We do this using the random samples of individual-level Census returns in Manson et al. (2017) for 1850 (a 1% random sample), 1900 (5%), 1950 (1%), and 2000 (5%). In each case we construct our measure of migration flows on the basis of where respondents aged 25-74 reside at the time of enumeration and where they were born.

[27] This estimation procedure is consistent with assuming that there is (classical) measurement error in the observed migration flows. We use a grid search algorithm to find $\hat{\kappa}_t^\mu$ and calculate our standard errors using a bootstrap procedure.

[28] While the geographic data used are the same as for the calculation of travel times $time_{ijt}$ used above, the mode-specific distance costs that we use for $freight_{ijt}$ are different. We obtain 1850 and 1900 costs of shipping from Donaldson & Hornbeck (2016), who, following Fogel (1964), estimate a cost of \$0.231 (in 1850 dollars) per ton-mile overland (where we again scale terrain via Naismith's rule), \$0.0063 per ton-

Finally, we estimate κ_t^τ by using a NLLS procedure that minimizes the difference between the model-predicted bilateral trade flows between sub-county pairs (given observed output Y_{it} and using equation 6) and the best available trade flow data for each year. For the years 2000 and 1950 we match the observed state-to-state trade flow data available from the 1997 Commodity Flow Survey and the 1949 rail waybill statistics collected by the Interstate Commerce Commission (as digitized by Crafts & Klein 2014), respectively. To our knowledge, however, there does not exist systematic intranational trade flow data in the U.S. for 1850 or 1900. To overcome this data limitation, we digitized the 1858 and 1900 *Chicago Commerce Reports* (Chicago Board of Trade 1859, 1901), and use the 1858 volume to inform our estimate for 1850. This source documents, for a range of important traded commodities, both (a) the local prices and (b) the quantities both arriving into and departing out of Chicago by mode of transit (i.e. particular rail line, canal, Great Lakes route, or wagon).[29] We then use FMM to calculate the mode of transit that would have been used (assuming cost-minimization) to travel to/from Chicago for every sub-county in the U.S., thereby partitioning the U.S. by mode of transit into Chicago.[30] For these years, our NLLS procedure then finds the κ_t^τ such that the observed value of Chicago imports/exports by mode of transit most closely matches the model-predicted Chicago imports/exports to/from the set of locations corresponding to that particular mode of transit.

3.2.2 Step #2: Recovering migration and trade market access terms

The goal of the second step is to invert a set of model equations in order to recover the inward and outward migration market access terms, Λ_{it} and Π_{it}, as well as the trade analogs, P_{it} and $\mathcal{P}_{it}$. This draws on the observed data on populations L_{it} and output $Y_{it} = w_{it}L_{it}$, in combination with the equilibrium structure of the model and the estimated migration and trade costs, $\widehat{M}_{ijt} = \exp\left(\widehat{\kappa}_t^\mu \times time_{ijt}\right)$ and $\widehat{T}_{ijt} = \exp\left(\widehat{\kappa}_t^\tau \times freight_{ijt}\right)$, from the previous step. To do so, we re-write the equilibrium system of equations (13)-(16) using equations (6)

mile via railroad, and \$0.0042 per ton-mile via water. In 1950, we reduce the costs of travel via railroad proportionately to the increase in the speed of travel from Gordon (2016) and then similarly update the costs of travel via road and water to match the relative costs via mode of travel estimated in Allen & Arkolakis (2014) (where we incorporate the fixed costs of shipping included in that study for the amount of a cross-country journey). Finally, to account for different costs across different types of roads, we assume the average speed of travel along road is 55 miles per hour and scale costs inversely proportional to the speed of each type of road.

[29]In 1858, we have price and quantity data for 18 commodities (flour, wheat, corn, oats, rye, barley, grass seed, beef cattle, live hog, dressed hog, hides, salt, wool, highwines, lath, shingles, lumber, and siding), allowing us to generate the total value of Chicago imports and exports by canal, lake, overland, and for 10 different rail lines. In 1900, we have price and quantity data for 36 commodities imported and/or exported by canal, lake, and for 23 different rail lines. See panel (a) of Appendix Figure C.5 for an example.

[30]Panel (b) of Figure C.5 depicts the resulting basin for each mode of transit into Chicago in 1850.

and (11), which yields, for all i:

$$\mathcal{P}_{it}^{1-\sigma} = \sum_j \widehat{T}_{ijt} \times Y_{jt} \times \left(P_{jt}^{1-\sigma}\right)^{-1}, \tag{22}$$

$$P_{it}^{1-\sigma} = \sum_j \widehat{T}_{jit} \times Y_{jt} \times \left(\mathcal{P}_{jt}^{1-\sigma}\right)^{-1}, \tag{23}$$

$$\left(\Lambda_{it}^{\theta}\right)^{-1} = \sum_j \widehat{M}_{jit} \times L_{jt-1} \times \left(\Pi_{jt}^{\theta}\right)^{-1}, \tag{24}$$

$$\Pi_{it}^{\theta} = \sum_j \widehat{M}_{ijt} \times L_{jt} \times \Lambda_{jt}^{\theta}. \tag{25}$$

The following proposition shows that the four remaining unknown variables—comprising the inward and outward trade and migration market access terms—in equations (13)-(16) are identified (up to an inconsequential scale factor), when raised to the exponents $\sigma - 1$ or θ, because this system of equations has a unique solution given data on Y_{it}, L_{it} and estimates of $\widehat{T}_{ijt}$ and $\widehat{M}_{ijt}$ from step #1.

Proposition 3. *Given observed data on $\{Y_{it}, L_{it}, L_{it-1}\}$ and given values of $\left\{\widehat{T}_{ijt}, \widehat{M}_{ijt}\right\}$ there exists a unique (up to scale) set of values of $\left\{\mathcal{P}_{it}^{\sigma-1}, P_{it}^{\sigma-1}, \Pi_{it}^{\theta}, \Lambda_{it}^{\theta}\right\}$ that satisfy equations (22)-(25).*

Proof. See Section A.5. □

Note that this inversion does not require any assumption regarding the value of the trade elasticity σ, migration elasticity θ, or degree of forward-looking behavior δ.

3.2.3 Step #3: Estimating the spillover elasticities

The third step of our estimation procedure uses the outputs of step #2 in order to estimate contemporaneous (α_1 and β_1) and historical (α_2 and β_2) spillover elasticities via an augmented Rosen-Roback procedure.

To see this, begin by substituting the productivity spillover function from equation (1) into the outward trade market access $\mathcal{P}_{it}$ from equation (7) and imposing $Y_{it} = w_{it}L_{it}$. This reveals the following (inverse) demand equation for labor in location i:

$$\ln w_{it} = \left[\alpha_1 \left(\frac{\sigma-1}{\sigma}\right) - \frac{1}{\sigma}\right] \ln L_{it} + \alpha_2 \left(\frac{\sigma-1}{\sigma}\right) \ln L_{it-1} + \frac{1}{\sigma} \ln \mathcal{P}_{it}^{1-\sigma} + \frac{\sigma-1}{\sigma} \ln \overline{A}_{it}. \tag{26}$$

In this expression, the inverse elasticity of labor demand combines the inverse elasticity of demand for goods from a location, $-\frac{1}{\sigma}$, with the contemporaneous productivity spillovers α_1;

this latter effect is moderated by $\frac{\sigma-1}{\sigma}$ because the location faces a downward-sloping demand curve for its output. Notably, with strong positive spillovers the labor demand curve can be upward-sloping. Also present in the demand equation are a set of shifters: (i) lagged population L_{it-1}, which raises productivity if there are historical productivity externalities (i.e. $\alpha_2 > 0$); (ii) the outward trade market access $\mathcal{P}_{it}$, which allows for the labor demand in location i to be high if its ability to sell goods to other locations is high; and (iii) the exogenous (and unobserved) component of productivity, $\overline{A}_{it}$. Importantly, this estimating equation describes a cross-sectional relationship that holds within any equilibrium, so it can be used for valid point estimation even if the model's parameters lie in a region for which multiplicity occurs.

The inverse labor supply curve can be obtained similarly. Substituting the amenity spillover function from equation (3) and the value function from equation (17) into the inward migration market access Λ_{it} from equation (12), and using $W_{it} = \left(\frac{w_{it}}{P_{it}} u_{it}\right)$, we obtain:

$$\ln w_{it} = \left(\frac{1}{\theta} - \beta_1\right) \ln L_{it} - \beta_2 \ln L_{it-1} + \frac{1}{\theta} \ln \Lambda_{it}^{\theta} + \frac{1}{1-\sigma} \ln P_{it}^{1-\sigma} - \frac{\delta}{\theta} \ln \Pi_{it+1}^{\theta} - \ln \overline{u}_{it}. \quad (27)$$

The inverse elasticity of labor supply combines the locational utility heterogeneity dispersion θ with the contemporaneous productivity spillovers β_1; analogously to the demand case, the elasticity of labor supply can be negative if such spillovers are positive and large. Shifters of the inverse labor supply curve comprise: (i) the lagged population in the location L_{it-1}, which matters to the extent that historical amenity externalities exist (i.e. $\beta_2 \neq 0$); (ii) the consumer cost-of-living P_{it}, which increases the nominal wage w_{it} that is required for a given amount of mobile workers to be willing to live in location i; (iii) the present discounted value of a location one period in the future Π_{it+1} (but discounted by δ), which reduces the nominal wage w_{it} necessary for mobile workers to be willing to live in a location; (iv) the inbound supply of potential migrants from other nearby locations as captured by Λ_{it}; and (v) the exogenous (and unobserved) component of location i's amenity, $\overline{u}_{it}$.[31] Again, this equation allows parameter estimation to proceed even though equilibria may be multiple.

The locational demand-supply system in equations (26) and (27) generalizes that in the Rosen-Roback framework (c.f. Roback 1982, Glaeser & Gottlieb 2009, Kline & Moretti 2014,

[31]Estimation of equation (27) requires data on Π_{it+1}^{θ} in all years, which may not be available since in some contexts (such as ours) the final year t of interest for estimation ($t = 2000$ for us) may also be the last year with available data. However, because we can always, given knowledge of the elasticity parameters, solve the model one period forwards we can view $\Pi_{i,2050}^{\theta}$ as a (nonlinear) function of those parameters, allowing application of NLLS. In practice, we find that the values of $\ln \Pi_{i,2000}^{\theta}$ and $\ln \Pi_{i,2050}^{\theta}$ are highly correlated, so that a simple strategy of proxying for the missing final value of Π_{it+1}^{θ} with its last available value would be extremely accurate.

Hsieh & Moretti 2019) in several respects. First, it relaxes the assumption that locations produce a homogeneous and freely traded product (i.e. that σ is infinite and $\tau_{ijt} = 1$). Second, it relaxes the assumption that all workers have identical preferences across locations and face no costs of migrating (i.e. that θ is infinite, and $\mu_{ijt} = 1$). This added flexibility necessitates the inclusion of the contemporaneous and forward-looking market access terms $\mathcal{P}_{it}^{1-\sigma}$, $P_{it}^{1-\sigma}$, Π_{it+1}^{θ} and Λ_{it}^{θ} as demand and supply shifters, as recovered in step #2. Finally, it allows for historical populations L_{it-1} to affect contemporaneous labor demand and supply via historical spillovers (α_2 and β_2).

We turn now to three details of estimation in our context. First we augment the estimating equations (26) and (27) to include location fixed effects so that all time-invariant geographic elements of labor demand and supply are controlled for. We further add controls for an interaction between broad geographic region identifiers and year effects so that features of the spatial reorganization of the US economy (such as those highlighted by Kim & Margo 2014) are controlled for.[32] Including year effects in this way is also important since Proposition 3 clarifies how the included (log) market access terms are only identified up to an (additive) scale factor in each year.

Second, while equations (26) and (27) could, in principle, be used to identify both the spillovers parameters ($\alpha_1, \alpha_2, \beta_1$ and β_2) and the preference parameters (δ, σ and θ), in practice the cross-spatial variation is not well suited to estimating the latter precisely because they are coefficients on market access variables that are, in their nature, highly correlated over space. We therefore use values of the preference parameters obtained in related U.S.-based studies: we set the intertemporal discount factor $\delta = 0.0535$ $(=\left(\frac{1}{1.0603}\right)^{50})$ to match the 6.03% average annualized return on wealth in the U.S. from 1870-2015 as measured by Jordà et al. (2019), the elasticity of substitution $\sigma = 9$ to match Donaldson & Hornbeck (2016), and the migration elasticity $\theta = 4$ to match Monte et al. (2018).[33] However, in what follows we assess the robustness of the results to alternative parameter values.

Finally, as with any demand-supply system, OLS estimates of the parameters in equations (26) and (27) would generically suffer from simultaneity bias. We therefore use an instrumental variable (IV) procedure that draws on the insight in Roback (1982) that observable components of amenity changes that are uncorrelated with productivity changes would be valid instruments for estimating the demand equation (26)—and vice versa for the supply

[32] To construct regions, we draw a box around the continental U.S. (in the Mercator projection) and, beginning from the southwest corner of the box, overlay squares on top of the box, each of which has an area equal to one tenth of the area of the box. This partitions the continental U.S. into 14 different regions.

[33] Donaldson & Hornbeck (2016) estimate a trade elasticity of 8.22 (implying $\sigma = 9.22$) when focusing on intranational trade in the U.S. during the late 19th century. Monte et al. (2018) estimate a location choice elasticity across U.S. counties of 3.30 over a five-year period, albeit in a static framework abstracting from bilateral migration costs.

equation (27). Given the 50-year time intervals and location fixed effects that we use for estimation, and the goal of estimating both contemporaneous and lagged spillovers, these instruments must derive from relatively long-run changes to the U.S. economy.

For the demand equation, we follow Barreca et al. (2016) who note that technological advances like air conditioning and more effective heating systems have made extreme hot and cold climates more bearable (delivering greater amenity value) throughout our sample period. Accordingly, our IVs consist of a linear time trend interacted with the average maximum temperature in the warmest month and the average minimum temperature in the coldest month (and their squared values to allow for nonlinearities) in each location. We obtain such data from WorldClim.org.

For the labor supply equation instruments, we leverage two major changes in U.S. agriculture over the past 200 years. The first is the increased use of more intensive cultivation practices (e.g. mechanization, fertilizer, genetic modification of seeds, etc), which raised land productivity. Following Bustos et al. (2016), we measure the extent to which locations could take advantage of this higher-intensity cultivation as the differential potential yield under low and high intensity cultivation, according to the FAO-GAEZ agroclimatic model of crop suitability (Fischer et al. 2008). Our first IV interacts this differential yield for corn, the dominant crop throughout our period, with a linear time trend.[34] The second major change that we exploit is a shift in world demand that has altered which crops are grown in the U.S., most notably soy.[35] To proxy for which locations saw the greatest gain in (revenue) productivity from this shift, we use the FAO-GAEZ predicted difference in potential yield between soy and wheat (a crop for which demand has remained relatively constant over time) and interact this with a linear time trend.[36] Together, these two sets of supply-equation instruments leverage heterogeneity in geographical exposure to both within-crop and across-crop changes among the three most important food crops for U.S. agriculture.

Finally, when estimating the demand equation (26) we use the climate amenity-based IVs, but additionally control for the agricultural productivity variables (in order to reduce residual variation and the risk that our amenity-based IVs are correlated with unobserved productivity variation). Analogously, our estimation of the supply equation (27) includes

[34]To allow for within-location heterogeneity in agroclimatic suitability, we include both the mean differential yield for corn and the standard deviation of the differential yield as instruments.

[35]Virtually absent in 1900, soy trailed only corn in terms of both value and acreage in 2000. Roth (2018), for example, argues that much of this rise is due to rising demand for U.S. exports of soy to Asia.

[36]In 1909, wheat was cultivated on 14.7% of harvested acres allocated to principal crops; in 2000, the figure was 17.2%; see USDA (2003). In practice, we use the high- and low-intensity scenarios for soy and wheat, respectively, to reflect the fact that the former was grown predominantly in a more technologically advanced era. As with the first labor supply instrument, we include both the mean soy-wheat differences and the standard deviation of the differences as instruments.

controls for the climate amenity variables.

To conclude the three-step procedure we note that, conditional on obtaining consistent estimates of the elasticity parameters, equations (26) and (27) allow recovery of the geographic fundamentals $\left\{\bar{A}_{it}, \bar{u}_{it}\right\}$ as well. Combined with the earlier estimates of $\{T_{ijt}, M_{ijt}\}$ from step one all model parameters are thereby identified.

3.3 Estimation results

We begin with the estimation of trade and migration costs in step #1, as reported in Table 1. Panel (a) presents the results for migration costs. We obtain $\{\widehat{\kappa}^{\mu}_{1850}, \widehat{\kappa}^{\mu}_{1900}, \widehat{\kappa}^{\mu}_{1950}, \widehat{\kappa}^{\mu}_{2000}\} = \{0.02, 0.06, 0.07, 0.07\}$ (with bootstrapped standard errors of $\{0.002, 0.001, 0.002, 0.001\}$), indicating migration costs have become more responsive to travel times over the past 150 years.[37] These magnitudes suggest that, for example in 1850, doubling the distance of migration via rail from 500 to 1,000 miles caused migration flows to decline by roughly a third, whereas in 2000 a doubling of distance caused migration flows to decline by roughly half.

Turning to trade costs, Panel (b) of Table 1 reports our estimates. In this case we find that $\{\widehat{\kappa}^{\tau}_{1850}, \widehat{\kappa}^{\tau}_{1900}, \widehat{\kappa}^{\tau}_{1950}, \widehat{\kappa}^{\tau}_{2000}\} = \{0.7, 0.5, 0.3, 0.4\}$ (with bootstrapped standard errors of $\{0.4, 0.4, 0.01, 0.01\}$). This implies that intra-U.S. trade costs have become less responsive to freight rates over the past 150 years.[38] The estimated magnitudes mean, for example, that doubling the distance of goods shipped by water from 500 to 1,000 miles caused trade flows to decline by roughly three quarters in 1850, whereas a similar doubling of distance caused trade flows to fall by roughly half in 2000.

Given the estimated trade and migration costs and the observed distributions of population and output, in step #2 we apply Proposition 3 to recover the market access terms $\{\mathcal{P}^{\sigma-1}_{it}, P^{\sigma-1}_{it}, \Pi^{\theta}_{it}, \Lambda^{\theta}_{it}\}$. For example, Appendix Figure C.6 depicts what the recovered inward market access parameter Λ^{θ}_{it} implies for the present discounted value of residence V^{θ}_{it} (since $\Lambda^{\theta}_{it} = V^{\theta}_{it}\left(\frac{L_{it}}{\bar{L}}\right)^{-1}$). One pattern on display is that the present discounted value of residing in relatively densely populated locations has increased over time, resulting in an increasingly concentrated distribution of V^{θ}_{it}.

Does this pattern arise from agglomeration forces or from changes in the underlying geography? To answer this question, we turn to step #3. The parameter values implied by our

[37] That agents have become more responsive to migration costs is consistent with the evidence of Kaplan & Schulhofer-Wohl (2017), who find that U.S. interstate migration rates have fallen over the past 20 years despite declines in travel costs.

[38] Consistent with our estimates for the 20th century, Disdier & Head (2008) conduct a meta-analysis of 103 different papers estimating the relationship between trade flows and distance and find that distance has a larger impact on trade after 1970 than before 1970, although most of the analyzed papers estimate the gravity relationship after 1970 (and none of the papers examine the gravity relationship before 1870).

2SLS estimates of the labor demand equation (26) are reported in Table 2.[39] We present both OLS and 2SLS estimates for three values of the elasticity of substitution ($\sigma = \{5, 9, 14\}$). Regardless of the chosen elasticity of substitution, we estimate a large positive contemporaneous agglomeration spillover α_1 and a small (and statistically insignificant) negative historical agglomeration effect α_2. At our preferred value of $\sigma = 9$, the 2SLS estimates are $\widehat{\alpha}_1 = 0.19\ (SE = 0.040)$ and $\widehat{\alpha}_2 = -0.041\ (SE = 0.045)$.[40,41]

Table 3 displays analogous 2SLS estimates of the parameters in the locational labor supply equation (27). As in the labor demand equation, we report the estimated amenity spillover parameters for a number of combinations of possible trade and migration elasticities spanning the range of values from the literature ($\sigma \in \{5, 9, 14\} \times \theta \in \{2, 4, 6\}$).[42] For brevity we only report the 2SLS estimates. Across all possible combinations, we estimate negative (but statistically insignificant) contemporaneous amenity spillovers β_1 and positive historical amenity spillover β_2, exactly as one would expect from the presence of a durable housing stock (see Appendix B.2). In our preferred specification, with $\sigma = 9$ and $\theta = 4$, we estimate $\widehat{\beta}_1 = -0.26\ (SE = 0.265)$ and $\widehat{\beta}_2 = 0.31\ (SE = 0.178)$. The similar spillover sizes (in absolute value) is consistent with housing that is durable at 50-year time scales.

What do these estimated spillovers imply for the degree of persistence and the possibility of path dependence? To answer this question, we return to Figure 1, where our preferred estimates from Tables 2 and 3 and are illustrated in the context of the parameter thresholds identified in Corollary 1, Proposition 1, and Proposition 2 (evaluated at our preferred values of σ, θ, and δ). The red star in Panel (a) indicates the location of the contemporaneous spillover estimates, $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ (and the red oval indicates the 95% confidence interval, or CI, for these estimates). From Corollary 1, its location in the yellow region indicates that the dynamic path of the economy's (bounded) equilibrium is unique—that is, given any initial distribution of population $\{L_{i0}\}$ and known evolution of geography $\{\tau_{ijt}, \mu_{ijt}, A_{it}, u_{it}\}$, we can uniquely determine the evolution of the economy. However, we know from Proposition 1 that the red star's location near the boundary of the parameter region in which uniqueness

[39] In Tables 2 and 3, the reported standard errors are two-way clustered at the location level and at the county-year level. First-stage estimates are presented in Appendix Table C.1 and Appendix Figure C.7 maps the spatial patterns of the predicted change in population from the first-stage regressions.

[40] Our estimate of the parameter α_1 is similar to the value of 0.2 obtained (for the manufacturing sector) by Kline & Moretti (2014), though it is smaller than the range of values (1.25-3.1) implied by the estimates obtained (again for manufacturing) by Greenstone et al. (2010) (as discussed in Kline & Moretti 2014). Our estimate of $\alpha_1 + \alpha_2 = 0.15$ is slightly higher than the value (0.09) estimated by Bleakley & Lin (2012).

[41] As reported in Table 2, the minimal first-stage Sanderson & Windmeijer (2016) F-statistic (taken across the two first-stage equations) in this regression is 58.1, indicating that finite-sample 2SLS bias is unlikely. The same is true for our labor supply equation estimates in Table 3.

[42] The choice of δ has only a minuscule effect on these results so Table 3 reports estimates of β_1 and β_2 obtained while using our preferred value of δ only.

is assured suggests the possibility of very persistent historical shocks. Finally, Panel (b) illustrates the parameter space for $\alpha_{ss} \equiv \alpha_1 + \alpha_2$ and $\beta_{ss} \equiv \beta_1 + \beta_2$, the combination of contemporaneous and historical spillovers that matters for characterizing steady states as per Proposition 2. Our estimates are indicated by the green star. Evidently, this point estimate (along with much of its accompanying CI) lies inside the (blue) region that, according to Proposition 2, suggests the possibility of multiple steady-states, and hence the possibility that shocks could exhibit path dependence.

4 Persistence and path dependence in the U.S.

We have just seen in Figure 1 how our model, when estimated on U.S. data from 1800-2000, features unique equilibria, but nevertheless the possibility of both long-lived persistence and path dependence. We now seek to quantify such phenomena by considering how the U.S. economy would look, both today and in the future, under alternative historical conditions.

4.1 Dynamic model evolution

Prior to studying how alternative historical conditions would have affected the evolution of the U.S. economy, there is value in first exploring how the model's dynamic path would have evolved in a hypothetical U.S. economy without any evolution of productivity or amenity differences over space. This allows us to assess the role that trade, migration, and agglomeration forces—along with the estimated evolution of the trade and migration costs—play in shaping the evolution of our model economy.

To do so, we calibrate the model to the initial year, $t_0 = 1850$, and then simulate it forwards while holding $\left\{\bar{A}_{it}, \bar{u}_{it}\right\}$ fixed at their t_0 values.[43] This generates a simulated stream of values for predicted population $\hat{L}_{it}$ in each location and year $t \geq t_0$. We then compare this simulated path to the actual observed path of population L_{it} by estimating the regression

$$\ln L_{it} = \beta \ln \hat{L}_{it} + \delta_i + \delta_t + \varepsilon_{it},$$

which includes location and year fixed effects so that the coefficient β provides a comparison of relative changes in the spatial distribution of the economy with such changes in the model with fixed productivities and amenities. We then repeat this exercise for output, Y_{it} and $\hat{Y}_{it}$.

[43] All simulations in this section use a finite-horizon economy (with length T and $\Pi_{i,T+1} = 1$) to approximate the infinite-horizon economy of Section 2. Our choice of $T = 3500$ is driven by the fact that in all simulations the economy appears to be in steady-state (up to numerical precision) by no later than $t = 3000$, and that varying our choice of T beyond that point is inconsequential.

Column 1 of Table 4 presents the results from this exercise (using population in Panel A and output in Panel B). We see a positive and statistically significantly correlation between observed and hypothetical changes in the spatial distribution of economy activity. And the (within) R^2 values indicate that a model with productivities and amenities held fixed to their 1850 values accounts for 11% of the observed variation in the changes in the spatial distribution of population, and 13% of that for output, over the ensuing 150 years. Columns 2-7 go on to illustrate that these findings are similar across a variety of subsets of locations, including those with initially high and low populations (columns 2 and 3), and those in Northern, Southern, and Eastern locations (columns 4, 5, and 7), athough Western locations (column 6) appear to behave differently in this regard (as is perhaps unsurprising given the paucity of economic data for those locations in 1850).

This exercise highlights how a large share of the evolution of population and output in the US economy since 1850 reflects productivity and amenity shocks.[44] Such a finding motivates our next step: a counterfactual exercise that explores alternative historical scenarios based on a certain set of spatial rearrangements of these shocks.

4.2 The effect of history on contemporary outcomes

How different would the U.S. spatial economy look today if historical conditions had been different? To answer this question, we need to compare actual historical conditions to counterfactual alternatives. While one could imagine many counterfactual histories of interest, we focus on one that draws inspiration from the vagaries of relative industrial success that struck America's communities at the turn of the 20th century. This period—known as the Technological Revolution or the Second Industrial Revolution—was a period of rapid productivity growth across a number of different industries due to the widespread adoption of technological innovations such as the internal combustion engine and electrification.[45]

Crucially for us, the adoption of these innovations varied across locations within the United States, often for reasons that may plausibly have involved elements of historical "luck". For example, Detroit's rise as the "Motor City" may owe something to the fact that Henry Ford happened to be born on a nearby farm. Or perhaps Buffalo became the "City of Light" in more than just a name because it was chosen to host the Pan-American Exposition at a time (1901) when Thomas Edison desired to demonstrate his newly invented AC power, and did so by adorning Buffalo's Exposition buildings with light bulbs.

[44]By contrast, this exercise does not provide a sense of the model's "fit" since, by design, the model's estimated productivity and amenity values $\{\bar{A}_{it}, \bar{u}_{it}\}$ exactly match the observed data on population and output in all locations and years.

[45]See e.g. Landes (2003).

Examples like these suggest that relatively similar locations may (or may not) have been the fortunate recipients of positive productivity shocks in a time of technological change. To study the consequences of such hypotheticals, we generate a set of counterfactual histories in which productivity fundamentals are randomly swapped between pairs of similar locations. For example, what if Cincinnati (with a population of 330,000 in 1900) had been chosen instead of Buffalo (population 350,000 in 1900) as the site of 1901's Exposition?

To operationalize this idea, albeit in an abstract manner, we carry out a set of B simulations, each indexed by b, as follows. First, we rank all locations in terms of their observed population in 1900, $L_{i,1900}$. Second, we form pairs p of locations based on their nearest neighbor in this ranked distribution, starting at the top; for example, Erie County, NY (home to Buffalo) and Hamilton County, OH (home to Cincinnati) occupy ranks 11 and 12 in the distribution.[46] Third, in simulation b we draw (independently) for each pair p a random variable S_p (with realization $s_p^{(b)}$ in simulation b) that is distributed $Bernoulli(1/2)$. When $s_p^{(b)} = 1$, we swap the values of the fundamental productivity in 1900 (i.e. $\overline{A}_{i,1900}$) among the two locations i within pair p; and when $s_p^{(b)} = 0$ we leave the pair unchanged. Fourth, we then simulate the model forwards from 1900 onward while holding fixed all other exogenous locational characteristics in the model (i.e. $L_{i,1850}, \overline{u}_{i,1900}$, and the entire path of $\overline{A}_{it}$ and $\overline{u}_{it}$ for $t > 1900$) at their values estimated in Section 3.3. We also set $\overline{A}_{it} = \overline{A}_{i,2000}$ and $\overline{u}_{i,t} = \overline{u}_{i,2000}$ and hold them fixed for all $t > 2000$. This generates a stream of counterfactual predictions for all the endogenous variables in the model (which we denote as $L_{it}^{(b)}, V_{it}^{(b)}$, etc.) at all dates $t \geq 1900$, though in practice we stop at $T = 3500$, or 30 generations after the year 2000 in which the economy's fundamentals become time-invariant.

We then repeat these four steps for all B simulations (and set $B = 100$ in practice). We will also conduct on an additional $(B+1)^{th}$ simulation (the output of which we label as, for example, $L_{it}^{(F)}$, for "factual") in which there are no swaps at all. This corresponds, for $t \leq 2000$, to the factual path taken by variables such as L_{it} in the data. For years $t > 2000$ this exercise therefore simulates forward a model that (by design) fits the past data perfectly.

To summarize, each of these "swap" counterfactual history simulations holds everything in the model constant apart from the fundamental sources of productivity in 1900, $\overline{A}_{i,1900}$. And even the $\overline{A}_{i,1900}$ *distribution* is held exactly constant (not just on aggregate but also that across the $N/2$ pairs of locations). The only thing being perturbed in any counterfactual history is the within-pair assignment of productivity in 1900 among pairs of locations that

[46] Other examples of pairs include the counties that are home to Worcester and San Francisco (ranks 17 and 18), Providence and Baltimore (ranks 21 and 22), New Haven and New Orleans (ranks 23 and 24), and Louisville and Minneapolis (ranks 27 and 28). Due to the odd number of locations, that with the smallest 1900 population—a subset of Craig County, VA—is without a partner. We therefore leave its productivity unchanged in every simulation.

are as close as possible to one another in terms of their 1900 populations.

4.2.1 Fragility and resilience

We begin by examining how the spatial distribution of the U.S. economy in the year 2000 varies across different counterfactual histories. Which locations tend to be fragile in the face of these shocks and which tend to be resilient? Figure 4 shows a plot of the standard deviation of the population $\ln L_{i,2000}^{(b)}$ (as well as the PDV $\ln V_{i,2000}^{(b)}$) across the B simulations against the factual population in 2000, $L_{i,2000}^{(F)}$. There are two key take-aways. First, small historical shocks 100 years in the past have substantial impacts on nearly all locations, with the average standard deviation across simulations in log population of 0.38 across simulations. Second, while no location is immune to these shocks, locations that are in reality more populated exhibit greater resilience to historical shocks; for example, the top quartile of locations by factual population in 2000 had an average standard deviation of 0.35 compared to 0.42 for the bottom quartile.

Similar patterns are on display for the (log) present discounted value, $\ln V_{i,2000}^{(b)}$, but in a substantially dampened fashion throughout the distribution, as we would expect due to the spatial smoothing facilitated by trade and migration. However, even the PDV is hardly impervious to 100 year-old shocks; for example, the largest 25 locations have an average standard deviation of $\ln V_{i,2000}^{(b)}$ of 0.07.

4.2.2 Luck

Having seen the large effects of historical shocks on modern outcomes on display in Figure 4, a natural question is how the factual history compares to other possible histories—that is, how lucky was "our" particular history? To evaluate this, we compare the factual spatial distribution of present discounted values $V_{i,2000}^{(F)}$ to the distributions of all other B simulations. Figure 5 displays the (population-weighted) median and interquartile range of the log PDV across all histories, relative to our own. As is evident, our history was relatively lucky, with the median person enjoying a larger PDV from their residence in the factual scenario than in 75 of the 100 other simulations. Moreover, the magnitudes concerned are substantial: for example, the median person in the factual economy has a PDV that is 8.4% greater than that in the tenth lowest alternative history (although also 2.1% worse than then tenth highest).

What makes our history lucky relative to the alternative histories considered? To shed light on this question, we examine how the factual year 2000 spatial distribution of economic activity differed from alternative histories. To do so, for each history b, we order each location by either its population or its PDV. Then, for each location, we calculate the fraction of alternative histories for which the factual rank of that location exceeded the alternative history rank. For example, in reality, Minneapolis, MN was the 32nd most populated location in the year 2000; we calculate the fraction of the b counterfactual histories for which Minneapolis was ranked lower than 32nd. Figure 6 presents the results. As is evident, locations in New England, New York, and the upper Midwest—commonly referred to as the "rust belt"—had greater population rankings (panel a) and PDVs (panel b) in the factual history than in most other counterfactual histories considered. On the other hand, locations in the Great Plains, Southwest and Rocky Mountain states were not so lucky.[47]

4.2.3 Persistence

The two previous results find that small historical shocks, even a century ago, can play a large role in determining the spatial distribution of contemporary U.S. economic activity, suggesting the presence of substantial persistence. We now quantify the extent of such persistence directly. While the exogenous changes in our simulations are to productivities $\overline{A}_{i,1900}$, a useful way to summarize the effect of these shocks can be obtained by noting that, from the perspective of any year $t > 1900$, the only impact of these shocks is to alter $\{L_{i,1900}\}$, the initial conditions of the model's only state variable. We therefore study the impact of a change in such initial conditions, rather than the underlying shocks to $\overline{A}_{i,1900}$ that altered these initial conditions, as follows.

For any generic "outcome" of interest, O_{it}, we use the data generated by the model simulations $b = 1...B$ in order to estimate the regression

$$\ln O_{it}^{(b)} = \delta_{it}^{O} + \eta_{it}^{O} \ln L_{i,1900}^{(b)} + \varepsilon_{it}^{O(b)} \tag{28}$$

separately for each location i and time period $t > 1900$. Our interest lies in the *persistence elasticity* (for outcome O), denoted by η_{it}^{O}. This elasticity measures the average relationship, across the B simulations, in location i between that location's historical population $L_{i,1900}^{(b)}$ and its value for the outcome $O_{it}^{(b)}$ in some later period t. The error term $\varepsilon_{it}^{O(b)}$ in equation (28) is almost surely correlated with $L_{i,1900}^{(b)}$, for any outcome—as, for example, equation (20) makes clear when O_{it} represents population. But $\ln \overline{A}_{i,1900}^{(b)}$ can serve as a valid IV for consistent estimation of η_{it}^{O} given that it is randomly assigned (by design) and excludable

[47]This is partially due to the fact that rust belt locations tended to be the member of population-matched pairs with the higher value of $\overline{A}_{i,1900}$. For example, a regression of the variable plotted in Figure 6 (the fraction of alternative histories with a worse rank for a location than its factual rank) on the within-pair rank of $\overline{A}_{i,1900}$ has an R^2 of 0.31 for population and 0.48 for PDV. But evidently the majority of the variation in "luck" rests on factors beyond a location's own productivity draw, such as the spatial configuration of other locations' characteristics.

(given that, as discussed, population is the model's only state variable).

Figure 3 reports the distribution (across locations) of the estimated values of the persistence elasticity for population $\widehat{\eta}_{it}^{L}$ that corresponds to each of the years t = 1950-2500.[48] The results confirm substantial persistence of temporary shocks: for example, the median population elasticity $\widehat{\eta}_{i,2000}^{L}$, 100 years after the simulated shocks, is 0.37.[49] While there is considerable heterogeneity across locations owing to differences in migration and trade market access, even the fifth percentile value in the population elasticity distribution is 0.33. This suggests that in a dynamic economic geography model like the one developed here, it should be considered the norm, rather than the exception, to observe that a local event that raises a location's population at a point in time leads to centuries-long economic persistence.[50]

Also shown in Figure 3 is the distribution of the elasticities $\widehat{\eta}_{it}^{V}$, where the outcome O_{it} is the present discounted value $V_{it}^{(b)}$. As expected, the PDV elasticities $\widehat{\eta}_{it}^{V}$ are considerably lower than the local population elasticities $\widehat{\eta}_{it}^{L}$—because of trade and migration, the PDV draws on both local and nearby geographical advantage, so the local impact of local shocks is muted by spatial interactions and arbitrage. Still, the median elasticity in the year 2000, $\hat{\eta}_{i,2000}^{V}$, is 0.09, suggesting substantial persistence in PDV despite these attenuating forces.

4.3 The effect of history on future outcomes

Taken together, the previous results suggest that the distribution of economic activity today depends strongly on the vicissitudes of history because temporary shocks have long-lasting effects. But do the temporary shocks induced by our swap counterfactuals exhibit any permanent effects? That is, do we see evidence for path dependence?

To answer this question, we return to Figure 3, this time looking many centuries beyond the year 2000. While this is undoubtedly a heroic exercise, it provides a direct way to assess the path-dependent properties of our model. Figure 3 demonstrates a clear sense of (very slow) convergence to a unique steady-state. That is, while there are persistent effects of the historical shock for hundreds of years, by the year 2500—600 years after the shock occurred—the estimated persistence elasticities are essentially zero in all simulations.

[48] To account for the fact that the elasticities $\widehat{\eta}_{it}^{O}$ are estimates, in this figure we weight each location by the inverse of the square of the standard error of its estimate. The instruments are typically very strong, with a mean (median) first stage F-statistic of 153 (29) and 69% of locations' F-statistics exceeding 10.

[49] This value is very close to the (squared) partial elasticity of population persistence given in equation (20) of $\left(\frac{\alpha_2(\sigma-1)+\beta_2\sigma}{1+\frac{\sigma}{\theta}-(\alpha_1(\sigma-1)+\beta_1\sigma)}\right)^2 \approx 0.37$, implying that, for the median location, and on average across simulations, the equilibrium effects of the market access terms in (20) are approximately zero.

[50] As a point of comparison, Peters (2022) uses variation in refugee assignment across locations in post-World War II Germany to estimate a 50-year population persistence elasticity of 0.88, which is actually larger than the equivalent predicted by our model (0.61 on average).

In addition, we find that by the year 3000 there is no simulation in which the correlation between its distribution of log population and that in the factual history is smaller than 0.9997. That is, there is no evidence of these small temporary shocks resulting in permanent changes to the distribution of the U.S. spatial economy.[51]

However, as Proposition 2 highlights, the possibility of path dependence depends on the strength of the combined contemporaneous and historical productivity spillovers $\alpha_1+\alpha_2$ and amenity spillovers $\beta_1+\beta_2$. The above simulations were conducted by using the point estimates of these parameters (the green star in Figure 1), without regard to the uncertainty that accompanies these estimates (the dashed green ellipse, indicating the 95% CI, in Figure 1). A natural question is whether the above finding, that our historical 1900 swap counterfactuals exhibit no path-dependent effects, continues to be true throughout the CI.

To investigate, we repeat the above simulations across a range of values of $\alpha_1+\alpha_2$ (and analogously, $\beta_1+\beta_2$) within its CI. In order to reduce the number of cases to consider, we focus on increasing the historical spillover parameter α_2 (while holding α_1 constant) towards the upper limit of the 95% CI of $\alpha_1+\alpha_2$.[52] This mimics the example in Section 2.4: holding contemporaneous spillovers α_1 constant, increasing the strength of historical spillovers α_2 can make path dependence more likely. For each alternative value of α_2 considered, we re-calculate the underlying distribution of productivities and amenities $\left\{\bar{A}_{it},\bar{u}_{it}\right\}$ to exactly match the observed distribution of economic activity, re-perform the 100 alternative histories—along with the one factual history—where we randomly swap productivities in 1900 between similar locations, and re-simulate the entire evolution of the economy from 1900 onwards.[53] For each simulation b, we calculate the (population weighted) average (log) present discounted value in the year 3000, $\ln\bar{V}_{3000}^{(b)}\equiv\sum_i\left(\frac{L_i}{\bar{L}}\right)\ln V_{i,3000}^{(b)}$, in order to quantify how such long-run aggregate welfare depends on small shocks in the distant past.

Figure 7, panel (a) presents the results. Raising α_2 increases the strength of the model's agglomeration forces, so it is not surprising to see that the long-run average of welfare in all simulations increases as well. More surprisingly, however, this figure demonstrates that this small increase in the strength of historical spillovers leads to a bifurcation with substantial welfare consequences: minor historical shocks can lead to many alternative long-run spatial

[51]Of course, this does not imply that this model economy, at these parameter values, definitely exhibits a unique steady-state, only that the types of shocks we consider do not appear to be large enough to cause the state variable (the distribution of population) to move from the basin of attraction associated with one steady-state to that of a potential alternative steady-state. Given the size of the model's state space we are unaware of a feasible algorithm that could determine the uniqueness of steady-states in cases (such as ours) where the sufficient condition in Proposition 2 is violated.

[52]Unsurprisingly, exploring the lower half of the CI, where spillovers are weaker, shows no evidence for path dependence.

[53]To facilitate comparisons across values of α_2, simulation b always uses the same values of $s_p^{(b)}$ regardless of the value of α_2 used.

distributions, each associated with very different levels of aggregate long-run welfare.

For example, at our point estimate of $\widehat{\alpha}_2 = -0.041$ we have seen that all 100 alternative perturbations of 1900 fundamentals result in the same steady-state as that of the factual, unperturbed economy. However, increasing α_2 to 0.045—an increase of 4/3 of the standard error of $\alpha_1 + \alpha_2$—leads these 101 alternatives to end up in three different spatial distributions. Another increase of 1/3 of a standard error admits an additional long-run spatial distribution, and a further increase to two standard errors results in eight possible spatial distributions. In each case we see that the difference between the average ln PDV in the best and worst spatial distributions is large. For example, at the largest value of α_2 that we consider, this difference is approximately a factor of two (0.68 log points). Finally, panel (b) of Figure 7 illustrates that analogous results—bifurcation, often with real welfare consequences—obtain for the case of historical amenity spillovers β_2 as well.

4.4 Discussion

The results in this section convey a number of lessons about how we might expect history to matter in a dynamic economic geography model when it is estimated to fit long-run U.S. data. We have seen how merely swapping the productivity fundamentals of similarly-sized locations in 1900—while holding fixed all other exogenous features before, during and after the year 1900—can set in motion a wide range of long-run consequences. Local shocks have large effects on their local economies, and these effects continue to leave their trace on local outcomes over many centuries. Indeed, these effects can be so long-lived that one might conclude (when looking at impacts of shocks on the scale of a few centuries, say) that they are permanent, providing evidence for an economy with multiple steady-states.

As suggested by Proposition 2, whether the particular counterfactual swaps we consider do have genuinely permanent consequences—that is, that the economy exhibits multiple steady-states and that our counterfactual swaps cause the economy to cross from one basin of attraction into another—depends on where one looks within reasonable segments of the parameter space. In particular, we do not see this behavior at the point estimates of our historical spillovers parameters, but do see it at modestly higher values of those parameters (well within our estimated confidence intervals). In this sense, our estimates for the U.S. spatial economy straddle the bifurcation boundary between an economy that displays path-dependent effects from even relatively mundane historical events and one that doesn't.

Given this, one might presume that the welfare consequences of the path dependence we study would be minor—that if our swap counterfactual shocks are barely large enough to reach the basin of a different steady-state then they could hardly be expected to reach one with meaningfully different aggregate welfare properties. Our results firmly reject such a presumption. They are consistent with an economic geography in which small historical events can have substantial consequences, not only for the spatial location of economic activity but for its aggregate efficiency as well.

5 Conclusion

It is not hard to look at the geographic patterns of economic activity around us and believe both that agglomeration forces are at work and that they may even be strong enough to cause a self-reinforcing clustering of economic activity. This opens up the possibility that there are many such spatial configurations in which mobile factors could settle—some good, some bad—as well as the potential for historical accidents, such as initial conditions or long-defunct technological shocks, to play a long-lived or even permanent role in determining the distribution and efficiency of spatial allocations.

This paper has sought to develop a dynamic and forward-looking economic geography framework that can be used to characterize and quantify these possibilities. We have derived conditions on how contemporaneous and historical agglomeration spillovers in production and amenities govern: (i) the existence and uniqueness of equilibria; (ii) the duration of persistence of shocks around a steady-state; and (iii) the scope for multiple steady-states and hence path dependence. A particularly rich region of the model's parameter space—and one that our application to the U.S. from 1800 onwards suggests is very much a possibility—is where equilibria are unique and easy to solve for, persistence lasts many centuries, and minor perturbations in historical conditions can lead the economy towards distinct steady-states with substantial differences in overall efficiency. One implication of this parameter region is that temporary events in many domains may leave large and long-lived geographical traces. The design of place-based policy will also be subtle in the presence of such features.

While we have developed this paper's empirical and theoretical tools with applications to economic geography in mind, they could be applied to other areas in which increasing returns and coordination failures, and hence multiplicity and path dependence, have long appeared as objects of theoretical interest that lack a corresponding amount of high-dimensional quantification and simulation. Applications could include: urban phenomena such as residential segregation, sorting, and "tipping" dynamics (Schelling 1971; Card et al. 2008; and Lee & Lin 2018); traditional "big push" models of development (Rosenstein-Rodan 1943; Murphy et al. 1989; and Krugman & Venables 1995); technology adoption in the presence of network effects and switching costs (David 1985; and Farrell & Klemperer 2007); and dynamic phenomena in political economy such as those surveyed in Acemoglu & Robinson (2005).

References

Acemoglu, D. & Robinson, J. A. (2005), *Economic Origins of Dictatorship and Democracy*, MIT press.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M. & Wolf, N. (2015), 'The economics of density: Evidence from the berlin wall', *Econometrica* **83**(6), 2127–2189.

Allen, T. & Arkolakis, C. (2014), 'Trade and the topography of the spatial economy', *The Quarterly Journal of Economics* **129**(3), 1085–1140.

Allen, T., Arkolakis, C. & Li, X. (2021), On the equilibrium properties of network models with heterogeneous agents. NBER working paper no. 27837.

Anderson, J. E. & Van Wincoop, E. (2003), 'Gravity with gravitas: A solution to the border puzzle', *American Economic Review* **93**(1), 170–192.

Armington, P. S. (1969), 'A theory of demand for products distinguished by place of production', *International Monetary Fund Staff Papers* **16**, 159–178.

Artuç, E., Chaudhuri, S. & McLaren, J. (2010), 'Trade shocks and labor adjustment: A structural empirical approach', *American Economic Review* **100**(3), 1008–45.

Atack, J. (2015), 'Historical geographic information systems (gis) database of steamboat-navigated rivers during the nineteenth century in the united states'.
URL: *https://my.vanderbilt.edu/jeremyatack/data-downloads/*

Atack, J. (2016), 'Historical geographic information systems (gis) database of u.s. railroads for 1850, 1900, and 1911'.
URL: *https://my.vanderbilt.edu/jeremyatack/data-downloads/*

Baldwin, R. E. (1999), 'Agglomeration and endogenous capital', *European Economic Review* **43**(2), 253–280.

Baldwin, R. E. (2001), 'Core-periphery model with forward-looking expectations', *Regional Science and Urban Economics* **31**(1), 21–49.

Baldwin, R. E. & Forslid, R. (2000), 'The core–periphery model and endogenous growth: Stabilizing and destabilizing integration', *Economica* **67**(267), 307–324.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G. & Robert-Nicoud, F. (2011), Economic geography and public policy, *in* 'Economic Geography and Public Policy', Princeton University Press.

Barreca, A., Clay, K., Deschenes, O., Greenstone, M. & Shapiro, J. S. (2016), 'Adapting to climate change: The remarkable decline in the us temperature-mortality relationship over the twentieth century', *Journal of Political Economy* **124**(1), 105–159.

Bazzi, S., Fiszbein, M. & Gebresilasse, M. (2020), 'Frontier culture: The roots and persistence of "rugged individualism" in the united states', *Econometrica* **88**(6), 2329–2368.

Bleakley, H. & Lin, J. (2012), 'Portage and path dependence', *Quarterly Journal of Economics* **127**(2), 587–644.

Bleakley, H. & Lin, J. (2015), 'History and the sizes of cities', *American Economic Review: Papers & Proceedings* **105**(5), 558–563.

Bustos, P., Caprettini, B. & Ponticelli, J. (2016), 'Agricultural productivity and structural transformation: Evidence from brazil', *American Economic Review* **106**(6), 1320–65.

Caliendo, L., Dvorkin, M. & Parro, F. (2019), 'Trade and labor market dynamics: General equilibrium analysis of the china trade shock', *Econometrica* **87**(3), 741–835.

Card, D., Mas, A. & Rothstein, J. (2008), 'Tipping and the dynamics of segregation', *Quarterly Journal of Economics* **133**(1), 177–218.

Chicago Board of Trade (1859), *Annual report of the trade and commerce of Chicago for the year ended December 31, 1858*.

Chicago Board of Trade (1901), *Annual report of the trade and commerce of Chicago for the year ended December 31, 1900*.

Choi, J. & Shim, Y. (2021), 'Technology adoption and late industrialization'.

Crafts, N. & Klein, A. (2014), 'Geography and intra-national home bias: U.s. domestic trade in 1949 and 2007', *Journal of Economic Geography* **15**(3), 477–497.

David, P. A. (1985), 'Clio and the economics of qwerty', *American Economic Review, Papers and Proceedings* **75**(2), 332–337.

Davis, D. & Weinstein, D. (2002), 'Bones, bombs, and break points: The geography of economic activity', *American Economic Review* **92**(5), 1269–1289.

Davis, D. & Weinstein, D. (2008), 'A search for multiple equilibria in urban industrial structure', *Journal of Regional Science* **48**(1), 29–65.

Dekle, R., Eaton, J. & Kortum, S. (2008), 'Global rebalancing with gravity: Measuring the burden of adjustment', *IMF Staff Papers* **55**(3), 511–540.

Dell, M. (2010), 'The persistent effects of peru's mining mita', *Econometrica* **78**(6), 1863–1903.

Dell, M. & Olken, B. (2020), 'The development effects of the extractive colonial economy: The dutch cultivation system in java', *Review of Economic Studies* **87**(1), 164–203.

Deneckere, R. J. & Judd, K. L. (1992), Cyclical and chaotic behavior in a dynamic equilibrium model, with implications for fiscal policy, *in* J. Benhabib, ed., 'Cycles and Chaos in Economic Equilibrium', Princeton University Press.

Desmet, K., Nagy, D. K. & Rossi-Hansberg, E. (2018), 'The geography of development', *Journal of Political Economy* **126**(3), 903–983.

Desmet, K. & Rossi-Hansberg, E. (2014), 'Spatial development', *American Economic Review* **104**(4), 1211–43.

Dippel, C. (2014), 'Forced coexistence and economic development: Evidence from native american reservations', *Econometrica* **82**(6), 2131–2165.

Disdier, A.-C. & Head, K. (2008), 'The puzzling persistence of the distance effect on bilateral trade', *Review of Economics and Statistics* **90**(1), 37–48.

Donaldson, D. & Hornbeck, R. (2016), 'Railroads and american economic growth: A "market access" approach', *Quarterly Journal of Economics* **131**(2), 799–858.

Ellingsen, S. (2021), Long-distance trade and long-term persistence. University of Pompeu Fabra Working Paper.

Farrell, J. & Klemperer, P. (2007), Coordination and lock-in: Competition with switching costs and network effects, *in* M. Armstrong & R. Porter, eds, 'Handbook of Industrial Organization', Vol. 3, Elsevier, pp. 1967–2072.

Feigenbaum, J., Lee, J. & Mezzanotti, F. (2022), 'Capital destruction and economic growth: The effects of sherman's march, 1850-1920', *American Economics Journal: Applied Economics* **14**(4), 301–342.

Fischer, G., Nachtergaele, F., Prieler, S., Van Velthuizen, H., Verelst, L. & Wiberg, D. (2008), 'Global agro-ecological zones assessment for agriculture (gaez 2008)', *IIASA, Laxenburg, Austria and FAO, Rome, Italy* **10**.

Fogel, R. W. (1964), *Railroads and American Economic Growth: Essays in Econometric History*, Baltimore: Johns Hopkins University Press.

Fujita, M., Krugman, P. & Venables, A. J. (1999), *The Spatial Economy: Cities, Regions, and International Trade*, MIT Press.

Fukao, K. & Benabou, R. (1993), 'History versus expectations: a comment', *The Quarterly Journal of Economics* **108**(2), 535–542.

Glaeser, E. L. (2008), *Cities, Agglomeration and Spatial Equilibrium*, Oxford University Press.

Glaeser, E. L. & Gottlieb, J. D. (2009), 'The wealth of cities: Agglomeration economies and spatial equilibrium in the united states', *Journal of economic literature* **47**(4), 983–1028.

Glaeser, E. L., Kerr, S. P. & Kerr, W. R. (2015), 'Entrepreneurship and urban growth: An empirical assessment with historical mines', *Review of Economics and Statistics* **97**(2), 498–520.

Gohberg, I., Lancaster, P. & Rodman, L. (2005), *Matrix polynomials*, Springer.

Gordon, R. J. (2016), *The rise and fall of American growth*, Princeton University Press.

Greenstone, M., Hornbeck, R. & Moretti, E. (2010), 'Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings', *Journal of Political Economy* **118**(3), 536–598.

Hanlon, W. (2017), 'Temporary shocks and persistent effects in urban economies: Evidence from british cities after the u.s. civil war', *Review of Economics and Statistics* **99**(1), 67–79.

Henderson, V., Squires, T., Storeygard, A. & Weil, D. N. (2018), 'The global spatial distribution of economic activity: Nature, history, and the role of trade', *Quarterly Journal of Economics* **133**(1), 357–406.

Herrendorf, B., Valentinyi, A. & Waldmann, R. (2000), 'Ruling out multiplicity and indeterminacy: the role of heterogeneity', *Review of Economic Studies* **67**(2), 295–307.

Hornbeck, R. & Keniston, D. (2017), 'Creative destruction: Barriers to urban growth and the great boston fire of 1872', *American Economic Review* **107**(6), 1365–1398.

Hornbeck, R. & Naidu, S. (2014), 'When the levee breaks: Black migration and economic development in the american south', *American Economic Review* **104**(3), 963–990.

Hsieh, C.-T. & Moretti, E. (2019), 'Housing constraints and spatial misallocation', *American Economic Journal: Macroeconomics* **11**(2), 1–39.

Jaworski, T. & Kitchens, C. T. (2019), 'National policy for regional development: Historical evidence from appalachian highways', *Review of Economics and Statistics* **101**(5), 777–790.

Jedwab, R. & Moradi, A. (2016), 'The permanent effects of transportation revolutions in poor countries: Evidence from africa', *Review of Economics and Statistics* **98**(2), 268–284.

Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M. & Taylor, A. M. (2019), 'The rate of return on everything, 1870–2015', *Quarterly Journal of Economics* **134**(3), 1225–1298.

Kaplan, G. & Schulhofer-Wohl, S. (2017), 'Understanding the long-run decline in interstate migration', *International Economic Review* **58**(1), 57–94.

Kim, S. & Margo, R. A. (2014), Historical perspectives on us economic geography, *in* J. V. Henderson & J.-F. Thisse, eds, 'Handbook of regional and urban economics', Vol. 4, Elsevier, pp. 2981–3019.

Kleinman, B., Liu, E. & Redding, S. J. (2021), Dynamic spatial general equilibrium, Technical report, National Bureau of Economic Research.

Kline, P. & Moretti, E. (2014), 'Local economic development, agglomeration economies and the big push: 100 years of evidence from the tennessee valley authority', *Quarterly Journal of Economics* **129**, 275–331.

Krugman, P. (1980), 'Scale economies, product differentiation, and the pattern of trade', *American Economic Review* **70**(5), 950–959.

Krugman, P. (1991), 'History versus expectations', *Quarterly Journal of Economics* **106**(2), 651–667.

Krugman, P. & Venables, A. J. (1995), 'Globalization and the inequality of nations', *Quarterly Journal of Economics* **110**(4), 857–880.

Landes, D. S. (2003), *The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present*, Cambridge University Press.

Lee, S. & Lin, J. (2018), 'Natural amenities, neighborhood dynamics, and persistence in the spatial distribution of income', *Review of Economic Studies* **85**(1), 663–694.

Manson, S., Schroeder, J., Van Riper, D., Ruggles, S. et al. (2017), 'Ipums national historical geographic information system: Version 12.0 [database]', *Minneapolis: University of Minnesota* **39**.

Matsuyama, K. (1991), 'Increasing returns, industrialization, and indeterminacy of equilibrium', *Quarterly Journal of Economics* **106**(2), 617–650.

Michaels, G. & Rauch, F. (2018), 'Resetting the urban network: 117-2012', *Economic Journal* **128**(608), 378–412.

Monte, F., Redding, S. J. & Rossi-Hansberg, E. (2018), 'Commuting, migration and local employment elasticities', *American Economic Review* **108**(12), 3855–90.

Murphy, K. M., Shleifer, A. & Vishny, R. W. (1989), 'Industrialization and the big push', *Journal of Political Economy* **97**(5), 1003–1026.

Nagy, D. K. (forthcoming), 'Hinterlands, city formation and growth: evidence from the us westward expansion', *Review of Economic Studies* .

Nunn, N. (2008), Slavery, inequality, and economic development in the americas: An examination of the engerman-sokoloff hypothesis, *in* E. Helpman, ed., 'Institutions and Economic Performance', Harvard University Press, pp. 148–180.

Nunn, N. (2014), Historical development, *in* P. Aghion & S. Durlauf, eds, 'Handbook of Economic Growth', Vol. 2, Elsevier, pp. 347–402.

Ottaviano, G. I. (2001), 'Monopolistic competition, trade, and endogenous spatial fluctuations', *Regional Science and Urban Economics* **31**(1), 51–77.

Ottaviano, G., Tabuchi, T. & Thisse, J.-F. (2002), 'Agglomeration and trade revisited', *International Economic Review* pp. 409–435.

Pellegrina, H. S., Sotelo, S. et al. (2021), Migration, specialization, and trade: Evidence from the brazilian march to the west, Research Seminar in International Economics, Gerald R. Ford School of Public Policy.

Peters, M. (2022), 'Market size and spatial growth—evidence from germany's post-war population expulsions', *Econometrica* **90**(5), 2357–2396.

Puga, D. (1999), 'The rise and fall of regional inequalities', *European Economic Review* **43**(2), 303–334.

Rauch, J. E. (1993), 'Does history matter only when it matters little? the case of city-indu try location', *Quarterly Journal of Economics* **108**(3), 843–867.

Redding, S. J. & Rossi-Hansberg, E. A. (2017), 'Quantitative spatial economics', *Annual Review of Economics* **9**(1).

Redding, S., Sturm, D. & Wolf, N. (2011), 'History and industrial location: Evidence from german airports', *Review of Economics and Statistics* **93**(3), 814–831.

Roback, J. (1982), 'Wages, rents, and the quality of life', *The Journal of Political Economy* pp. 1257–1278.

Robert-Nicoud, F. (2005), 'The structure of simple 'new economic geography' models (or, on identical twins)', *Journal of Economic Geography* **5**(2), 201–234.

Rosen, S. (1979), Wage-based indexes of urban quality of life, *in* P. Mieszkowski & M. Straszheim, eds, 'Current Issues in Urban Economics', Johns Hopkins University Press, pp. 74–104.

Rosenstein-Rodan, P. N. (1943), ‘Problems of industrialisation of eastern and south-eastern europe’, *Economic Journal* **June-September**, 202–11.

Roth, M. (2018), *Magic Bean: The Rise of Soy in America*, University Press of Kansas.

Sanderson, E. & Windmeijer, F. (2016), ‘A weak instrument f-test in linear iv models with multiple endogenous variables’, *Journal of Econometrics* **190**(2), 212–221.

Schelling, T. C. (1971), ‘Dynamic models of segregation’, *Journal of Mathematical Sociology* **1**(July), 143–186.

Sequeira, S., Nunn, N. & Qian, N. (2020), ‘Immigrants and the making of america’, *Review of Economic Studies* **87**(1), 382–419.

Sethian, J. (1996), ‘A fast marching level set method for monotonically advancing fronts’, *Proceedings of the National Academy of Sciences* **93**(4), 1591–1595.

Sims, C. A. (2002), ‘Solving linear rational expectations models’, *Computational economics* **20**(1-2), 1.

Tisseur, F. & Meerbergen, K. (2001), ‘The quadratic eigenvalue problem’, *SIAM review* **43**(2), 235–286.

Tsitsiklis, J. (1995), ‘Efficient algorithms for globally optimal trajectories’, *Automatic Control, IEEE Transactions on* **40**(9), 1528–1538.

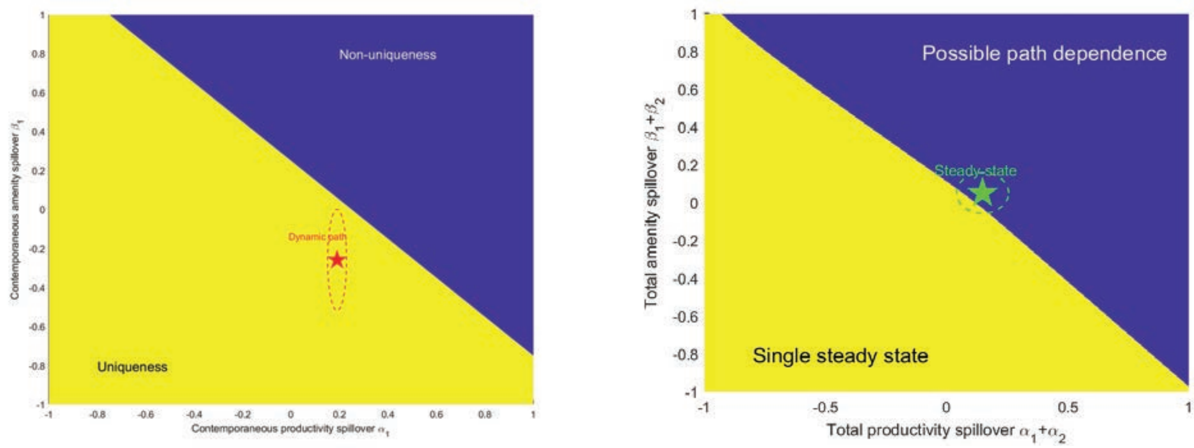
USDA (2003), Historical track records. National Agricultural Statistics Service.

Voth, H.-J. (2021), Persistence–myth and mystery, *in* ‘The handbook of historical economics’, Elsevier, pp. 243–267.

Figure 1: Uniqueness, path dependence, and agglomeration spillover estimates

(a) Uniqueness of the transition path

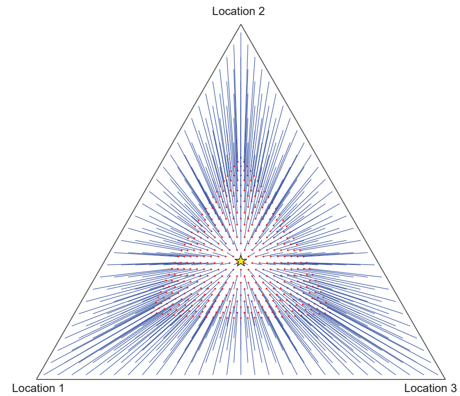
(b) Possibility of path dependence



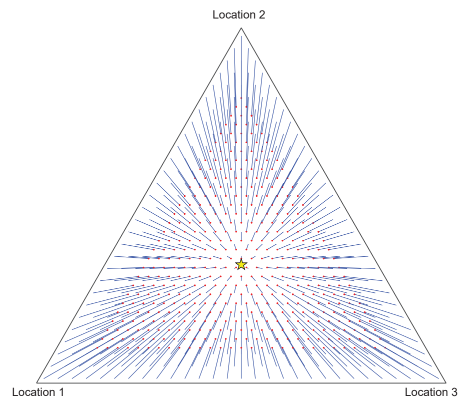
Notes: This figure illustrates what the productivity and amenity spillovers estimated in Section 3.3 imply for the equilibrium properties of the model from Corollary 1 and Proposition 2 (when evaluated at our preferred values of σ, θ, and δ). In panel (a), we show that the estimated pair of contemporaneous productivity and amenity spillovers $\widehat{\alpha}_1$ and $\widehat{\beta}_1$—indicated by the red star—are in the yellow region, which from Corollary 1 implies that the transition path of the economy is unique. In panel (b), we show that the estimated combination of contemporaneous and historical productivity and amenity spillovers $\widehat{\alpha}_1 + \widehat{\alpha}_2$ and $\widehat{\beta}_1 + \widehat{\beta}_2$—indicated by the green star—lies in the blue region, indicating the possibility of path dependence following Proposition 2. In both panels, 95% confidence intervals are shown with dashed lines.

Figure 2: Persistence and path dependence in a three-location economy

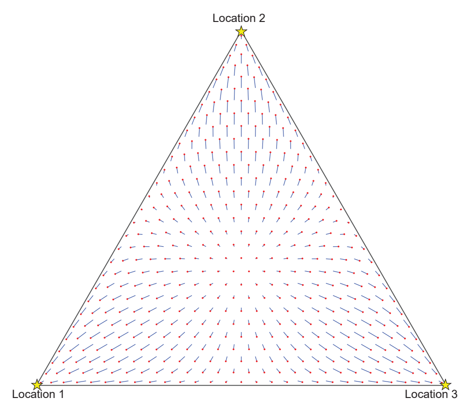
(a) Low persistence ($\alpha_1 = -0.2$, $\alpha_2 = 0$)



(b) High persistence ($\alpha_1 = 0$, $\alpha_2 = 0$)

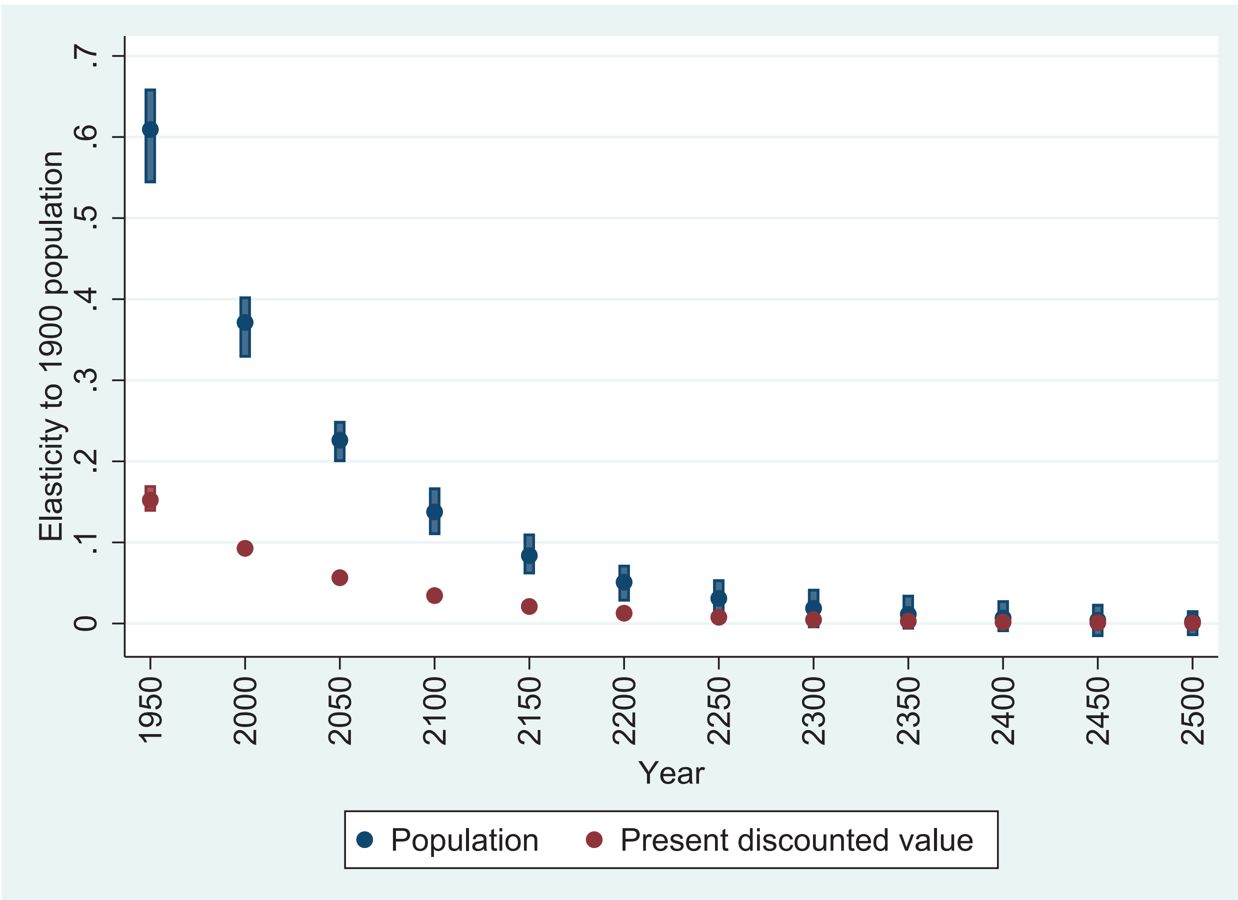


(c) Path dependence ($\alpha_1 = 0$, $\alpha_2 = 0.2$)



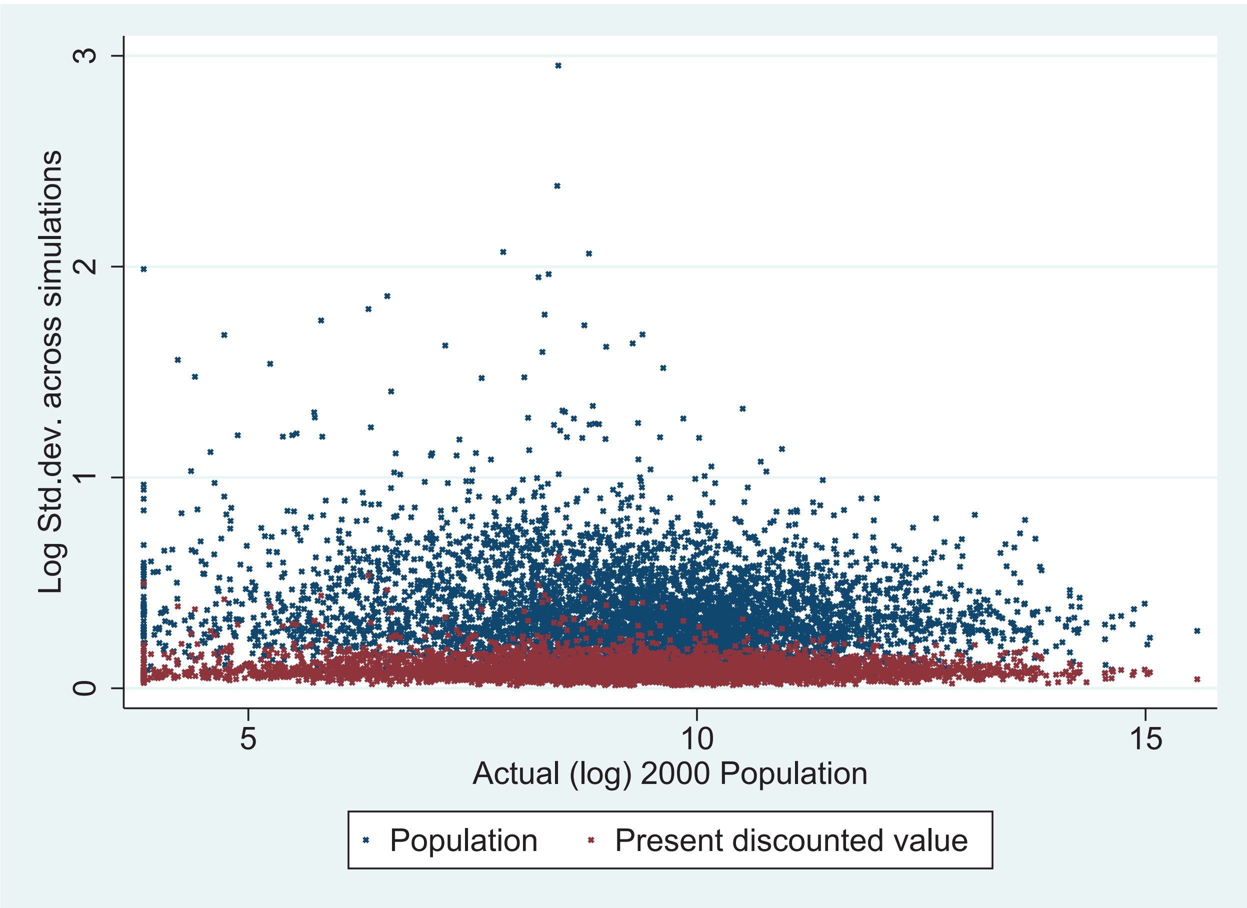
Notes: These figures illustrate phase diagrams for a three-location example economy. Blue arrows (with motion towards the red tips) indicate the change in the unique (following Corollary 1) equilibrium distribution of population from one period to the next, with yellow stars denoting stable steady-states. Panel (a) has relatively low values for both contemporaneous productivity spillovers α_1 and historical spillovers α_2; following Proposition 1, its persistence is therefore relatively low (blue arrows are long). Panel (b) increases the value of α_1, giving rise to longer persistence (shorter arrows). Panel (c) then increases the value of α_2, which, following Proposition 2, results in multiple steady-states (three yellow stars). See Section 2.4 for details.

Figure 3: How persistent are historical shocks?



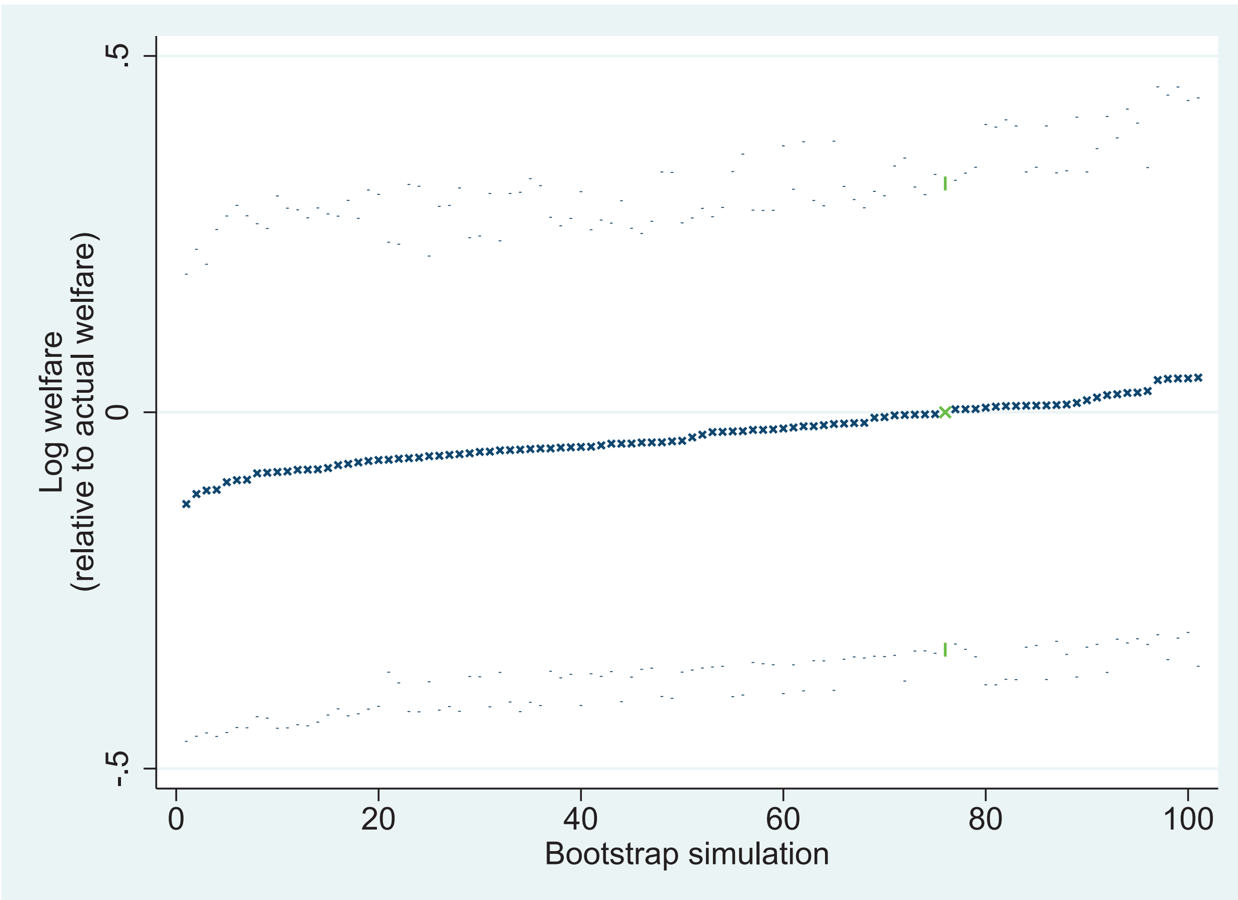
Notes: This figure shows the distribution of estimated elasticities of the local persistence elasticity, $\widehat{\eta}_{it}^{O}$ for two outcomes "O" (population $L_{i,t}$ or present discounted value V_{it}), across all locations i and for each indicated year t. Following equation (28), η_{it}^{O} is obtained from a regression of $\ln O_{it}^{(b)}$ on $L_{i,1900}^{(b)}$ across 100 simulations b, separately by location-year, using the (randomly assigned) value of exogenous productivity $\overline{A}_{i,1900}^{(b)}$ as an IV. Each simulated history randomly shuffles the realized exogenous productivity in the year 1900 between all pairs of locations, where pairs are assigned to locations with the closest 1900 populations. The dots indicate the median estimated elasticity $\widehat{\eta}_{it}^{O}$ across all locations (and the bar indicates the 5-95% range) in a given year, weighting elasticity estimates by the inverse of the square the estimate's standard error.

Figure 4: How resilient are locations to historical shocks?



Notes: This figure plots the standard deviation of log population $\ln L_{i,2000}^{(b)}$ (and log present discounted value $\ln V_{i,2000}^{(b)}$) in the year 2000, across 100 different simulations b of alternative historical conditions, against each location's actual year 2000 log population $\ln L_{i,2000}$. Each simulated history randomly shuffles the realized exogenous productivity $\overline{A}_{i,1900}^{(b)}$ in the year 1900 between all pairs of locations, where pairs are assigned to locations with the closest 1900 populations.

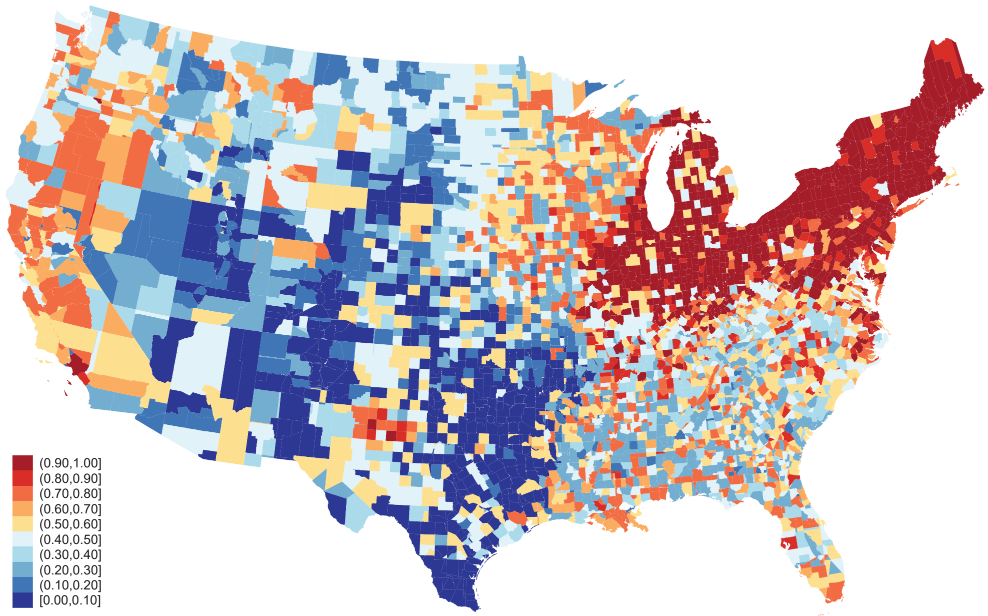
Figure 5: How lucky was our particular history?



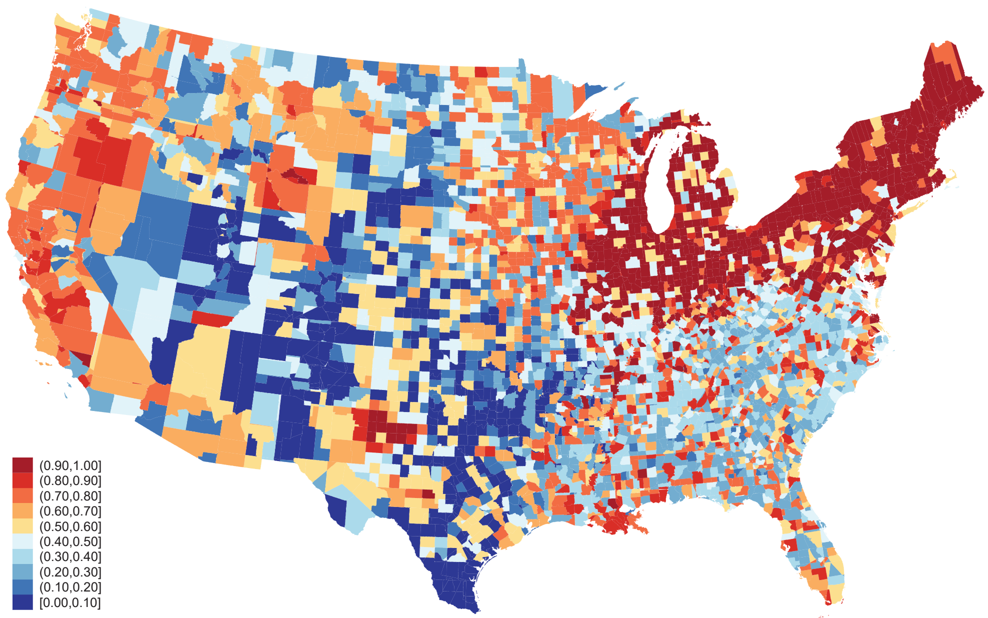
Notes: This figure compares the present discounted value (PDV, i.e. V_{it}) of our factual history to 100 different simulations of alternative historical conditions. Simulations are ordered by their median (log) PDV (indicated with a blue x), where the factual history's PDV (indicated in green) is normalized to zero. The upper and lower quartiles of the PDV across locations are also indicated (with light blue dashes). Each location is weighted by its population so that the PDV reflects the median (and upper/lower quartiles) of each agent in the economy. Each simulated history b randomly shuffles the realized exogenous productivity $\overline{A}_{i,1900}^{(b)}$ in the year 1900 between all pairs of locations, where pairs are assigned to locations with the closest 1900 populations.

Figure 6: How lucky were different locations?

(a) Population



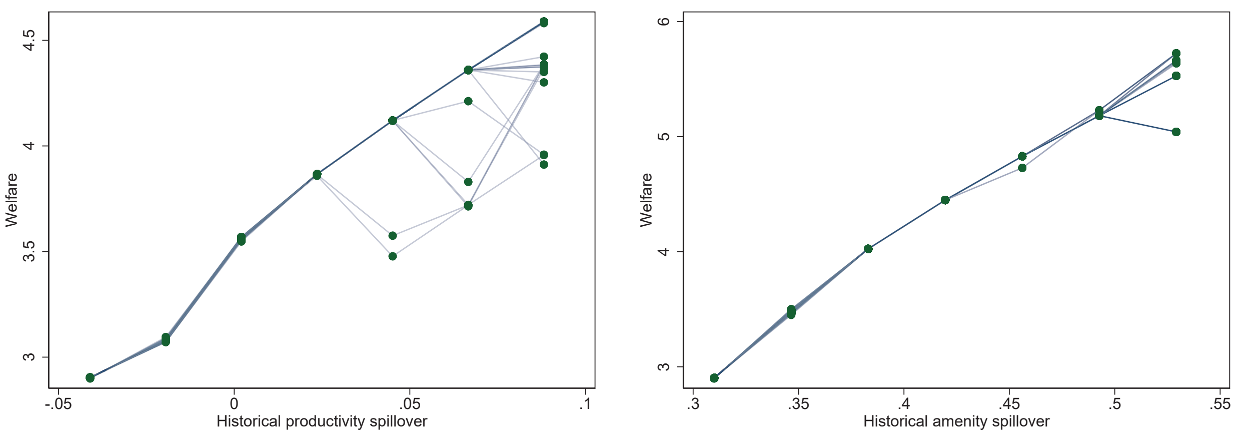
(b) Present discounted value



Notes: This figure compares the actual distribution of economic activity in the year 2000 to 100 different simulations of alternative historical conditions. Each simulated history b randomly shuffles the realized exogenous productivity $\overline{A}_{i,1900}^{(b)}$ in the year 1900 between all pairs of locations, where pairs are assigned to locations with the closest 1900 populations. For each simulated history, we calculate the rank of each location (in terms of its population in panel (a) or PDV in panel (b)) relative to all other locations. The two panels of the figure show the fraction of simulated histories for which each location exceeds that rank in its actual history.

Figure 7: The possibilities of path dependence

(a) Increasing the historical productivity spillover (α_2) (b) Increasing the historical amenity spillover (β_2)



Notes: This figure shows how the population weighted average (log) PDV in the year 3000 for each counterfactual history b, $\ln \bar{V}_{3000}^{(b)}$, changes as we increase the historical productivity spillover α_2 (panel a) or the historical amenity spillover β_2 (panel b). The range considered runs from the point estimate reported in Tables 2 and 3 ($\widehat{\alpha}_2 = -0.04$ and $\widehat{\beta}_2 = 0.31$, respectively) to a value that is larger by twice the value of the standard error reported in Tables 2 and 3 ($SE(\widehat{\alpha}_2) = 0.05$ and $SE(\widehat{\beta}_2) = 0.18$, respectively).

Table 1: Estimated trade and migration costs

	(1)	(2)	(3)	(4)
	1850	1900	1950	2000
		Panel (a): Migration costs		
Effect of travel times on migration flows (κ_t^μ)	0.019*** (0.002)	0.055*** (0.001)	0.073*** (0.002)	0.066*** (0.001)
Source	1850 Census (1% sample)	1900 Census (5% sample)	1950 Census (1% sample)	2000 Census (5% sample)
Observations	8,019	39,641	3,164	16,257
		Panel (b): Trade costs		
Effect of freight costs on trade flows (κ_t^τ)	0.68 (0.428)	0.52 (0.403)	0.27*** (0.013)	0.36*** (0.010)
Source	1858 Chicago Commerce Report	1900 Chicago Commerce Report	1949 Crafts-Klein Rail Flow Data	1997 Commodity Flow Survey
Observations	10	25	1,906	2,123

Notes: Non-linear least squares estimates. For migration costs, we find the value of κ_t^μ so that the sub-county to sub-county migration flows predicted by the model, given the observed population distribution and calculated travel times, most closely matches observed state-to-county flows as revealed in Census data. For trade costs, we find the value of κ_t^τ so that the sub-county to sub-county trade flows predicted by the theory given the observed income distribution and calculated freight costs most closely matches the value of outflows from Chicago by mode of transit and railroad line (for 1850 and 1900), the state-to-state rail shipments (for 1950), and the state-to-state trade flows (for 2000); see the text for details. Bootstrapped standard errors reported in parentheses. Stars indicate statistical significance: * p<.10 ** p<.05 *** p<.01.

Table 2: Estimated productivity spillovers

	(1)	(2)	(3)	(4)	(5)	(6)
Estimated parameters:						
Contemporaneous productivity spillover (α_1)	0.305***	0.334***	0.183***	0.190***	0.143***	0.142***
	(0.039)	(0.045)	(0.035)	(0.040)	(0.034)	(0.039)
Historical productivity spillover (α_2)	-0.002	-0.091*	-0.002	-0.041	-0.002	-0.025
	(0.007)	(0.051)	(0.006)	(0.045)	(0.006)	(0.043)
Estimator	OLS	IV	OLS	IV	OLS	IV
Assumed parameters:						
Elasticity of substitution (σ)	5	5	9	9	13	13
Min SW 1st-stage F-stat		58.1		58.1		58.1
Observations	15,764	15,764	15,764	15,764	15,764	15,764

Notes: OLS and IV estimates of equation (26). Each observation is a sub-county from 1850, 1900, 1950 or 2000. The dependent variable is the (log) value of output per unit labor minus the appropriately (σ-) scaled (log) outward trade market access term. IV regressions use instruments formed from the interaction of linear time trends with two time-invariant geographic variables based on long-run climatic averages: (1) the maximum temperature in the warmest month (and its square); and (2) the minimum temperature in the coldest month (and its square). All regressions additionally control for: (i) the instruments used for the estimation of the amenity spillovers (when estimating equation 27) and (ii) sub-county and region-year (for 14 broad geographic regions, as detailed in the text). The sample is all sub-counties in all years where geographic instruments and contemporaneous/lagged population values are observed, but regressions are weighted by the inverse of the number of sub-counties in the sample in a given year. Standard errors (two-way clustered at the sub-county and county-year levels) are reported in parentheses. Stars indicate statistical significance: * p<.10 ** p<.05 *** p<.01.

Table 3: Estimated amenity spillovers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimated parameters:									
Contemporaneous amenity spillover (β_1)	-0.129	-0.326	-0.392	-0.061	-0.258	-0.324	-0.039	-0.236	-0.301
	(0.266)	(0.264)	(0.263)	(0.267)	(0.265)	(0.265)	(0.267)	(0.266)	(0.266)
Historical amenity spillover (β_2)	0.320*	0.314*	0.313*	0.315*	0.310*	0.308*	0.314*	0.309*	0.307*
	(0.180)	(0.178)	(0.178)	(0.180)	(0.178)	(0.178)	(0.180)	(0.179)	(0.179)
Assumed parameters:									
Elasticity of substitution (σ)	5	5	5	9	9	9	13	13	13
Migration elasticity (θ)	2	4	6	2	4	6	2	4	6
Min SW 1st-stage F-stat	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2	11.2
Observations	15,764	15,764	15,764	15,764	15,764	15,764	15,764	15,764	15,764

Notes: Two-stage least squares estimates of equation (27). Each observation is a sub-county from 1850, 1900,1950 or 2000. The dependent variable is the (log) value of output per unit labor adjusted by the appropriately scaled (log) market access terms: minus inward trade market access (scaled by σ), minus inward migration market access (scaled by θ), and plus outward migration market access (scaled by θ and δ). All regressions use instruments formed from the interaction of linear time trends with two time-invariant geographic variables: (1) the difference between high intensity and low intensity agro-climatic potential yields of maize; and (2) the difference between high intensity agro-climatic potential yields of soy and the low intensity potential yields of wheat. All regressions additionally control for: (i) the instruments used for the estimation of the productivity spillovers (when estimating equation 26) and (ii) sub-county and region-year (for 14 broad geographic regions, as detailed in the text). The sample is all sub-counties in all years where geographic instruments and contemporaneous/lagged population values are observed, but regressions are weighted by the inverse of the number of sub-counties in the sample in a given year. Standard errors (two-way clustered at the sub-county and county-year levels) are reported in parentheses. Stars indicate statistical significance: * p<.10 ** p<.05 *** p<.01.

Table 4: Dynamic model evolution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Panel A: Population (log)				
Predicted population (log)	0.171***	0.166***	0.041***	0.212***	0.114***	-0.022	0.212***
	(0.023)	(0.016)	(0.005)	(0.029)	(0.017)	(0.065)	(0.029)
R-squared (within)	0.112	0.075	0.042	0.178	0.055	0.001	0.178
			Panel B: Output (log)				
Predicted output (log)	0.241***	0.243***	0.078***	0.310***	0.152***	0.078	0.310***
	(0.035)	(0.029)	(0.010)	(0.041)	(0.025)	(0.070)	(0.041)
R-squared (within)	0.128	0.089	0.062	0.215	0.059	0.004	0.215
Observations	19,900	9,952	9,948	9,948	9,952	9,952	9,948
Sample	All	$L_{i,1850}$<median	$L_{i,1850}$>median	Northern	Southern	Western	Eastern

Notes: Ordinary least squares estimates. Each observation is a sub-county-year pair. The dependent variable is the actual (log) population (panel A) or actual (log) output (panel B). The independent variable is the predicted (log) population (top panel) or predicted (log) output (bottom panel) from a model calibrated to exactly match the spatial distribution of economic activity in 1850, and simulated forward holding constant the productivities and amenities in all locations at their 1850 values. All regressions control for year and sub-county fixed effects. In column 1, all sub-counties are included in the sample; columns 2 and 3 split the sample between sub-counties with above and below the median observed population in the year 1850; respectively; columns 4 and 5 split the sample between sub-counties above and below the median latitude, respectively; columns 6 and 7 split the sample between sub-counties above and below the median longitude, respectively. Standard errors two-way clustered at the sub-county and county-year level are reported in parentheses. Stars indicate statistical significance: * p<.10 ** p<.05 *** p<.01.

Persistence and Path Dependence in the Spatial Economy: Online Appendix

Treb Allen and Dave Donaldson

Table of Contents

A Online Appendix: Proofs

A.1 Proof of Theorem 1

This theorem provides a sufficient condition for the uniqueness of a bounded equilibrium solution to equation (18), which we restate here for convenience:

$$x_{i,h,t} = \sum_{j=1}^{N} K_{ij,h,t} \prod_{h'=1}^{H} \left(x_{j,h',t}\right)^{\varepsilon^{\mathtt{j,t}}_{h,h'}} \left(x_{j,h',t+1}\right)^{\varepsilon^{\mathtt{j,t+1}}_{h,h'}} \left(x_{i,h',t+1}\right)^{\varepsilon^{\mathtt{i,t+1}}_{h,h'}} \left(x_{j,h',t-1}\right)^{\varepsilon^{\mathtt{j,t-1}}_{h,h'}} \left(x_{i,h',t-1}\right)^{\varepsilon^{\mathtt{i,t-1}}_{h,h'}}. \tag{A.1}$$

This system of equations holds for all $i \in \{1, ..., N\} \equiv \mathcal{N}$, $h \in \{1, ..., H\} \equiv \mathcal{H}$, and $t \in \{1, ..., \infty\} \equiv \mathcal{T}$. The values of $\{x_{i,h,t}\}_{i,h,t} \in \mathbb{R}^{N\times H}_{++} \times ....$ are unknown, whereas those of $\{K_{ij,h,t}\}_{ij,h,t} \in \mathcal{K} \subseteq \mathbb{R}^{N^2\times H}_{+} \times ....$, $\left\{\varepsilon^{\mathtt{j,t}}_{h,h'}, \varepsilon^{\mathtt{j,t+1}}_{h,h'}, \varepsilon^{\mathtt{i,t+1}}_{h,h'}, \varepsilon^{\mathtt{j,t-1}}_{h,h'}, \varepsilon^{\mathtt{i,t-1}}_{h,h'}\right\}_{h,h'} \in \mathbb{R}^{5(H\times H)}$, and, for some subset $\tilde{\mathcal{H}} \subseteq \mathcal{H}$, the initial conditions $\{x_{i,h,t}\}_{h\in\tilde{\mathcal{H}},i\in\mathcal{N},t=0} \in \mathbb{R}^{N\times\tilde{H}}_{++}$ are finite and given, where $\tilde{H}$ is the dimension of $\tilde{\mathcal{H}}$. Theorem 1 is as follows:

Theorem 1. *Consider the inhomogeneous linear second-order difference equation,*

$$\left(\left|\mathbf{E}^{\mathtt{i,t-1}}\right| + \left|\mathbf{E}^{\mathtt{j,t-1}}\right|\right) \boldsymbol{\mu}_{t-1} - \left(\mathbf{I} - \left|\mathbf{E}^{\mathtt{j,t}}\right|\right) \boldsymbol{\mu}_t + \left(\left|\mathbf{E}^{\mathtt{j,t+1}}\right| + \left|\mathbf{E}^{\mathtt{i,t+1}}\right|\right) \boldsymbol{\mu}_{t+1} = \boldsymbol{b}_t, \tag{A.2}$$

where: the absolute value operator $|\cdot|$ is taken element-wise; $\mathbf{I}$ denotes the $H \times H$ identity matrix; the $H \times H$ matrices $\mathbf{E}^{\mathtt{i,t-1}}$, $\mathbf{E}^{\mathtt{j,t-1}}$, $\mathbf{E}^{\mathtt{j,t}}$, $\mathbf{E}^{\mathtt{j,t+1}}$, and $\mathbf{E}^{\mathtt{i,t+1}}$ are given and correspond to the values defined in equation (18); the sequence $\boldsymbol{b}_t$ is given for all $t \in \mathcal{T}$; the initial conditions $\mu_{\tilde{h},0} = 0$ for all $\tilde{h} \in \tilde{\mathcal{H}}$; and $\boldsymbol{\mu}_t$ is unknown for all $t > 0$ and for all $\tilde{h} \notin \tilde{\mathcal{H}}$ at $t = 0$. Then there is at most one bounded equilibrium solution to equation (18) if the following two conditions hold:

(a) In the case where $\boldsymbol{b}_t = \mathbf{0}$ for all $t \in \mathcal{T}$, the unique solution to (A.2) is $\boldsymbol{\mu}_t = \mathbf{0}$ for all $t \in \mathcal{T}$.

(b) In the case where $\boldsymbol{b}_t \geq \mathbf{0}$ (and where at least one element of the inequality is strict) for all $t \in \mathcal{T}$, there exists no solution to (A.2) of the form $\boldsymbol{\mu}_t \geq \mathbf{0}$ for all $t \in \mathcal{T}$.

Proof. Consider any two bounded solutions that satisfy equation (18), which we will call $\{x_{i,h,t}\}_{i,h,t}$ and $\{y_{i,h,t}\}_{i,h,t}$. Define $z_{i,h,t} \equiv y_{i,h,t}/x_{i,h,t}$. We will show that, under the conditions of the Theorem, $z_{i,h,t} = 1$ for all i, h, t, i.e. $x_{i,h,t} = y_{i,h,t}$.

From equation (18), we have:

$$z_{i,h,t} = \sum_{j=1}^{N} F_{ij,h,t} \prod_{h'=1}^{H} z_{j,h',t}^{\varepsilon^{\mathtt{j,t}}_{h,h'}} z_{j,h',t+1}^{\varepsilon^{\mathtt{j,t+1}}_{h,h'}} z_{i,h',t+1}^{\varepsilon^{\mathtt{i,t+1}}_{h,h'}} z_{j,h',t-1}^{\varepsilon^{\mathtt{j,t-1}}_{h,h'}} z_{i,h',t-1}^{\varepsilon^{\mathtt{i,t-1}}_{h,h'}}, \tag{A.3}$$

where $F_{ij,h,t} \equiv \dfrac{K_{ij,h,t} \prod_{h'=1}^{H} x_{j,h',t}^{\varepsilon^{\mathtt{j,t}}_{h,h'}} x_{j,h',t+1}^{\varepsilon^{\mathtt{j,t+1}}_{h,h'}} x_{i,h',t+1}^{\varepsilon^{\mathtt{i,t+1}}_{h,h'}} x_{j,h',t-1}^{\varepsilon^{\mathtt{j,t-1}}_{h,h'}} x_{i,h',t-1}^{\varepsilon^{\mathtt{i,t-1}}_{h,h'}}}{\sum_{j=1}^{N} K_{ij,h,t} \prod_{h'=1}^{H} x_{j,h',t}^{\varepsilon^{\mathtt{j,t}}_{h,h'}} x_{j,h',t+1}^{\varepsilon^{\mathtt{j,t+1}}_{h,h'}} x_{i,h',t+1}^{\varepsilon^{\mathtt{i,t+1}}_{h,h'}} x_{j,h',t-1}^{\varepsilon^{\mathtt{j,t-1}}_{h,h'}} x_{i,h',t-1}^{\varepsilon^{\mathtt{i,t-1}}_{h,h'}}}$.

Because $\{x_{i,h,t}\}_{i,h,t}$ and $\{y_{i,h,t}\}_{i,h,t}$ are both bounded, $z_{i,h,t}$ is also bounded, i.e. there exists a set of finite scalars $\{m_{h,t}, M_{h,t}\}_{h,t}$ such that we have $m_{h,t} \leq z_{i,h,t} \leq M_{h,t}$ for all i. Define $\tilde{\mu}_{h,t} \equiv M_{h,t}/m_{h,t}$ as the ratio of these bounds and $\mu_{h,t} \equiv \ln \tilde{\mu}_{h,t}$ as its log. Note that by construction $\tilde{\mu}_{h,t} \geq 1$ and $\mu_{h,t} \geq 0$. Using the fact that $\sum_{j=1}^{N} F_{ij,h,t} = 1$, equation (A.3) implies:

$$\tilde{\mu}_{h,t} \leq \prod_{h'=1}^{H} \tilde{\mu}_{h',t}^{\left|\varepsilon^{\mathtt{j,t}}_{h,h'}\right|} \tilde{\mu}_{h',t+1}^{\left|\varepsilon^{\mathtt{j,t+1}}_{h,h'}\right| + \left|\varepsilon^{\mathtt{i,t+1}}_{h,h'}\right|} \tilde{\mu}_{h',t-1}^{\left|\varepsilon^{\mathtt{j,t-1}}_{h,h'}\right| + \left|\varepsilon^{\mathtt{i,t-1}}_{h,h'}\right|},$$

which, by taking logs, can be written in matrix notation as:

$$\boldsymbol{\mu}_t \leq \left(\left|\mathbf{E}^{\mathtt{i,t-1}}\right| + \left|\mathbf{E}^{\mathtt{j,t-1}}\right|\right) \boldsymbol{\mu}_{t-1} + \left|\mathbf{E}^{\mathtt{j,t}}\right| \boldsymbol{\mu}_t + \left(\left|\mathbf{E}^{\mathtt{j,t+1}}\right| + \left|\mathbf{E}^{\mathtt{i,t+1}}\right|\right) \boldsymbol{\mu}_{t+1}, \tag{A.4}$$

which is equivalent to the inhomogeneous linear second-order difference equation (A.2) for some $\boldsymbol{b}_t \geq \mathbf{0}$. By condition (b) of Theorem 1, no $\boldsymbol{b}_t \geq \mathbf{0}$, $\boldsymbol{b}_t \neq \mathbf{0}$ ensures $\boldsymbol{\mu}_t \geq \mathbf{0}$ for all $t \in \mathcal{T}$. Hence, the only admissible solution to equation (A.4) is for the inequality to hold with equality. By condition (a) of Theorem 1, this is the unique solution. Hence $z_{i,h,t} = 1$ for all i, h, and t, as required. □

Remark 1. Conditions (a) and (b) of Theorem 1 can be verified by using existing results from the study of inhomogeneous linear second-order difference equations. Consider the following such equation (which is equivalent to (19) but moved forward a period to respect conventional timing assumptions):

$$|\boldsymbol{A}|\,\boldsymbol{x}_t - (\boldsymbol{I} - |\boldsymbol{B}|)\,\boldsymbol{x}_{t+1} + |\boldsymbol{C}|\,\boldsymbol{x}_{t+2} = \boldsymbol{b}_t, \tag{A.5}$$

Following Tisseur & Meerbergen (2001) and Gohberg et al. (2005), consider the following quadratic eigenvalue problem (QEP). Define the $H \times H$ matrix polynomial $\mathbf{Q}(\lambda)$ as:

$$\mathbf{Q}(\lambda) = \lambda^2\,|\boldsymbol{C}| - \lambda\,(\boldsymbol{I} - |\boldsymbol{B}|) + |\boldsymbol{A}|$$

and let $\Lambda(\mathbf{Q}) \equiv \{\lambda \in \mathbb{C} | \det \mathbf{Q}(\lambda) = 0\}$ be the set of quadratic eigenvalues of $\mathbf{Q}(\lambda)$. When $|\boldsymbol{C}|$ is singular (as will indeed be the case when we apply this Remark to Corollary 1), we have $r < 2H$ eigenvalues, to which we add $2H - r$ infinite eigenvalues. Let the matrices $(\boldsymbol{X}, \boldsymbol{J})$ form the Jordan pair corresponding to $\mathbf{Q}(\lambda)$. That is, $\boldsymbol{J}$ is a $2H \times 2H$ matrix (or, equivalently, an $r \times r$ block matrix) that contains the eigenvalues and their algebraic multiplicities of $\mathbf{Q}(\lambda)$ and $\boldsymbol{X}$ is an $H \times 2H$ matrix containing the corresponding Jordan chains. Furthermore, decompose $(\boldsymbol{X}, \boldsymbol{J})$ into a finite Jordan pair $(\boldsymbol{X}_F, \boldsymbol{J}_F)$ corresponding to the finite eigenvalues and an infinite Jordan pair $(\boldsymbol{X}_\infty, \boldsymbol{J}_\infty)$ corresponding to the infinite eigenvalues, where $\boldsymbol{J}_\infty$ is a Jordan matrix formed of Jordan blocks setting the eigenvalue $\lambda = 0$, i.e. $\boldsymbol{X} = [\boldsymbol{X}_F, \boldsymbol{X}_\infty]$ and $\boldsymbol{J} = \boldsymbol{J}_F \oplus \boldsymbol{J}_\infty$ (where "$\oplus$" denotes the direct sum, i.e. $\boldsymbol{A} \oplus \boldsymbol{B} \equiv \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}$). Finally, define the resolvent matrix $\boldsymbol{Z} = [\boldsymbol{Z}_F, \boldsymbol{Z}_\infty]$ as follows:

$$\begin{pmatrix} \boldsymbol{Z}_F \\ \boldsymbol{Z}_\infty \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{J}_\infty \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_F & \boldsymbol{X}_\infty \\ |\boldsymbol{C}|\,\boldsymbol{X}_F \boldsymbol{J}_F & -|\boldsymbol{A}|\,\boldsymbol{X}_\infty \boldsymbol{J}_\infty - (\boldsymbol{I} - |\boldsymbol{B}|)\,\boldsymbol{X}_\infty \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \boldsymbol{I} \end{pmatrix}.$$

The resolvent along with the Jordan pair forms the Jordan triple. Then, from Theorem 8.3 of Gohberg et al. (2005), the general solution of (19) is:

$$\boldsymbol{\mu}_0 = \boldsymbol{X}_F \boldsymbol{a} - \sum_{t=0}^{v-1} \boldsymbol{X}_\infty \boldsymbol{J}_\infty^t \boldsymbol{Z}_\infty \boldsymbol{b}_t, \tag{A.6}$$

$$\boldsymbol{\mu}_t = \boldsymbol{X}_F \boldsymbol{J}_F^t \boldsymbol{a} - \sum_{\tau=0}^{v-1} \boldsymbol{X}_\infty \boldsymbol{J}_\infty^t \boldsymbol{Z}_\infty \boldsymbol{b}_{t+\tau} + \sum_{\tau=0}^{t-1} \boldsymbol{X}_F \boldsymbol{J}_F^{t-\tau-1} \boldsymbol{Z}_F \boldsymbol{b}_\tau \text{ for } t \geq 1, \tag{A.7}$$

where v is a positive integer such that $\boldsymbol{J}_\infty^v = \boldsymbol{0}$ and $\boldsymbol{a}$ is an arbitrary vector.

The general solution of (19) allows us to derive sufficient conditions that imply conditions (a) and (b) of Theorem 1. If (i) the number of eigenvalues $\lambda \in \Lambda(\mathbf{Q})$ inside the unit circle is equal to $\tilde{H}$, i.e. the number of dimensions of $\left\{x_{i,\tilde{h},0}\right\}$ that is given, then condition (a) of Theorem 1 is satisfied. Moreover, since any bounded solution must be in the span of the eigenvectors associated with these eigenvalues and $\mu_{\tilde{h},0} = 0$ for all $\tilde{h} \in \mathcal{H}$, we have $\boldsymbol{\mu}_0 = \boldsymbol{0}$. If in addition (ii) $\boldsymbol{J}_\infty = \boldsymbol{0}$ and (iii) $\boldsymbol{X}_F \boldsymbol{Z}_F \leq 0$, then from equation A.6, $\boldsymbol{\mu}_0 = \boldsymbol{X}_F \boldsymbol{a}$. Since $\boldsymbol{\mu}_0 = \boldsymbol{0}$, as long as (iv) the Jordan chains corresponding to eigenvalues with moduli less than one are linearly independent, we have $\boldsymbol{a} = \boldsymbol{0}$. Because $\boldsymbol{a} = \boldsymbol{0}$, from equation A.7, $\boldsymbol{\mu}_1 = \boldsymbol{X}_F \boldsymbol{Z}_F \boldsymbol{b}_0$. But because $\boldsymbol{X}_F \boldsymbol{Z}_F \leq 0$, the only $\boldsymbol{b}_0 \geq 0$ that ensure $\boldsymbol{X}_F \boldsymbol{Z}_F \boldsymbol{b}_0 \geq \boldsymbol{0}$ are those such that $\boldsymbol{X}_F \boldsymbol{Z}_F \boldsymbol{b}_0 = \boldsymbol{0}$ and hence $\boldsymbol{b}_0 = \boldsymbol{0}$. Since this implies $\boldsymbol{\mu}_1 = \boldsymbol{0}$, we can use the same argument to show that $\boldsymbol{b}_1 = \boldsymbol{0}$ and, proceeding inductively forward for all $t \in \mathcal{T}$, that condition (b) is satisfied. Hence, conditions (i)-(iv) imply conditions (a) and (b) of Theorem 1.

A.2 Proof of Corollary 1

We first restate the Corollary:

Corollary 1. *Suppose that the matrices of elasticities $\mathbf{E}^{\mathtt{j,t}}$, $\mathbf{E}^{\mathtt{j,t+1}}$, $\mathbf{E}^{\mathtt{i,t+1}}$, $\mathbf{E}^{\mathtt{j,t-1}}$, and $\mathbf{E}^{\mathtt{i,t-1}}$ described in Theorem 1 are as follows:*

$$\mathbf{E}^{\mathtt{j,t}} = \begin{pmatrix} \tilde{\sigma}\left(1+\alpha_1\sigma+\beta_1\left(\sigma-1\right)\right) & \left(1-\sigma\right)\tilde{\sigma} & 0 \\ 0 & \theta & 0 \\ 0 & 0 & -\theta \end{pmatrix} \mathbf{\Gamma}^{-1},$$

$$\mathbf{E}^{\mathtt{j,t+1}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \delta\theta \\ 0 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1},\ \mathbf{E}^{\mathtt{i,t+1}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \delta\theta \end{pmatrix} \mathbf{\Gamma}^{-1},$$

$$\mathbf{E}^{\mathtt{j,t-1}} = \begin{pmatrix} \tilde{\sigma}\left(\alpha_2\sigma+\beta_2\left(\sigma-1\right)\right) & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1},\ \mathbf{E}^{\mathtt{i,t-1}} = \begin{pmatrix} \tilde{\sigma}\left(\alpha_2\left(\sigma-1\right)+\beta_2\sigma\right) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{\Gamma}^{-1},$$

where $\tilde{\sigma} \equiv \left(\sigma-1\right)/\left(2\sigma-1\right)$. If these matrices satisfy conditions (a) and (b) of Theorem 1, then for any initial population $\{L_{i0}\}$ and geography $\{\bar{A}_{it} > 0, \bar{u}_{it} > 0, \tau_{ijt} = \tau_{jit}, \mu_{ijt} > 0\}$, there exists at most one bounded equilibrium in the model described by equations (13)-(17).

Proof. In order to apply Theorem 1 to our case at hand, we proceed in three steps to transform our system of equations (13)–(17) into the form of equation 18.

First, we impose the symmetry of trade costs. When trade costs are symmetric, Allen & Arkolakis (2014) show that the origin and destination fixed effects of the gravity trade equation are equal up to scale. That is if $X_{ijt} = K_{ijt}\gamma_{it}\delta_{jt}$, $K_{ijt} = K_{jit}$, and $\sum_j X_{ijt} = \sum_j X_{jit}$, there exists a $\kappa_t > 0$ such that:[54]

$$\gamma_{it} = \kappa_t \delta_{it}.$$

From equation (4), this implies:

$$\begin{aligned} w_{it}^{1-\sigma} A_{it}^{\sigma-1} &= \kappa_t P_{it}^{\sigma-1} w_{it} L_{it} \iff \\ w_{it} &= \kappa_t^{\frac{1}{1-2\sigma}} W_{it}^{\tilde{\sigma}} \bar{u}_{it}^{-\tilde{\sigma}} \bar{A}_{it}^{\tilde{\sigma}} L_{it}^{\left(\alpha_1-\beta_1+\frac{1}{1-\sigma}\right)\tilde{\sigma}} \left(L_{i,t-1}\right)^{\left(\alpha_2-\beta_2\right)\tilde{\sigma}} \end{aligned}$$

where $\tilde{\sigma} \equiv \frac{\sigma-1}{2\sigma-1}$, and we have used the spillover functions with notation $A_{it} = \bar{A}_{it} L_{it}^{\alpha_1} \left(L_{i,t-1}\right)^{\alpha_2}$ and $u_{it} = \bar{u}_{it} L_{it}^{\beta_1} \left(L_{i,t-1}\right)^{\beta_2}$. As a result, we can combine the first two equilibrium conditions (13) and (14) into the following single condition:

$$\begin{aligned} L_{it}^{\tilde{\sigma}\left(1-\alpha_1(\sigma-1)-\beta_1\sigma\right)} W_{it}^{\tilde{\sigma}\sigma} &= \sum_j K_{ij,1,t} \left(L_{i,t-1}\right)^{\tilde{\sigma}\left(\alpha_2(\sigma-1)+\beta_2\sigma\right)} \left(L_{j,t-1}\right)^{\tilde{\sigma}\left(\alpha_2\sigma+\beta_2(\sigma-1)\right)} \\ &\quad \times L_{jt}^{\tilde{\sigma}\left(1+\alpha_1\sigma+\beta_1(\sigma-1)\right)} W_{jt}^{-(\sigma-1)\tilde{\sigma}}, \end{aligned} \tag{A.8}$$

where $K_{ij,1,t} \equiv \tau_{ijt}^{1-\sigma} \bar{A}_{it}^{(\sigma-1)\tilde{\sigma}} \bar{u}_{it}^{\tilde{\sigma}\sigma} \bar{u}_{jt}^{(\sigma-1)\tilde{\sigma}} \bar{A}_{jt}^{\tilde{\sigma}\sigma}$. We note that given $\{L_{i0}\}_{i\in\mathcal{N}}$ and the initial geography $K_{ij,1,0}$, equation (A.8) uniquely identifies $\{W_{i0}\}_{i\in\mathcal{N}}$. In what follows, we then take $\tilde{H} = 2$, i.e. the initial distribution population and welfare is taken as given.

Second, we use equation (17) to write $V_{it} = W_{it}\Pi_{it+1}^{\delta}$ and substitute this expression into equations (15),

[54]The exact scale (κ) is determined by the aggregate labor market clearing condition. However, the scale can be ignored by first solving for the "scaled" labor (i.e. imposing the scalar is equal to one) and then recovering the scale by imposing the labor market clearing condition. This does not affect any of the other equilibrium equations below, as they are all homogeneous of degree 0 with respect to labor.

(16), and (A.8), yielding:

$$\begin{aligned}
L_{it}^{\tilde{\sigma}(1-\alpha_1(\sigma-1)-\beta_1\sigma)} W_{it}^{\tilde{\sigma}\sigma} &= \sum_j (K_{ij,1,t} \left(L_{i,t-1}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \left(L_{j,t-1}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
&\quad \times L_{jt}^{\tilde{\sigma}(1+\alpha_1\sigma+\beta_1(\sigma-1))} W_{jt}^{-(\sigma-1)\tilde{\sigma}}) \\
\Pi_{it}^{\theta} &= \sum_j K_{ij,2,t} W_{jt}^{\theta} \Pi_{jt+1}^{\delta\theta} \\
L_{it} W_{it}^{-\theta} &= \sum_j K_{ij,3,t} L_{jt-1} \Pi_{jt}^{-\theta} \Pi_{it+1}^{\delta\theta},
\end{aligned}$$

where $K_{ij,2,t} \equiv \mu_{ij,t}^{-\theta}$ and $K_{ij,3,t} \equiv \mu_{ji,t}^{-\theta}$. The purpose of this step (along with the previous step) is to reduce the complexity of the equilibrium system to $3 \times N$ equations with $3 \times N$ unknowns $\{L_{it}, W_{it}, \Pi_{it}\}$ for each time period $t \in \{1, ...\}$.

Third, we impose the following change of variables:

$$\begin{aligned}
x_{i,1,t} &\equiv L_{i,t}^{\tilde{\sigma}(1-\alpha_1(\sigma-1)-\beta_1\sigma)} \left(W_{i,t}\right)^{\tilde{\sigma}\sigma} \\
x_{i,2,t} &\equiv \Pi_{i,t}^{\theta} \\
x_{i,3,t} &\equiv L_{i,t} W_{i,t}^{-\theta}
\end{aligned}$$

where:

$$\begin{pmatrix} \ln x_{1,i,t} \\ \ln x_{2,i,t} \\ \ln x_{3,i,t} \end{pmatrix} = \underbrace{\begin{pmatrix} \tilde{\sigma}\left(1-\alpha_1\left(\sigma-1\right)-\beta_1\sigma\right) & \tilde{\sigma}\sigma & 0 \\ 0 & 0 & \theta \\ 1 & -\theta & 0 \end{pmatrix}}_{\equiv \mathbf{\Gamma}} \begin{pmatrix} \ln L_{it} \\ \ln W_{it} \\ \ln \Pi_{it} \end{pmatrix}$$

provides the one-to-one mapping between $\{x_{i,1,t}, x_{i,2,t}, x_{i,3,t}\}$ and the original endogenous variables $\{L_{i,t}, W_{i,t}, \Pi_{i,t}\}$, which then allows us to re-write the equilibrium system of equations as follows:

$$\begin{aligned}
x_{i,1,t} &= \sum_{j=1}^{N} (K_{ij,1,t} \left(x_{i,1,t-1}^{\gamma_{11}^{-1}} x_{i,2,t-1}^{\gamma_{12}^{-1}} x_{i,3,t-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \left(x_{j,1,t-1}^{\gamma_{11}^{-1}} x_{j,2,t-1}^{\gamma_{12}^{-1}} x_{j,3,t-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
&\quad \times \left(x_{j,1,t}^{\gamma_{11}^{-1}} x_{j,2,t}^{\gamma_{12}^{-1}} x_{j,3,t}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(1+\alpha_1\sigma+\beta_1(\sigma-1))} \left(x_{j,1,t}^{\gamma_{21}^{-1}} x_{j,2,t}^{\gamma_{22}^{-1}} x_{j,3,t}^{\gamma_{23}^{-1}}\right)^{(1-\sigma)\tilde{\sigma}} \qquad (A.9) \\
x_{i,2,t} &= \sum_{j=1}^{N} \mu_{ij,t}^{-\theta} \left(x_{j,1,t}^{\gamma_{21}^{-1}} x_{j,2,t}^{\gamma_{22}^{-1}} x_{j,3,t}^{\gamma_{23}^{-1}}\right)^{\theta} \left(x_{j,1,t+1}^{\gamma_{21}^{-1}} x_{j,2,t+1}^{\gamma_{22}^{-1}} x_{j,3,t+1}^{\gamma_{23}^{-1}}\right)^{\delta\theta} \qquad (A.10) \\
x_{i,3,t} &= \sum_{j=1}^{N} \mu_{ji,t}^{-\theta} \left(x_{j,1,t-1}^{\gamma_{11}^{-1}} x_{j,2,t-1}^{\gamma_{12}^{-1}} x_{j,3,t-1}^{\gamma_{13}^{-1}}\right) \left(x_{j,1,t}^{\gamma_{31}^{-1}} x_{j,2,t}^{\gamma_{32}^{-1}} x_{j,3,t}^{\gamma_{33}^{-1}}\right)^{-\theta} \left(x_{i,1,t+1}^{\gamma_{21}^{-1}} x_{i,2,t+1}^{\gamma_{22}^{-1}} x_{i,3,t+1}^{\gamma_{23}^{-1}}\right)^{\delta\theta}, \qquad (A.11)
\end{aligned}$$

where (with some abuse of notation), we denote the $\{m, n\}$ element of the 3×3 matrix $\mathbf{\Gamma}^{-1}$ as γ_{mn}^{-1}.

As claimed above, the system of equations (A.9)-(A.11) are a special case of equation (18), where:

$$x_{i,h,t} = \sum_{j=1}^{N} K_{ij,h,t} \prod_{h'=1}^{H} x_{j,h',t}^{\varepsilon_{h,h'}^{\mathtt{j,t}}} x_{j,h',t+1}^{\varepsilon_{h,h'}^{\mathtt{j,t+1}}} x_{i,h',t+1}^{\varepsilon_{h,h'}^{\mathtt{i,t+1}}} x_{j,h',t-1}^{\varepsilon_{h,h'}^{\mathtt{j,t-1}}} x_{i,h',t-1}^{\varepsilon_{h,h'}^{\mathtt{i,t-1}}}, \qquad (A.12)$$

where the matrices $\mathbf{E}^{\mathtt{j,t}}$, $\mathbf{E}^{\mathtt{j,t+1}}$, $\mathbf{E}^{\mathtt{i,t+1}}$, $\mathbf{E}^{\mathtt{j,t-1}}$, and $\mathbf{E}^{\mathtt{i,t-1}}$ are defined as above, as required. □

A.3 Proof of Proposition 1

We first restate the proposition:

Proposition 1. *Consider any initial population* $\{L_{i0}\}$ *and time-invariant geography* $\{\bar{A}_i > 0, \bar{u}_i > 0, \tau_{ij} =$

$\tau_{ji}, \mu_{ij} > 0\}$*. Suppose that* $\delta = 0$ *and* $\rho\left(\left|\mathbf{E}^{\mathtt{j,t}}\right|\right) < 1$*, where* $\mathbf{E}^{\mathtt{j,t}}$ *is defined in Corollary 1, so that the dynamic equilibrium is unique. Then the following relationship holds:*

$$\begin{pmatrix} \ln \chi_{L,t} \\ \ln \chi_{V,t} \\ \ln \chi_{\Pi,t} \end{pmatrix} \leq \left|\mathbf{\Gamma}^{-1}\right| \left(\mathbf{I} - \left|\mathbf{E}^{\mathtt{j,t}}\right|\right)^{-1} \mathbf{G} \left|\mathbf{\Gamma}\right| \begin{pmatrix} \ln \chi_{L,t-1} \\ \ln \chi_{V,t-1} \\ \ln \chi_{\Pi,t-1} \end{pmatrix}, \tag{A.13}$$

and $\mathbf{G}$ *is a 3-by-3 matrix whose first two rows are strictly positive (with values that depend on the parameters* $\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma$ *and* θ*, as fully defined below and whose second row consists entirely of zeroes.*

Proof. In the case where $\delta = 0$ so that $V_{it} = W_{it}$, the equilibrium of the dynamic model corresponds to the set of endogenous variables $\{L_{it}, V_{it}, \Pi_{it}\}$ that solve the following system of equations given exogenous parameters $\left\{\bar{A}_{it}, \bar{u}_{it}, \tau_{ij}, \mu_{ij}, L_{it-1}\right\}$.

We have the (combined) trade equation:

$$L_{it}^{\tilde{\sigma}(1-\alpha_1(\sigma-1)-\beta_1)} V_{it}^{\tilde{\sigma}\sigma} = \sum_j F_{ij} L_{jt}^{\tilde{\sigma}(1+(\sigma-1)\beta_1+\alpha_1\sigma)} V_{jt}^{(1-\sigma)\tilde{\sigma}} \left(L_{i,t-1}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \left(L_{j,t-1}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))}, \tag{A.14}$$

the value of living in a particular location (multilateral migration resistance):

$$\Pi_{it}^{\theta} = \sum_j \mu_{ij}^{-\theta} V_{jt}^{\theta}, \tag{A.15}$$

and the population law of motion:

$$L_{it} V_{it}^{-\theta} = \sum_j \mu_{ji}^{-\theta} \Pi_{jt}^{-\theta} L_{jt-1}, \tag{A.16}$$

We take as given the population at time $t = 0$, i.e. $\{L_{i0}\}$. The proof of Proposition 1 proceeds in five steps.

Step #1: Redefine the system We begin by redefining the left-hand side of the equilibrium equations:

$$\begin{aligned} x_{it} &\equiv L_{it}^{\tilde{\sigma}(1-\alpha_1(\sigma-1)-\beta_1)} V_{it}^{\tilde{\sigma}\sigma} \\ y_{it} &\equiv \Pi_{it}^{\theta} \\ z_{it} &\equiv L_{it} V_{it}^{-\theta}, \end{aligned}$$

or equivalently:

$$\begin{pmatrix} \ln x_{it} \\ \ln y_{it} \\ \ln z_{it} \end{pmatrix} = \underbrace{\begin{pmatrix} \tilde{\sigma}\left(1 - \alpha_1\left(\sigma - 1\right) - \beta_1\right) & \tilde{\sigma}\sigma & 0 \\ 0 & 0 & \theta \\ 1 & -\theta & 0 \end{pmatrix}}_{=\mathbf{\Gamma}} \begin{pmatrix} \ln L_{it} \\ \ln V_{it} \\ \ln \Pi_{it} \end{pmatrix} \iff$$

$$\mathbf{\Gamma}^{-1} \begin{pmatrix} \ln x_{it} \\ \ln y_{it} \\ \ln z_{it} \end{pmatrix} = \begin{pmatrix} \ln L_{it} \\ \ln V_{it} \\ \ln \Pi_{it} \end{pmatrix}.$$

With a slight abuse of notation, let γ_{kl}^{-1} denote the $\langle k,l\rangle^{th}$ component of $\mathbf{\Gamma}^{-1}$. We can then re-write the system of equations as:

$$
\begin{aligned}
x_{it} =& \sum_j F_{ij} \left(x_{jt}^{\gamma_{11}^{-1}} y_{jt}^{\gamma_{12}^{-1}} z_{jt}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(1+(\sigma-1)\beta_1+\alpha_1\sigma)} \left(x_{jt}^{\gamma_{21}^{-1}} y_{jt}^{\gamma_{22}^{-1}} z_{jt}^{\gamma_{23}^{-1}}\right)^{(1-\sigma)\tilde{\sigma}} \left(x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
&\times \left(x_{it-1}^{\gamma_{11}^{-1}} y_{it-1}^{\gamma_{12}^{-1}} z_{it-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \\
y_{it} =& \sum_j \mu_{ij}^{-\theta} \left(x_{jt}^{\gamma_{21}^{-1}} y_{jt}^{\gamma_{22}^{-1}} z_{jt}^{\gamma_{23}^{-1}}\right)^{\theta} \\
z_{it} =& \sum_j \mu_{ji}^{-\theta} \left(x_{jt}^{\gamma_{31}^{-1}} y_{jt}^{\gamma_{32}^{-1}} z_{jt}^{\gamma_{33}^{-1}}\right)^{-\theta} \left(x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}}\right)^{1},
\end{aligned}
$$

or, equivalently:

$$
x_{it} = \sum_j F_{ij} x_{jt}^{A_{11}} y_{jt}^{A_{12}} z_{jt}^{A_{13}} \left(x_{it-1}^{\gamma_{11}^{-1}} y_{it-1}^{\gamma_{12}^{-1}} z_{it-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \left(x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \tag{A.17}
$$

$$
y_{it} = \sum_j \mu_{ij}^{-\theta} x_{jt}^{A_{21}} y_{jt}^{A_{22}} z_{jt}^{A_{23}} \tag{A.18}
$$

$$
z_{it} = \sum_j \mu_{ji}^{-\theta} x_{jt}^{A_{31}} y_{jt}^{A_{32}} z_{jt}^{A_{33}} x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}}, \tag{A.19}
$$

where:

$$
\mathbf{A} = \underbrace{\begin{pmatrix} \tilde{\sigma}\left(1+(\sigma-1)\beta_1+\alpha_1\sigma\right) & (1-\sigma)\,\tilde{\sigma} & 0 \\ 0 & \theta & 0 \\ 0 & 0 & -\theta \end{pmatrix}}_{=\mathbf{B}} \mathbf{\Gamma}^{-1}.
$$

Note that $\mathbf{A} = \mathbf{E}^{\mathsf{j},\mathsf{t}}$, as defined in Proposition 1. Equations (A.17)-(A.19) constitute the redefined system.

Step #2: Re-write the system in terms of changes We can further re-write equations (A.17)-(A.19) as:

$$
\begin{aligned}
x_{it} =& \sum_j F_{ij} x_{jt}^{A_{11}} y_{jt}^{A_{12}} z_{jt}^{A_{13}} \left(x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \left(x_{it-1}^{\gamma_{11}^{-1}} y_{it-1}^{\gamma_{12}^{-1}} z_{it-1}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \iff \\
x_{it} =& \sum_j F_{ij} \left(\frac{x_{jt}}{x_{j,t-1}}\right)^{A_{11}} \left(\frac{y_{jt}}{y_{j,t-1}}\right)^{A_{12}} \left(\frac{z_{j,t}}{z_{j,t-1}}\right)^{A_{13}} \left(\left(\frac{x_{j,t-1}}{x_{j,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{j,t-1}}{y_{j,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{j,t-1}}{z_{j,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
&\times \left(\left(\frac{x_{i,t-1}}{x_{i,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{i,t-1}}{y_{i,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{i,t-1}}{z_{i,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \\
&\times x_{j,t-1}^{A_{11}} y_{j,t-1}^{A_{12}} z_{j,t-1}^{A_{13}} \left(x_{jt-2}^{\gamma_{11}^{-1}} y_{jt-2}^{\gamma_{12}^{-1}} z_{jt-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
&\times \left(x_{it-2}^{\gamma_{11}^{-1}} y_{it-2}^{\gamma_{12}^{-1}} z_{it-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)},
\end{aligned} \tag{A.20}
$$

or equivalently,

$$y_{it} = \sum_j \mu_{ij}^{-\theta} x_{jt}^{A_{21}} y_{jt}^{A_{22}} z_{jt}^{A_{23}} \iff$$

$$y_{it} = \sum_j \mu_{ij}^{-\theta} \left(\frac{x_{jt}}{x_{j,t-1}}\right)^{A_{21}} \left(\frac{y_{jt}}{y_{j,t-1}}\right)^{A_{22}} \left(\frac{z_{j,t}}{z_{j,t-1}}\right)^{A_{23}} x_{j,t-1}^{A_{21}} y_{j,t-1}^{A_{22}} z_{j,t-1}^{A_{23}}. \tag{A.21}$$

$$z_{it} = \sum_j \mu_{ji}^{-\theta} x_{jt}^{A_{31}} y_{jt}^{A_{32}} z_{jt}^{A_{33}} x_{jt-1}^{\gamma_{11}^{-1}} y_{jt-1}^{\gamma_{12}^{-1}} z_{jt-1}^{\gamma_{13}^{-1}} \iff$$

$$\begin{aligned} z_{it} = &\sum_j \mu_{ji}^{-\theta} \left(\frac{x_{jt}}{x_{j,t-1}}\right)^{A_{31}} \left(\frac{y_{j,t}}{y_{j,t-1}}\right)^{A_{32}} \left(\frac{z_{j,t}}{z_{j,t-1}}\right)^{A_{33}} \left(\frac{x_{j,t-1}}{x_{j,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{j,t-1}}{y_{j,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{j,t-1}}{z_{j,t-2}}\right)^{\gamma_{13}^{-1}} \\ &\times x_{j,t-1}^{A_{31}} y_{j,t-1}^{A_{32}} z_{j,t-1}^{A_{33}} x_{jt-2}^{\gamma_{11}^{-1}} y_{jt-2}^{\gamma_{12}^{-1}} z_{jt-2}^{\gamma_{13}^{-1}}. \end{aligned} \tag{A.22}$$

Equations (A.20)-A.22 are then the redefined system in changes.

Step #3: Bound the changes Define the following constants:

$$\begin{aligned} M_{x,t} &\equiv \max_j \frac{x_{j,t}}{x_{j,t-1}},\ M_{y,t} \equiv \max_j \frac{y_{j,t}}{y_{j,t-1}},\ M_{z,t} \equiv \max_j \frac{z_{j,t}}{z_{j,t-1}} \\ m_{x,t} &\equiv \min_j \frac{x_{j,t}}{x_{j,t-1}},\ m_{y,t} \equiv \min_j \frac{y_{j,t}}{y_{j,t-1}},\ m_{z,t} \equiv \min_j \frac{z_{j,t}}{z_{j,t-1}} \\ \mu_{x,t} &\equiv \frac{M_{x,t}}{m_{x,t}},\ \mu_{y,t} \equiv \frac{M_{y,t}}{m_{y,t}},\ \mu_{z,t} \equiv \frac{M_{z,t}}{m_{z,t}}. \end{aligned}$$

Let us bound $\{x_{it}\}$ from above first:

$$\begin{aligned} x_{it} = &\sum_j F_{ij} \left(\frac{x_{jt}}{x_{j,t-1}}\right)^{A_{11}} \left(\frac{y_{jt}}{y_{j,t-1}}\right)^{A_{12}} \left(\frac{z_{j,t}}{z_{j,t-1}}\right)^{A_{13}} \left(\left(\frac{x_{j,t-1}}{x_{j,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{j,t-1}}{y_{j,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{j,t-1}}{z_{j,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\ &\times \left(\left(\frac{x_{i,t-1}}{x_{i,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{i,t-1}}{y_{i,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{i,t-1}}{z_{i,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \\ &\times x_{j,t-1}^{A_{11}} y_{j,t-1}^{A_{12}} z_{j,t-1}^{A_{13}} \left(x_{jt-2}^{\gamma_{11}^{-1}} y_{jt-2}^{\gamma_{12}^{-1}} z_{jt-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \left(x_{it-2}^{\gamma_{11}^{-1}} y_{it-2}^{\gamma_{12}^{-1}} z_{it-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \implies \end{aligned}$$

$$\begin{aligned} \frac{x_{it}}{x_{i,t-1}} \leq &\frac{M_{x,t}^{A_{11}\mathbf{1}\{A_{11}\geq 0\}}}{m_{x,t}^{-A_{11}\mathbf{1}\{A_{11}<0\}}} \frac{M_{y,t}^{A_{12}\mathbf{1}\{A_{12}\geq 0\}}}{m_{y,t}^{-A_{12}\mathbf{1}\{A_{12}<0\}}} \frac{M_{z,t}^{A_{13}\mathbf{1}\{A_{13}\geq 0\}}}{m_{z,t}^{-A_{13}\mathbf{1}\{A_{13}<0\}}} \frac{M_{x,t-1}^{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{m_{x,t-1}^{-\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\ &\times \frac{M_{x,t-1}^{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{m_{x,t-1}^{-\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}} \frac{M_{y,t-1}^{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{m_{y,t-1}^{-\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\ &\times \frac{M_{y,t-1}^{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{m_{y,t-1}^{-\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}} \frac{M_{z,t-1}^{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{m_{z,t-1}^{-\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\ &\times \frac{M_{z,t-1}^{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{m_{z,t-1}^{-\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}}. \end{aligned} \tag{A.23}$$

Similarly, we can bound $\{x_{i,t}\}$ from below:

$$
\begin{aligned}
x_{it} = & \sum_j F_{ij} \left(\frac{x_{jt}}{x_{j,t-1}}\right)^{A_{11}} \left(\frac{y_{jt}}{y_{j,t-1}}\right)^{A_{12}} \left(\frac{z_{j,t}}{z_{j,t-1}}\right)^{A_{13}} \left(\left(\frac{x_{j,t-1}}{x_{j,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{j,t-1}}{y_{j,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{j,t-1}}{z_{j,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
& \times \left(\left(\frac{x_{i,t-1}}{x_{i,t-2}}\right)^{\gamma_{11}^{-1}} \left(\frac{y_{i,t-1}}{y_{i,t-2}}\right)^{\gamma_{12}^{-1}} \left(\frac{z_{i,t-1}}{z_{i,t-2}}\right)^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \qquad x_{j,t-1}^{A_{11}} y_{j,t-1}^{A_{12}} z_{j,t-1}^{A_{13}} \left(x_{jt-2}^{\gamma_{11}^{-1}} y_{jt-2}^{\gamma_{12}^{-1}} z_{jt-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))} \\
& \times \left(x_{it-2}^{\gamma_{11}^{-1}} y_{it-2}^{\gamma_{12}^{-1}} z_{it-2}^{\gamma_{13}^{-1}}\right)^{\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)} \qquad \Longrightarrow \\
\frac{x_{it}}{x_{i,t-1}} \geq & \frac{m_{x,t}^{A_{11}\mathbf{1}\{A_{11}\geq 0\}}}{M_{x,t}^{-A_{11}\mathbf{1}\{A_{11}<0\}}} \frac{m_{y,t}^{A_{12}\mathbf{1}\{A_{12}\geq 0\}}}{M_{y,t}^{-A_{12}\mathbf{1}\{A_{12}<0\}}} \frac{m_{z,t}^{A_{13}\mathbf{1}\{A_{13}\geq 0\}}}{M_{z,t}^{-A_{13}\mathbf{1}\{A_{13}<0\}}} \frac{m_{x,t-1}^{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{M_{x,t-1}^{-\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\
& \times \frac{m_{x,t-1}^{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{M_{x,t-1}^{-\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}} \frac{m_{y,t-1}^{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{M_{y,t-1}^{-\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\
& \times \frac{m_{y,t-1}^{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{M_{y,t-1}^{-\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}} \frac{m_{z,t-1}^{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\geq 0\right\}}}{M_{z,t-1}^{-\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))<0\right\}}} \\
& \times \frac{m_{z,t-1}^{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\geq 0\right\}}}{M_{z,t-1}^{-\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\mathbf{1}\left\{\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)<0\right\}}}.
\end{aligned} \tag{A.24}
$$

Combining equations (A.23) and (A.24) (dividing the maximum by the minimum) implies:

$$
\begin{aligned}
\mu_{x,t} \leq & \mu_{x,t}^{|A_{11}|} \mu_{y,t}^{|A_{12}|} \mu_{z,t}^{|A_{13}|} \\
& \times \mu_{x,t-1}^{\left|\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\right|+\left|\gamma_{11}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\right|} \mu_{y,t-1}^{\left|\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\right|+\left|\gamma_{12}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\right|} \\
& \times \mu_{z,t-1}^{\left|\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2\sigma+\beta_2(\sigma-1))\right|+\left|\gamma_{13}^{-1}\tilde{\sigma}(\alpha_2(\sigma-1)+\beta_2\sigma)\right|}.
\end{aligned}
$$

Proceeding similarly for $\{y_{it}\}$ and $\{z_{it}\}$ yield, respectively:

$$
\begin{aligned}
\mu_{y,t} &\leq \mu_{x,t}^{|A_{21}|} \mu_{y,t}^{|A_{22}|} \mu_{z,t}^{|A_{23}|} \\
\mu_{z,t} &\leq \mu_{x,t}^{|A_{31}|} \mu_{y,t}^{|A_{32}|} \mu_{z,t}^{|A_{33}|} \mu_{x,t-1}^{\left|\gamma_{11}^{-1}\right|} \mu_{y,t-1}^{\left|\gamma_{12}^{-1}\right|} \mu_{z,t-1}^{\left|\gamma_{13}^{-1}\right|}.
\end{aligned}
$$

Step #4: Combining the bounds Combining the three inequalities and taking logs yields:

$$
\begin{aligned}
\begin{pmatrix} \ln \mu_{x,t} \\ \ln \mu_{y,t} \\ \ln \mu_{z,t} \end{pmatrix} \quad & \leq \quad \underbrace{\begin{pmatrix} |A_{11}| & |A_{12}| & |A_{13}| \\ |A_{21}| & |A_{22}| & |A_{23}| \\ |A_{31}| & |A_{32}| & |A_{33}| \end{pmatrix}}_{\equiv |\mathbf{A}|} \begin{pmatrix} \ln \mu_{x,t} \\ \ln \mu_{y,t} \\ \ln \mu_{z,t} \end{pmatrix} + \mathbf{G} \begin{pmatrix} \ln \mu_{x,t-1} \\ \ln \mu_{y,t-1} \\ \ln \mu_{z,t-1} \end{pmatrix} \iff \\
(\mathbf{I} - |\mathbf{A}|) \begin{pmatrix} \ln \mu_{x,t} \\ \ln \mu_{y,t} \\ \ln \mu_{z,t} \end{pmatrix} \quad & \leq \quad \mathbf{G} \begin{pmatrix} \ln \mu_{x,t-1} \\ \ln \mu_{y,t-1} \\ \ln \mu_{z,t-1} \end{pmatrix},
\end{aligned} \tag{A.25}
$$

for all $t > 0$, where $\mathbf{G}$ is a 3×3 matrix whose first row has elements $G_{1l} = \left|\gamma_{1l}^{-1}\tilde{\sigma}\left(\alpha_2\sigma + \beta_2\left(\sigma - 1\right)\right)\right| + \left|\gamma_{1l}^{-1}\tilde{\sigma}\left(\alpha_2\left(\sigma - 1\right) + \beta_2\sigma\right)\right|$, whose third row has elements $G_{3l} = \left|\gamma_{1l}^{-1}\right|$, and whose second row is a vector of zeroes.

Because $\rho\left(|\mathbf{A}|\right) < 1$, $(\mathbf{I} - |\mathbf{A}|)$ is an M-matrix and is invertible, which in turn implies that its inverse $(\mathbf{I} - |\mathbf{A}|)^{-1}$ is strictly positive. As a result, we can multiply both sides of equation (A.25) by $(\mathbf{I} - |\mathbf{A}|)^{-1}$

while preserving the inequality, which yields:

$$\begin{pmatrix}\ln\mu_{x,t}\\ \ln\mu_{y,t}\\ \ln\mu_{z,t}\end{pmatrix} \leq (\mathbf{I}-|\mathbf{A}|)^{-1}\,\mathbf{G}\begin{pmatrix}\ln\mu_{x,t-1}\\ \ln\mu_{y,t-1}\\ \ln\mu_{z,t-1}\end{pmatrix}. \tag{A.26}$$

Step #5: Converting the bounds to bounds in $\{L_{it}, V_{it}, \Pi_{it}\}$ space Finally, we convert the bound (A.26) back into $\{L_{it}, V_{it}, \Pi_{it}\}$ space. To do so, recall that:

$$\mathbf{\Gamma}^{-1}\begin{pmatrix}\ln x_{it}\\ \ln y_{it}\\ \ln z_{it}\end{pmatrix} = \begin{pmatrix}\ln L_{it}\\ \ln V_{it}\\ \ln \Pi_{it}\end{pmatrix}$$

so that, for example, we have:

$$\begin{aligned}
&\mu_{L,t} \equiv \frac{\max_i L_{i,t}/L_{i,t-1}}{\min_i L_{i,t}/L_{i,t-1}} \iff\\
&\mu_{L,t} = \frac{\max_i x_{i,t}^{\gamma_{11}^{-1}}\, y_{i,t}^{\gamma_{12}^{-1}}\, z_{i,t}^{\gamma_{13}^{-1}} / x_{i,t-1}^{\gamma_{11}^{-1}} y_{i,t-1}^{\gamma_{12}^{-1}} z_{i,t-1}^{\gamma_{13}^{-1}}}{\min_i x_{i,t}^{\gamma_{11}^{-1}}\, y_{i,t}^{\gamma_{12}^{-1}}\, z_{i,t}^{\gamma_{13}^{-1}} / x_{i,t-1}^{\gamma_{11}^{-1}} y_{i,t-1}^{\gamma_{12}^{-1}} z_{i,t-1}^{\gamma_{13}^{-1}}} \implies\\
&\mu_{L,t} \leq \frac{\max_i\left((x_{i,t}/x_{i,t-1})^{\gamma_{11}^{-1}}\right)\times\max_i\left((y_{i,t}/y_{i,t-1})^{\gamma_{12}^{-1}}\right)\times\max_i\left((z_{i,t}/z_{i,t-1})^{\gamma_{13}^{-1}}\right)}{\min_i\left((x_{i,t}/x_{i,t-1})^{\gamma_{11}^{-1}}\right)\times\min_i\left((y_{i,t}/y_{i,t-1})^{\gamma_{12}^{-1}}\right)\times\min_i\left((z_{i,t}/z_{i,t-1})^{\gamma_{13}^{-1}}\right)} \iff\\
&\mu_{L,t} \leq \left(\frac{\max_i (x_{i,t}/x_{i,t-1})}{\min_i (x_{i,t}/x_{i,t-1})}\right)^{\left|\gamma_{11}^{-1}\right|}\times\left(\frac{\max_i (y_{i,t}/y_{i,t-1})}{\min_i (y_{i,t}/y_{i,t-1})}\right)^{\left|\gamma_{12}^{-1}\right|}\times\left(\frac{\max_i (z_{i,t}/z_{i,t-1})}{\min_i (z_{i,t}/z_{i,t-1})}\right)^{\left|\gamma_{13}^{-1}\right|} \iff\\
&\mu_{L,t} \leq \mu_{x,t}^{\left|\gamma_{11}^{-1}\right|}\mu_{y,t}{}^{\left|\gamma_{12}^{-1}\right|}\mu_{z,t}^{\left|\gamma_{13}^{-1}\right|}.
\end{aligned}$$

Proceeding similarly for $\mu_{V,t}$ and $\mu_{\Pi,t}$ yields:

$$\begin{pmatrix}\ln\mu_{L,t}\\ \ln\mu_{V,t}\\ \ln\mu_{\Pi,t}\end{pmatrix} \leq \left|\mathbf{\Gamma}^{-1}\right|\begin{pmatrix}\ln\mu_{x,t}\\ \ln\mu_{y,t}\\ \ln\mu_{z,t}\end{pmatrix}. \tag{A.27}$$

An identical argument starting with the expression $\begin{pmatrix}\ln x_{it}\\ \ln y_{it}\\ \ln z_{it}\end{pmatrix} = \mathbf{\Gamma}\begin{pmatrix}\ln L_{it}\\ \ln V_{it}\\ \ln \Pi_{it}\end{pmatrix}$ yields:

$$\begin{pmatrix}\ln\mu_{x,t}\\ \ln\mu_{y,t}\\ \ln\mu_{z,t}\end{pmatrix} \leq |\mathbf{\Gamma}|\begin{pmatrix}\ln\mu_{L,t}\\ \ln\mu_{W,t}\\ \ln\mu_{\Pi,t}\end{pmatrix}. \tag{A.28}$$

Substituting bounds (A.27) and (A.28) into bound (A.26) and recalling that $\mathbf{A} = \mathbf{E}^{(j,t)}$ yields:

$$\begin{pmatrix}\ln\mu_{L,t}\\ \ln\mu_{V,t}\\ \ln\mu_{\Pi,t}\end{pmatrix} \leq \left|\mathbf{\Gamma}^{-1}\right|\left(\mathbf{I}-\left|\mathbf{E}^{(j,t)}\right|\right)^{-1}\mathbf{G}\,|\mathbf{\Gamma}|\begin{pmatrix}\ln\mu_{L,t-1}\\ \ln\mu_{V,t-1}\\ \ln\mu_{\Pi,t-1}\end{pmatrix},$$

as required. □

A.4 Proof of Proposition 2

We first restate the proposition:

Proposition 2. *For any time-invariant geography* $\left\{\bar{A}_i > 0, \bar{u}_i > 0, \tau_{ij} = \tau_{ji}, \mu_{ij} = \mu_{ji}\right\}$*, there exists a unique steady-state equilibrium if:*

$$\rho\left(\mathbf{B}\right) < 1,$$

where

$$\mathbf{B} \equiv \begin{pmatrix} \left|\frac{1-\frac{\sigma}{\theta}-\beta_{ss}+\alpha_{ss}\sigma+\beta_{ss}\sigma+\frac{1}{\theta}}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma}\right| & \left|\frac{(1+\delta)(\alpha_{ss}+1)\left(\frac{\sigma-1}{\theta}\right)}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma}\right| \\ \left|\frac{(2\sigma-1)/(\sigma-1)}{\left(\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma\right)}\right| & \left|\frac{1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma-\delta\frac{\sigma}{\theta}}{\frac{\sigma}{\theta}+1-\alpha_{ss}(\sigma-1)-\beta_{ss}\sigma}\right| \end{pmatrix}$$

and $\alpha_{ss} \equiv \alpha_1 + \alpha_2$ *and* $\beta_{ss} \equiv \beta_1 + \beta_2$.

Moreover, if $\rho\left(\mathbf{B}\right) > 1$*, then there exist many geographies for which there are multiple steady-states at each geography.*[55]

Proof. We now characterize the steady state equilibrium. Imposing the symmetry of trade costs (see Proposition 1), in the steady state equations (13)-(17) become:

$$W_i^{\sigma\tilde{\sigma}} L_i^{\tilde{\sigma}(1-(\alpha_1+\alpha_2)(\sigma-1)-\sigma(\beta_1+\beta_2))} = \sum_j \tau_{ij}^{1-\sigma}\left(\bar{A}_i\bar{u}_j\right)^{(\sigma-1)\tilde{\sigma}}\left(\bar{u}_i\bar{A}_j\right)^{\tilde{\sigma}\sigma} W_j^{(1-\sigma)\tilde{\sigma}} L_j^{\tilde{\sigma}(1+(\beta_1+\beta_2)(\sigma-1)+\sigma(\alpha_1+\alpha_2))} \tag{A.29}$$

$$V_i = W_i \Pi_i^{\delta} \tag{A.30}$$

$$\Pi_i^{\theta} \equiv \sum_j \mu_{ij}^{-\theta} V_j^{\theta} \tag{A.31}$$

$$L_i V_i^{-\theta} = \sum_j \mu_{ji}^{-\theta} \Pi_j^{-\theta} L_j. \tag{A.32}$$

Much like the symmetry of trade costs allowed us to simplify the equilibrium spatial distribution of economic activity in each period, the symmetry of migration costs allows us to simplify the steady state relationship between the distribution of labor, the value function of residing in each location, and option value of remaining in that location. In particular, if migration costs are symmetric and we are in the steady-state, we have: $\sum_i L_{ij} = \sum_j L_{ji}$, $L_{ij} = M_{ij} g_i d_j$, and $M_{ij} = M_{ji}$. So then it will be the case that:

$$g_i \propto d_i.$$

In our case, this implies:

$$V_i \Pi_i L_i^{-\frac{1}{\theta}} = \Omega^2,$$

which recall is our measure of steady-state welfare.

This simplifies our system of equations as follows:

$$W_i^{\tilde{\sigma}\sigma} L_i^{\tilde{\sigma}(1-(\alpha_1+\alpha_2)(\sigma-1)-\sigma(\beta_1+\beta_2))} = \sum_j \tau_{ij}^{1-\sigma} \bar{A}_i^{(\sigma-1)\tilde{\sigma}} \bar{u}_i^{\tilde{\sigma}} u_j^{(\sigma-1)\tilde{\sigma}} \bar{A}_j^{\tilde{\sigma}\sigma} W_j^{-(\sigma-1)\tilde{\sigma}} L_j^{\tilde{\sigma}(1+(\alpha_1+\alpha_2)\sigma+(\beta_1+\beta_2)(\sigma-1))}$$

$$L_i^{\frac{1}{1+\delta}} W_i^{-\frac{\theta}{1+\delta}} = \left(\Omega^2\right)^{-\theta\left(\frac{1-\delta}{1+\delta}\right)} \sum_j \mu_{ij}^{-\theta} W_j^{\frac{\theta}{1+\delta}} L_j^{\frac{\delta}{1+\delta}}$$

Let us order the endogenous variables as L, W. Define $\tilde{\alpha} \equiv \alpha_1 + \alpha_2$ and $\tilde{\beta} \equiv \beta_1 + \beta_2$ Then the matrix of LHS coefficients becomes:

$$\mathbf{\Gamma}_{ss} \equiv \begin{pmatrix} \tilde{\sigma}\left(1-\tilde{\alpha}\left(\sigma-1\right)-\tilde{\beta}\sigma\right) & \tilde{\sigma}\sigma \\ \frac{1}{1+\delta} & -\frac{\theta}{1+\delta} \end{pmatrix},$$

[55] If $\rho\left(\mathbf{B}\right) = 1$, there exists at most one steady-state.

and the matrix on the RHS coefficients becomes:

$$\tilde{\mathbf{A}}_{ss} \equiv \begin{pmatrix} \tilde{\sigma}\left(1+\tilde{\alpha}\sigma+\tilde{\beta}\left(\sigma-1\right)\right) & -\left(\sigma-1\right)\tilde{\sigma} \\ \frac{\delta}{1+\delta} & \frac{\theta}{1+\delta} \end{pmatrix}.$$

Hence, we have:

$$\mathbf{A}_{ss} \equiv \tilde{\mathbf{A}}_{ss}\mathbf{\Gamma}_{ss}^{-1}.$$

Defining $\mathbf{B} \equiv \tilde{\mathbf{A}}_{ss}\mathbf{\Gamma}_{ss}^{-1}$, from Theorem 1 of Allen et al. (2021), there exists a unique steady state if:

$$\rho\left(\mathbf{B}\right) < 1,$$

(and at most one steady state if $\rho\left(\mathbf{B}\right) = 1$; see part (ii) of Theorem 1 of Allen et al. (2021) and Remark 2), as required.

The second part of the proposition claims that there exists a geography for which if

$$\rho\left(\mathbf{B}\right) > 1,$$

then there exist multiple equilibria. For readability, we present it this result as a general lemma, under which our model clearly falls:

Lemma 1. *Consider the following mathematical system:*

$$x_{i,1} = \lambda_1 \sum_{j=1}^{N} K_{ij,1} x_{j,1}^{a_{11}} x_{j,2}^{a_{12}} \tag{A.33}$$

$$x_{i,2} = \lambda_2 \sum_{j=1}^{N} K_{ij,2} x_{j,1}^{a_{21}} x_{j,2}^{a_{22}}, \tag{A.34}$$

where $\left\{K_{ij,k}\right\}_{i,j\in\{1,\dots,N\}}^{k\in\{1,2\}}$ are the "kernels" of (exogenous) bilateral frictions, $\left\{a_{lk}\right\}_{l,k\in\{1,2\}}$ are (exogenous) elasticities, $\left\{x_{i,k}\right\}_{i\in\{1,\dots,N\}}^{k\in\{1,2\}}$ are (endogenous) strictly positive vectors and $\left\{\lambda_k\right\}_{k\in\{1,2\}}$ are either endogenous scalars determined by additional constraints or are exogenous. If the spectral radius of the 2×2 matrix $\mathbf{A}^p \equiv \left[\left|a_{kl}\right|\right]$ is greater than one, then there exist kernels $\left\{K_{ij,k}\right\}_{i,j\in\{1,\dots,N\}}^{k\in\{1,2\}}$ such that there are multiple solutions to equations (A.33) and (A.34).

The proof proceeds by construction. We begin by performing two transformations of the problem that simplifies the setup. First, we absorb the scalars into the endogenous variables. To do so, define $y_{i,k} = \left(\lambda_1^{d_{k,1}}\lambda_2^{d_{k,2}}\right) x_{i,k}$, where $\mathbf{D} = \left[d_{kl}\right] \equiv -\left(\mathbf{I}-\mathbf{A}\right)^{-1}$. Note that this is well defined as long as the spectral radius of $\mathbf{A}$ is not equal to one. It is straightforward to then show that the following equations:

$$y_{i,1} = \sum_j K_{ij,1} y_{j,1}^{a_{11}} y_{j,2}^{a_{12}}$$

$$y_{i,2} = \sum_j K_{ij,2} y_{j,1}^{a_{21}} y_{j,2}^{a_{22}}$$

are equivalent to equations (A.33) and (A.34). To see this, substitute in the definition of $y_{i,k}$, yielding:

$$\left(\lambda_1^{d_{11}}\lambda_2^{d_{12}}\right) x_{i,1} = \sum_j K_{ij,1} x_{j,1}^{a_{11}} \left(\lambda_1^{d_{11}}\lambda_2^{d_{12}}\right)^{a_{11}} x_{j,2}^{a_{12}} \left(\lambda_1^{d_{21}}\lambda_2^{d_{22}}\right)^{a_{12}}$$

$$\left(\lambda_1^{d_{21}}\lambda_2^{d_{22}}\right) x_{i,2} = \sum_j K_{ij,2} y_{j,1}^{a_{21}} \left(\lambda_1^{d_{11}}\lambda_2^{d_{12}}\right)^{a_{21}} x_{j,2}^{a_{22}} \left(\lambda_1^{d_{21}}\lambda_2^{d_{22}}\right)^{a_{22}},$$

which, rearranging, yields:

$$x_{i,1} = \lambda_1^{-d_{11}+a_{11}d_{11}+a_{12}d_{21}} \lambda_2^{-d_{12}+a_{11}d_{12}+a_{12}d_{22}} \sum_j K_{ij,1} x_{j,1}^{a_{11}} x_{j,2}^{a_{12}}$$
$$x_{i,2} = \lambda_1^{-d_{21}+a_{21}d_{11}+a_{22}d_{21}} \lambda_2^{-d_{22}+a_{21}d_{12}+a_{22}d_{22}} \sum_j K_{ij,2} x_{j,1}^{a_{21}} x_{j,2}^{a_{22}},$$

which, given the definition of $\mathbf{D}$, is equivalent to equations (A.33) and (A.34) as claimed.[56]

The second transformation is closely related to the "exact hat" algebra pioneered by Dekle et al. (2008) in the field of trade and considers a "normalized" system of equations around an observed equilibrium. Suppose we observe a steady-state solution $\{y_{i,k}\}_{i\in S,k\in\{1,2\}}$ that satisfies:

$$y_{i,1} = \sum_j K_{ij,1} y_{j,1}^{a_{11}} y_{j,2}^{a_{12}}$$
$$y_{i,2} = \sum_j K_{ij,2} y_{j,1}^{a_{21}} y_{j,2}^{a_{22}}.$$

We are interested in knowing whether there exists a different steady-state solution $\{x_{i,k}\}_{i\in S,k\in\{1,2\}}$ that also satisfies the same equations:

$$x_{i,1} = \sum_j K_{ij,1} x_{j,1}^{a_{11}} x_{j,2}^{a_{12}}$$
$$x_{i,2} = \sum_j K_{ij,2} x_{j,1}^{a_{21}} x_{j,2}^{a_{22}}$$

Define $z_{i,k} \equiv \frac{x_{i,k}}{y_{i,k}}$ and note that the previous equations can be written as:

$$z_{i,1} = \sum_j F_{ij,1} z_{j,1}^{a_{11}} z_{j,2}^{a_{12}} \tag{A.35}$$
$$z_{i,2} = \sum_j F_{ij,2} z_{j,1}^{a_{21}} z_{j,2}^{a_{22}}, \tag{A.36}$$

where $F_{ij,k} \equiv \left(\frac{K_{ij,k}}{y_{i,k}} y_{j,1}^{a_{k1}} y_{j,2}^{a_{k2}}\right)$. By construction, note that $z_{i,k} = 1$ is a solution to this system of equations. Moreover, the matrices $\mathbf{F}_k$ are stochastic, i.e.:

$$\sum_j F_{ij,k} = 1 \ \forall i \in \{1, ..., N\} \ k \in \{1, 2\}.$$

In what follows, we will search for stochastic matrices $\mathbf{F}_k$ that have two solutions: one in which $z_{i,k} = 1$ for all $i \in \{1, .., N\}$ and $k \in \{1, 2\}$ and another in which there exists a $z_{i,k} \neq 1$.

It turns out to do this requires $N = 4$. Choose any $m_k < 1 < M_k$ for $k \in \{1, 2\}$. Then we will construct a set of kernels that have the following solution:

$$\begin{pmatrix} z_{1,1} & z_{1,2} \\ z_{2,1} & z_{2,2} \\ z_{3,1} & z_{3,2} \\ z_{4,1} & z_{4,2} \end{pmatrix} = \begin{pmatrix} m_1^{\mathbf{1}\{a_{11}>0\}} M_1^{\mathbf{1}\{a_{11}\leq 0\}}; & m_2^{\mathbf{1}\{a_{12}>0\}} M_2^{\mathbf{1}\{a_{12}\leq 0\}} \\ m_1^{\mathbf{1}\{a_{21}>0\}} M_1^{\mathbf{1}\{a_{21}\leq 0\}}; & m_2^{\mathbf{1}\{a_{22}>0\}} M_2^{\mathbf{1}\{a_{22}\leq 0\}} \\ m_1^{\mathbf{1}\{a_{11}\leq 0\}} M_1^{\mathbf{1}\{a_{11}>0\}}; & m_2^{\mathbf{1}\{a_{12}\leq 0\}} M_2^{\mathbf{1}\{a_{12}>0\}} \\ m_1^{\mathbf{1}\{a_{21}\leq 0\}} M_1^{\mathbf{1}\{a_{21}>0\}}; & m_2^{\mathbf{1}\{a_{22}\leq 0\}} M_2^{\mathbf{1}\{a_{22}>0\}} \end{pmatrix}. \tag{A.37}$$

Before constructing the kernel, we note the following helpful properties.

[56] This follows because $\exp\left(\left(-\mathbf{D} + \mathbf{AD}\right) \ln \boldsymbol{\lambda}\right) = \exp\left(\left(-\left(\mathbf{I} - \mathbf{A}\right)\mathbf{D}\right) \ln \boldsymbol{\lambda}\right) = \lambda$.

First, define $\ln \mathbf{m} \equiv \begin{pmatrix} \ln m_1 \\ \ln m_2 \end{pmatrix}$, $\ln \mathbf{M} \equiv \begin{pmatrix} \ln M_1 \\ \ln M_2 \end{pmatrix}$, and the indicator matrix

$$\mathbf{P} \equiv \begin{pmatrix} \mathbf{1}\{a_{11} > 0\} & \mathbf{1}\{a_{12} > 0\} \\ \mathbf{1}\{a_{21} > 0\} & \mathbf{1}\{a_{22} > 0\} \end{pmatrix}$$

(for "positive"); and $\mathbf{E} \equiv \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Then note that we can bound $\mathbf{m}$ and $\mathbf{M}$ as follows:

$$\begin{aligned}
(\mathbf{A} \circ \mathbf{P}) \ln \mathbf{m} + (\mathbf{A} \circ (\mathbf{E} - \mathbf{P})) \ln \mathbf{M} \leq \ln \mathbf{m} \leq \ln \mathbf{M} \leq (\mathbf{A} \circ (\mathbf{E} - \mathbf{P})) \ln \mathbf{m} + (\mathbf{A} \circ \mathbf{P}) \ln \mathbf{M} &\iff \\
(\mathbf{A} \circ \mathbf{P}) \ln \mathbf{m} + (\mathbf{A} - (\mathbf{A} \circ \mathbf{P})) \ln \mathbf{M} \leq \ln \mathbf{m} \leq \ln \mathbf{M} \leq (\mathbf{A} - (\mathbf{A} \circ \mathbf{P})) \ln \mathbf{m} + (\mathbf{A} \circ \mathbf{P}) \ln \mathbf{M} &\iff \\
\mathbf{A} \ln \mathbf{M} - (\mathbf{A} \circ \mathbf{P}) (\ln \mathbf{M} - \ln \mathbf{m}) \leq \ln \mathbf{m} \leq \ln \mathbf{M} \leq \mathbf{A} \ln \mathbf{m} + (\mathbf{A} \circ \mathbf{P}) (\ln \mathbf{M} - \ln \mathbf{m}) &\iff \\
\ln \mathbf{B} - (\mathbf{A} \circ \mathbf{P}) (\ln \mathbf{M} - \ln \mathbf{m}) \leq \ln \mathbf{m} \leq \ln \mathbf{M} \leq \ln \mathbf{b} + (\mathbf{A} \circ \mathbf{P}) (\ln \mathbf{M} - \ln \mathbf{m}) &\iff \\
\ln \mathbf{B} - \ln \mathbf{D} \leq \ln \mathbf{m} \leq \ln \mathbf{M} \leq \ln \mathbf{b} + \ln \mathbf{D}, &
\end{aligned} \tag{A.38}$$

where $\mathbf{B} = \begin{pmatrix} M_1^{a_{11}} M_2^{a_{12}} \\ M_1^{a_{21}} M_2^{a_{22}} \end{pmatrix}$, $\ln \mathbf{D} \equiv (\mathbf{A} \circ \mathbf{P}) (\ln \mathbf{M} - \ln \mathbf{m}) = \begin{pmatrix} \ln \left(\frac{M_1}{m_1}\right)^{a_{11}\mathbf{1}\{a_{11}>0\}} \left(\frac{M_2}{m_2}\right)^{a_{12}\mathbf{1}\{a_{12}>0\}} \\ \ln \left(\frac{M_1}{m_1}\right)^{a_{21}\mathbf{1}\{a_{21}>0\}} \left(\frac{M_2}{m_2}\right)^{a_{22}\mathbf{1}\{a_{22}>0\}} \end{pmatrix}$ and $D_k \equiv \exp\left((\ln \mathbf{D})_k\right)$. We denote $\underline{m}_k = \frac{B_k}{D_k}, \bar{M}_k = b_k D_k$ as the lower and upper bounds for m_k and M_k (the values $z_{j,k}$ can take) from inequality A.38.

Second, we note the existence and uniqueness of weights that can be used to relate the $z_{j,k}$ ($j \in \{1, 2, 3, 4\}, k \in \{1, 2\}$) variables to other endogenous objects. In what follows, we define those weights for $z_{1.1}$, but the corresponding results also hold for all other j, k. Note that by definition

$$\begin{aligned}
z_{1,1} &= m_1^{\mathbf{1}\{a_{11}>0\}} M_1^{\mathbf{1}\{a_{11}\leq 0\}} \\
&= m_1 \left(\frac{M_1}{m_1}\right)^{\mathbf{1}\{a_{11}\leq 0\}} \\
&= M_1 \left(\frac{m_1}{M_1}\right)^{\mathbf{1}\{a_{11}>0\}}
\end{aligned}$$

and similarly,

$$\begin{aligned}
z_{1,2} &= m_2^{\mathbf{1}\{a_{12}>0\}} M_2^{\mathbf{1}\{a_{12}\leq 0\}} \\
&= m_2 \left(\frac{M_2}{m_2}\right)^{\mathbf{1}\{a_{12}\leq 0\}} \\
&= M_2 \left(\frac{m_2}{M_2}\right)^{\mathbf{1}\{a_{12}>0\}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
z_{1,1}^{a_{11}} z_{1,2}^{a_{12}} &= M_1^{a_{11}} M_2^{a_{12}} \left(\frac{m_1}{M_1}\right)^{a_{11}\mathbf{1}\{a_{11}>0\}} \left(\frac{m_2}{M_2}\right)^{a_{12}\mathbf{1}\{a_{12}>0\}} \\
&= \underline{m}_1
\end{aligned}$$

We can similarly work out the following:

$$\begin{aligned}
z_{2,1}^{a_{11}} z_{2,2}^{a_{12}} &= M_1^{a_{11}} M_2^{a_{12}} \left(\frac{m_1}{M_1}\right)^{a_{11}\mathbf{1}\{a_{21}>0\}} \left(\frac{m_2}{M_2}\right)^{a_{12}\mathbf{1}\{a_{22}>0\}} \\
&= \underline{m}_1 \left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{21}>0\} - \mathbf{1}\{a_{11}\leq 0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{22}>0\} - \mathbf{1}\{a_{12}\leq 0\})}
\end{aligned}$$

$$
\begin{aligned}
z_{3,1}^{a_{11}} z_{3,2}^{a_{12}} &= M_1^{a_{11}} M_2^{a_{12}} \left(\frac{m_1}{M_1}\right)^{a_{11}\mathbf{1}\{a_{11}\leq 0\}} \left(\frac{m_2}{M_2}\right)^{a_{12}\mathbf{1}\{a_{12}\leq 0\}} \\
&= m_1^{a_{11}} m_2^{a_{12}} \left(\frac{M_1}{m_1}\right)^{a_{11}\mathbf{1}\{a_{11}>0\}} \left(\frac{M_1}{m_1}\right)^{a_{12}\mathbf{1}\{a_{12}>0\}} \\
&= \bar{M}_1
\end{aligned}
$$

$$
\begin{aligned}
z_{4,1}^{a_{11}} z_{4,2}^{a_{12}} &= M_1^{a_{11}} M_2^{a_{12}} \left(\frac{m_1}{M_1}\right)^{a_{11}\mathbf{1}\{a_{21}\leq 0\}} \left(\frac{m_2}{M_2}\right)^{q_{12}\mathbf{1}\{a_{22}\leq 0\}} \\
&= \bar{M}_1 \left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{11}>0\}-\mathbf{1}\{a_{21}>0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{12}>0\}-\mathbf{1}\{a_{22}>0\})}
\end{aligned}
$$

From inequality A.38, we note that $\exists P_k, Q_k \in [0,1]$ such that

$$
m_k = P_k \underline{m}_k + (1-P_k)\,\bar{M}_k \tag{A.39}
$$

$$
M_k = Q_k \underline{m}_k + (1-Q_k)\,\bar{M}_k \tag{A.40}
$$

And from the definition of $z_{1.1}$ in A.37:

$$
\begin{aligned}
z_{1,1} &= \mathbf{1}\{a_{11}>0\}\, m_k + \mathbf{1}\{a_{11}\leq 0\}\, M_k \implies \\
&= \mathbf{1}\{a_{11}>0\}\left(P_k \underline{m}_k + (1-P_k)\,\bar{M}_k\right) + \mathbf{1}\{a_{11}\leq 0\}\left(Q_k \underline{m}_k + (1-Q_k)\,\bar{M}_k\right) \\
&= \left(\mathbf{1}\{a_{11}>0\}\, P_k + \mathbf{1}\{a_{11}\leq 0\}\, Q_k\right)\underline{m}_k + \left(\mathbf{1}\{a_{11}>0\}\,(1-P_k) + \mathbf{1}\{a_{11}\leq 0\}\,(1-Q_k)\right)\bar{M}_k \\
&= \omega_k^A \underline{m}_k + \left(1-\omega_k^A\right)\bar{M}_k
\end{aligned}
$$

We can similarly solve for the rest of the kernels of $z_{j,k}$. With all of these properties established, we have enough information to define our kernels:

$$
\mathbf{F}_1 = \begin{pmatrix} \omega_1^A & 0 & 1-\omega_1^A & 0 \\ \omega_1^B & 0 & 1-\omega_1^B & 0 \\ \omega_1^C & 0 & 1-\omega_1^C & 0 \\ \omega_1^D & 0 & 1-\omega_1^D & 0 \end{pmatrix}
$$

$$
\mathbf{F}_2 = \begin{pmatrix} 0 & \omega_2^A & 0 & 1-\omega_2^A \\ 0 & \omega_2^B & 0 & 1-\omega_2^B \\ 0 & \omega_2^C & 0 & 1-\omega_2^C \\ 0 & \omega_2^D & 0 & 1-\omega_2^D \end{pmatrix}.
$$

Note that the $z_{i,k} = 1$ for all $i \in \{1,..,4\}$ and $k \in \{1,2\}$ trivially satisfies the equilibrium system. But it is also straightforward to confirm that the proposed solution (A.37) is also an equilibrium. This is because every equation has a term of $\underline{m}_k$ and $\bar{M}_k$, which we know every endogenous variable is a weighted average of (see equations (A.39) and (A.40)).

Finally, we mention that there are many geographies that deliver this multiplicity for two reasons. First, the argument above holds for any choice of $m_k < 1 < M_k$. Second, it is straightforward to show that perturbations of the above kernel also generate multiple equilibria. Suppose we considered the perturbed system of equations:

$$
\mathbf{F}_1 = \begin{pmatrix} \omega_1^A - \kappa\varepsilon & \delta\varepsilon & 1-\omega_1^A-(1-\kappa)\,\varepsilon & (1-\delta)\,\varepsilon \\ \omega_1^B & 0 & 1-\omega_1^B & 0 \\ \omega_1^C & 0 & 1-\omega_1^C & 0 \\ \omega_1^D & 0 & 1-\omega_1^D & 0 \end{pmatrix},
$$

where $\varepsilon > 0$, $\kappa \in [0,1]$ and $\delta \in [0,1]$. The only restriction we place is that $\omega_1^A - \kappa\varepsilon > 0 \iff \kappa\varepsilon < \omega_1^A$ and $\left(1-\omega_1^A-(1-\kappa)\,\varepsilon\right) > 0 \iff \varepsilon\,(1-\kappa) < 1-\omega_1^A$. Note that both of these equations will hold for sufficiently small ε, as $\omega_k^A = \left(\mathbf{1}\{a_{11}>0\}\, P_k + \mathbf{1}\{a_{11}\leq 0\}\, Q_k\right)$ and $P_k \in [0,1]$ and $Q_k \in [0,1]$. In what follows, we show for any choice of $\varepsilon > 0$ (that is sufficiently small to satisfy these inequalities) and any choice of $\delta \in [0,1]$, there exists a $\kappa \in [0,1]$ that ensures the multiplicity still holds.

Then the relevant equation becomes:

$$
\begin{aligned}
z_{1,1} = &\left(\omega_1^A - \kappa\varepsilon\right)\underline{m}_1 + \delta\epsilon\underline{m}_1 \left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{21}>0\}-\mathbf{1}\{a_{11}\leq 0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{22}>0\}-\mathbf{1}\{a_{12}\leq 0\})} \\
&+ \left(1-\omega_1^A - (1-\kappa)\varepsilon\right)\bar{M}_1 + (1-\delta)\varepsilon\bar{M}_1 \left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{11}>0\}-\mathbf{1}\{a_{21}>0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{12}>0\}-\mathbf{1}\{a_{22}>0\})}
\end{aligned}
$$

Since $z_{1.1} = \omega_1^A\underline{m}_1 + \left(1-\omega_1^A\right)\bar{M}_1$, we have

$$
\begin{aligned}
\kappa\epsilon\underline{m}_1 + (1-\kappa)\epsilon\bar{M}_1 = &\delta\varepsilon\left(\underline{m}_1\left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{21}>0\}-\mathbf{1}\{a_{11}\leq 0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{22}>0\}-\mathbf{1}\{a_{12}\leq 0\})}\right) \\
&+ (1-\delta)\epsilon\bar{M}_1\left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{11}>0\}-\mathbf{1}\{a_{21}>0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{12}>0\}-\mathbf{1}\{a_{22}>0\})} \implies \\
\kappa\underline{m}_1 + (1-\kappa)\bar{M}_1 = &\delta G\underline{m}_1 + (1-\delta)\frac{1}{G}\bar{M}_1
\end{aligned}
$$

where $G \equiv \left(\frac{m_1}{M_1}\right)^{a_{11}(\mathbf{1}\{a_{21}>0\}-\mathbf{1}\{a_{11}\leq 0\})} \left(\frac{m_2}{M_2}\right)^{a_{12}(\mathbf{1}\{a_{22}>0\}-\mathbf{1}\{a_{12}\leq 0\})}$. Note that $G \geq 1$. Recall also that $\underline{m}_1$ and $\bar{M}_1$ are the lowest and highest values that can be achieved given the signs of the exponents. Since $G \geq 1$:

$$\underline{m}_1 \leq G\underline{m}_1, \frac{1}{G}\bar{M}_1 \leq \bar{M}_1,$$

Note also that $G\underline{m}_1 \leq \bar{M}_1$ and $\underline{m}_1 \leq \frac{1}{G}\bar{M}_1$. As a result, there exist constants (weights) $\lambda_1 \in [0,1]$ and $\lambda_2 \in [0,1]$ such that:

$$
\begin{aligned}
G\underline{m}_1 &= \lambda_1\underline{m}_1 + (1-\lambda_1)\bar{M}_1 \\
\frac{1}{G}\bar{M}_1 &= \lambda_2\underline{m}_1 + (1-\lambda_2)\bar{M}_1
\end{aligned}
$$

We now return to the above equation:

$$
\begin{aligned}
\kappa\underline{m}_1 + (1-\kappa)\bar{M}_1 &= \delta G\underline{m}_1 + (1-\delta)\frac{1}{G}\bar{M}_1 \iff \\
\kappa\underline{m}_1 + (1-\kappa)\bar{M}_1 &= \delta\left(\lambda_1\underline{m}_1 + (1-\lambda_1)\bar{M}_1\right) + (1-\delta)\left(\lambda_2\underline{m}_1 + (1-\lambda_2)\bar{M}_1\right) \iff \\
\kappa\underline{m}_1 + (1-\kappa)\bar{M}_1 &= (\delta\lambda_1 + (1-\delta)\lambda_2)\underline{m}_1 + (\delta(1-\lambda_1) + (1-\delta)(1-\lambda_2))\bar{M}_1 \qquad \text{(A.41)}
\end{aligned}
$$

Choose $\kappa \equiv \delta\lambda_1 + (1-\delta)\lambda_2$. Then

$$
\begin{aligned}
1-\kappa &= 1 - \delta\lambda_1 - (1-\delta)\lambda_2 \iff \\
1-\kappa &= 1 + \delta - \delta - \delta\lambda_1 - (1-\delta)\lambda_2 \iff \\
1-\kappa &= \delta(1-\lambda_1) + (1-\delta)(1-\lambda_2),
\end{aligned}
$$

so that equation (A.41) holds. Hence, for any choice of δ, we can find a κ that ensures the equilibrium still holds. Note that there is nothing in this argument that is particular to $z_{1,1}$. As a result, we can construct examples of multiple equilibria of the form:

$$
\mathbf{F}_1 = \begin{pmatrix}
\omega_1^A - \kappa_1^A\varepsilon_1^A; & \delta_1^A\varepsilon_1^A; & 1-\omega_1^A - \left(1-\kappa_1^A\right)\varepsilon_1^A; & \left(1-\delta_1^A\right)\varepsilon_1^A \\
\omega_1^B - \kappa_1^B\varepsilon_1^B; & \delta_1^B\varepsilon_1^B; & 1-\omega_1^B - \left(1-\kappa_1^B\right)\varepsilon_1^B; & \left(1-\delta_1^B\right)\varepsilon_1^B \\
\omega_1^C - \kappa_1^C\varepsilon_1^C; & \delta_1^C\varepsilon_1^C; & 1-\omega_1^C - \left(1-\kappa_1^C\right)\varepsilon_1^C; & \left(1-\delta_1^C\right)\varepsilon_1^C \\
\omega_1^D - \kappa_1^D\varepsilon_1^D; & \delta_1^D\varepsilon_1^D; & 1-\omega_1^D - \left(1-\kappa_1^D\right)\varepsilon_1^D; & \left(1-\delta_1^D\right)\varepsilon_1^D
\end{pmatrix}
$$

$$\mathbf{F}_2 = \begin{pmatrix} \delta_2^A \varepsilon_2^A; & \omega_2^A - \kappa_2^A \varepsilon_2^A; & \left(1-\delta_2^A\right)\varepsilon_2^A & 1-\omega_2^A - \left(1-\kappa_2^A\right)\varepsilon_2^A \\ \delta_2^B \varepsilon_2^B; & \omega_2^B - \kappa_2^B \varepsilon_2^B; & \left(1-\delta_2^B\right)\varepsilon_2^B & 1-\omega_2^B - \left(1-\kappa_2^B\right)\varepsilon_2^B \\ \delta_2^C \varepsilon_2^C; & \omega_2^C - \kappa_2^C \varepsilon_2^C; & \left(1-\delta_2^C\right)\varepsilon_2^C & 1-\omega_2^C - \left(1-\kappa_2^C\right)\varepsilon_2^C \\ \delta_2^D \varepsilon_2^D; & \omega_2^D - \kappa_2^D \varepsilon_2^D; & \left(1-\delta_2^D\right)\varepsilon_2^D & 1-\omega_2^D - \left(1-\kappa_2^D\right)\varepsilon_2^D \end{pmatrix},$$

for many different chosen values of $\left\{\varepsilon_k^l\right\}$ and $\left\{\delta_k^l\right\}$. □

A.5 Proof of Proposition 3

We first restate Proposition 3:

Proposition 3. *Given observed data on* $\{Y_{it}, L_{it}, L_{it-1}\}$ *and given values of* $\left\{\widehat{T}_{ijt}, \widehat{M}_{ijt}\right\}$ *there exists a unique (up to scale) set of values of* $\left\{\mathcal{P}_{it}^{\sigma-1}, P_{it}^{\sigma-1}, \Pi_{it}^{\theta}, \Lambda_{it}^{\theta}\right\}$ *that satisfy equations (22)-(25).*

Proof. Note that the four equations can be considered as two distinct systems of two equations, where the two systems of equations are:

$$\mathcal{P}_{it}^{1-\sigma} = \sum_j \widehat{T}_{ijt} \times Y_{jt} \times (P_{jt}^{1-\sigma})^{-1}$$
$$P_{it}^{1-\sigma} = \sum_j \widehat{T}_{jit} \times Y_{jt} \times (\mathcal{P}_{it}^{1-\sigma})^{-1},$$

and:

$$\left(\Lambda_{it}^{\theta}\right)^{-1} = \sum_j \widehat{M}_{jit} \times L_{jt-1} \times \left(\Pi_{jt}^{\theta}\right)^{-1}$$
$$\Pi_{it}^{\theta} = \sum_j \widehat{M}_{ijt} \times L_{jt} \times \Lambda_{jt}^{\theta}.$$

The first system of equations can be written as:

$$x_i = \sum_j K_{ij}^A y_j^{-1}$$
$$y_i = \sum_j K_{ij}^B x_j^{-1},$$

which has a corresponding LHS matrix of coefficients:

$$\boldsymbol{B} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and the matrix on the RHS coefficients becomes:

$$\boldsymbol{\Gamma} \equiv \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Hence, we have:

$$\mathbf{A} \equiv \boldsymbol{\Gamma}\mathbf{B}^{-1} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

The second system of equations can be written as:

$$x_i^{-1} = \sum_j K_{ij}^A y_j^{-1}$$
$$y_i = \sum_j K_{ij}^B x_j,$$

which has a corresponding LHS matrix of coefficients:

$$\boldsymbol{B} \equiv \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

and the matrix on the RHS coefficients becomes:

$$\mathbf{\Gamma} \equiv \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Hence, we have:

$$\mathbf{A} \equiv \mathbf{\Gamma}\mathbf{B}^{-1} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

In both systems, we have $\mathbf{A}^p = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. It is then straightforward to check that $\rho\left(\mathbf{A}^p\right) = 1$, as required. $\square$

B Online Appendix: Possible Microfoundations for Spillovers

Section 2 briefly discussed several microfoundations for the productivity and amenity spillover functions in equations (1) and (3), respectively. This appendix elaborates.

B.1 Productivity spillovers

We formalize two models—based on the persistence of local knowledge and the durability of investments, in turn—that provide examples of formal microfoundations for the productivity spillover function, $A_{it} = \bar{A}_{it} L_{it}^{\alpha_1} L_{it-1}^{\alpha_2}$.

B.1.1 Microfoundation #1: persistence of local knowledge

We follow Deneckere & Judd (1992). Suppose that firms can pay a fixed cost f_i (in terms of local labor) to create a new variety, over which they have monopoly rights for one period (the period in which they introduce the variety). In the subsequent period, the new variety exists but is produced under conditions of perfect competition. In the following period (two periods after its introduction), we assume the variety no longer exists (i.e. its value to consumers has fully depreciated). Finally, we assume that consumers have Cobb-Douglas preferences (within locations) over the the new varieties and the old varieties, and CES preferences across respectively.

Demand: Let Ω_{it}^{new} be the set of varieties created by monopolistically competitive firms in period t in location i and $\Omega_{i,t}^{old}$ be the set of varieties created in the previous period that are now produced under perfect competition. We assume that consumers have the following preferences:

$$C_{jt} = \left(\sum_i \left(\left(\left(\int_{\Omega_{it}^{new}} q_{ijt}(\omega)^{\frac{\rho-1}{\rho}} d\omega \right)^{\frac{\rho}{\rho-1}} \right)^{\chi} \left(\left(\int_{\Omega_{it}^{old}} q_{ijt}(\omega)^{\frac{\rho-1}{\rho}} d\omega \right)^{\frac{\rho}{\rho-1}} \right)^{1-\chi} \right)^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}},$$

where $q_{ijt}(\omega)$ is the quantity consumed in country j of variety ω from location i. Hence, ρ is the elasticity of substitution between varieties of a given type from a given location, χ is the Cobb-Douglas share of the CES composite of new varieties from a given location, and σ is the elasticity of substitution of the aggregate bundles (of new and old goods) across locations.

Given these preferences, the quantity a consumer in location j in period t will demand from firm ω in location i can be written as:

$$q_{ijt}(\omega) = \begin{cases} \chi p_{ijt}(\omega)^{-\rho} \left(P_{it}^{new}\right)^{\rho-1} \times \frac{\tau_{ijt}^{1-\sigma} \left((P_{it}^{new})^{\chi} \left(P_{it}^{old}\right)^{1-\chi} \right)^{1-\sigma}}{\sum_k \tau_{ijt}^{1-\sigma} \left(\left(P_{kt}^{new}\right)^{\chi} \left(P_{kt}^{old}\right)^{1-\chi} \right)^{1-\sigma}} Y_{jt} & \text{if } \omega \in \Omega_{it}^{new} \\ (1-\chi) p_{ijt}(\omega)^{-\rho} \left(P_{it}^{old}\right)^{\rho-1} \times \frac{\tau_{ijt}^{1-\sigma} \left((P_{it}^{new})^{\chi} \left(P_{it}^{old}\right)^{1-\chi} \right)^{1-\sigma}}{\sum_k \tau_{ijt}^{1-\sigma} \left(\left(P_{kt}^{new}\right)^{\chi} \left(P_{kt}^{old}\right)^{1-\chi} \right)^{1-\sigma}} Y_{jt} & \text{if } \omega \in \Omega_{it}^{old}, \end{cases} \tag{B.1}$$

where:

$$\left(P_{it}^{new}\right)^{1-\rho} \equiv \int_{\Omega_{it}^{new}} p_{ijt}(\omega)^{1-\rho} d\omega \tag{B.2}$$

$$\left(P_{it}^{old}\right)^{1-\rho} \equiv \int_{\Omega_{it}^{old}} p_{ijt}(\omega)^{1-\rho} d\omega \tag{B.3}$$

denote the price indices of the inner CES nests.

Supply: Let $c_{it} \equiv \frac{w_{it}}{\bar{A}_{it}}$ denote the marginal cost of production by a firm, where $\bar{A}_{it}$ is the (exogenous) productivity. The optimization problem faced by firm ω is:

$$\max_{\{q_{ijt}(\omega)\}_j} \sum_j \left(p_{ijt}(\omega) q_{ijt}(\omega) - c_{it}\tau_{ijt} q_{ijt}(\omega)\right) - w_{it} f_{it},$$

subject to consumer demand given by equation (B.1).

As a result, conditional on positive production (of which more below), the first order conditions imply:

$$p_{ijt}(\omega) = \frac{\rho}{\rho - 1} c_{it}\tau_{ijt}, \tag{B.4}$$

so that the price index across new varieties within a location is:

$$P_{it}^{new} \equiv \left(M_{it}^{new}\right)^{\frac{1}{1-\rho}} \left(\frac{\rho}{\rho - 1} c_{it}\right). \tag{B.5}$$

Profits of monopolistically competitive firms: The profits of a firm $\omega \in \Omega_{it}^{new}$ are:

$$\pi_{it}(\omega) \equiv \sum_j \left(p_{ijt}(\omega) - c_{it}\tau_{ijt}\right) q_{ijt}(\omega) - w_{it} f_{it}. \tag{B.6}$$

Substituting the consumer demand expression (B.1) and the price expression (B.4) into equation (B.6) yields:

$$\pi_{it}(\omega) = \chi \frac{1}{\rho} \left(\frac{\rho}{\rho - 1}\right)^{1-\rho} \sum_j \left(c_{it}\tau_{ijt}\right)^{1-\rho} \left(P_{it}^{new}\right)^{\rho - 1} \frac{\tau_{ijt}^{1-\sigma} \left(\left(P_{it}^{new}\right)^{\chi} \left(P_{it}^{old}\right)^{1-\chi}\right)^{1-\sigma}}{\sum_k \tau_{ijt}^{1-\sigma} \left(\left(P_{kt}^{new}\right)^{\chi} \left(P_{kt}^{old}\right)^{1-\chi}\right)^{1-\sigma}} Y_{jt} - w_{it} f_{it}$$

Noting that, from the consumer demand equation (B.1) and the price expression (B.4), the revenue a producer receives is:

$$r_{it}(\omega) \equiv \sum_j p_{ijt}(\omega) q_{ijt}(\omega) \iff$$

$$r_{it}(\omega) \left(\frac{\rho}{\rho - 1}\right)^{\rho - 1} \frac{1}{\chi} = \sum_j \left(c_{it}\tau_{ijt}\right)^{1-\rho} \left(P_{it}^{new}\right)^{\rho - 1} \frac{\tau_{ijt}^{1-\sigma} \left(\left(P_{it}^{new}\right)^{\chi} \left(P_{it}^{old}\right)^{1-\chi}\right)^{1-\sigma}}{\sum_k \tau_{ijt}^{1-\sigma} \left(\left(P_{kt}^{new}\right)^{\chi} \left(P_{kt}^{old}\right)^{1-\chi}\right)^{1-\sigma}} Y_{jt}, \tag{B.7}$$

it is apparent that variable profits are simply equal to revenue divided by the elasticity of substitution:

$$\pi_{it}(\omega) + w_{it} f_{it} = \frac{1}{\rho} r_{it}(\omega). \tag{B.8}$$

Free entry: From the free entry condition, total profits of a firm are zero, i.e. $\pi_{it}(\omega) = 0$. Applying the free entry condition to equation (B.8) yields:

$$w_{it} f_{it} = \frac{1}{\rho} r_{it}(\omega) \tag{B.9}$$

Substituting equation (B.9) into equation (B.7) yields:

$$\sum_j \tau_{ijt}^{1-\rho} w_{it}^{-\rho} A_{it}^{\rho - 1} \left(P_{it}^{new}\right)^{\rho - 1} \frac{\tau_{ijt}^{1-\sigma} \left(\left(P_{it}^{new}\right)^{\chi} \left(P_{it}^{old}\right)^{1-\chi}\right)^{1-\sigma}}{\sum_k \tau_{ijt}^{1-\sigma} \left(\left(P_{kt}^{new}\right)^{\chi} \left(P_{kt}^{old}\right)^{1-\chi}\right)^{1-\sigma}} Y_{jt} = \frac{1}{\chi} \left(\frac{\rho}{\rho - 1}\right)^{\rho - 1} \rho f_{it}, \tag{B.10}$$

where we use the fact that $c_{it} = w_{it}/A_{it}$.

Perfectly competitive varieties: The price charged for the perfectly competitive varieties $\omega \in \Omega_{it}^{new}$ is simply the marginal cost:

$$p_{ijt}(\omega) = \tau_{ijt} c_{it} \ \forall \omega \in \Omega_{it}^{new},$$

so that:

$$P_{it}^{old} = \left(M_{it}^{old}\right)^{\frac{1}{1-\rho}} c_{it}, \tag{B.11}$$

where $M_{it}^{old} \equiv \left|\Omega_{it}^{old}\right|$ denotes the measure of existing varieties.

Labor market clearing: Labor market clearing requires that the total labor used by all firms (for entry and production of the new varieties as well as production of the existing varieties) must equal to the total number of workers in the location, $L_{i,t}$. The total amount of labor required by new varieties is:

$$\begin{aligned} L_{it}^{new} &= \int\limits_{\Omega_{it}^{new}} \left(\sum_j \tau_{ijt} \frac{q_{ijt}(\omega)}{\bar{A}_{it}} + f_i\right) d\omega \iff \\ L_{it}^{new} &= \rho f_{it} M_{it}^{new}, \end{aligned}$$

where $M_{it}^{new} \equiv \left|\Omega_{it}^{new}\right|$ denotes the measure of new varieties and we have used the free entry equation (B.10). Similarly, the total amount of labor required by old varieties is:

$$\begin{aligned} L_{it}^{old} &= \int\limits_{\Omega_{it}^{old}} \left(\sum_j \tau_{ijt} \frac{q_{ijt}(\omega)}{\bar{A}_{it}}\right) d\omega \iff \\ L_{it}^{old} &= M_{it}^{new} \frac{1-\chi}{\chi} \rho f_{it}, \end{aligned}$$

where we have used the equations for the old and new variety price indices from equations (B.5) and (B.11).

Total labor used by all firms is hence:

$$\begin{aligned} L_{it}^{new} + L_{it}^{old} &= L_{it} \iff \\ M_{it}^{new} &= \chi \frac{L_{it}}{\rho f_{it}}, \end{aligned} \tag{B.12}$$

so that the measure of new firms is proportional to the labor supply.

The productivity microfoundation: Combining the old and new variety price indices from equations (B.5) and (B.11) yields:

$$\left(\left(P_{it}^{new}\right)^{\chi} \left(P_{it}^{old}\right)^{1-\chi}\right)^{1-\sigma} = \left(c_{it}\right)^{1-\sigma} \frac{\rho}{\rho-1}^{(1-\sigma)\chi} \left(M_{it}^{new}\right)^{\chi\left(\frac{1-\sigma}{1-\rho}\right)} \left(M_{it}^{old}\right)^{(1-\chi)\left(\frac{1-\sigma}{1-\rho}\right)}.$$

Total trade flows from i to j at time t are determined by simply aggregating across all firms of both types. The total trade of new varieties is thus:

$$\begin{aligned} X_{ijt}^{new} &= \int_{\Omega_{it}^{new}} p_{ijt}(\omega) q_{ijt}(\omega) d\omega \iff \\ X_{ijt}^{new} &= \chi \frac{\left(\tau_{ijt} c_{it}\right)^{1-\sigma} \left(M_{it}^{new}\right)^{\chi\left(\frac{1-\sigma}{1-\rho}\right)} \left(M_{it}^{old}\right)^{(1-\chi)\left(\frac{1-\sigma}{1-\rho}\right)}}{\sum_k \left(\tau_{kjt} c_{kt}\right)^{1-\sigma} \left(M_{kt}^{new}\right)^{\chi\left(\frac{1-\sigma}{1-\rho}\right)} \left(M_{kt}^{old}\right)^{(1-\chi)\left(\frac{1-\sigma}{1-\rho}\right)}} Y_{jt}. \end{aligned}$$

Similarly, the total trade of existing varieties is:

$$X_{ijt}^{old} = \int_{\Omega_{it}^{old}} p_{ijt}(\omega)\, q_{ijt}(\omega)\, d\omega \iff$$
$$X_{ijt}^{old} = (1-\chi)\frac{(\tau_{ijt}c_{it})^{1-\sigma}\left(M_{it}^{new}\right)^{\chi\left(\frac{1-\sigma}{1-\rho}\right)}\left(M_{it}^{old}\right)^{(1-\chi)\left(\frac{1-\sigma}{1-\rho}\right)}}{\sum_k (\tau_{kjt}c_{kt})^{1-\sigma}\left(M_{kt}^{new}\right)^{\chi\left(\frac{1-\sigma}{1-\rho}\right)}\left(M_{kt}^{old}\right)^{(1-\chi)\left(\frac{1-\sigma}{1-\rho}\right)}} Y_{jt}.$$

Hence, total trade flows are:

$$X_{ijt} = X_{ijt}^{new} + X_{ijt}^{old} \iff$$
$$X_{ijt} = \tau_{ijt}^{1-\sigma} w_{it}^{1-\sigma} A_{it}^{\sigma-1} P_{jt}^{\sigma-1} Y_{jt},$$

where:

$$P_{jt}^{1-\sigma} \equiv \sum_k \tau_{kjt}^{1-\sigma} w_{kt}^{1-\sigma} A_{kt}^{\sigma-1}$$

and:

$$A_{it} \equiv \bar{A}_{it} f_{it}^{\frac{1}{\rho-1}} \times L_{it}^{\alpha_1} \times L_{it-1}^{\alpha_2},$$

and $\alpha_1 \equiv \frac{\chi}{\rho-1}$ and $\alpha_2 \equiv \frac{1-\chi}{\rho-1}$, as claimed.

B.1.2 Microfoundation #2: durable investments in local productivity

Setup: In each location i, there is a measure of firms that compete a la Bertrand. Firms can hire workers either to produce or to innovate, where the total quantity produced in location i at time t depends on the amount of labor used in the production L_{it}, the amount of land H_{it}, the amount of innovation ϕ_{it} and some productivity shifter B_{it}:

$$Q_{it} = \phi_{it}^{\gamma_1} B_{it} L_{it}^{\mu} H_{it}^{1-\mu} \iff$$
$$q_{it} = \phi_{it}^{\gamma_1} B_{it} l_{it}^{\mu},$$

where in what follows we focus on the output per unit land q_{it} and the labor per unit land l_{it}. We assume the parameters satisfy $\mu < 1$ (due to the diminishing marginal product of labor per unit land) and $\gamma_1 < 1$ (due to the diminishing marginal product of innovation).

To employ a level of innovation ϕ_{it}, a firm must hire $\nu\phi_{it}^{\xi}$ additional units of labor, where $\xi < \gamma_1/(1-\mu)$. We assume that innovation today has an affect on the level of productivity tomorrow so that:

$$B_{it} = \phi_{it-1}^{\widetilde{\delta}\gamma_1} \bar{B}_{it}, \tag{B.13}$$

where $\bar{B}_{it}$ is an exogenous shock and $\widetilde{\delta} < 1$ indicates the extent to which innovation decays from one period to the next. We assume the cost per unit of land r_{it} is determined by a competitive auction, so that firms obtain zero profits.

Profit maximization: Even though innovations today affect innovations in future periods, because firms earn zero profits in the future, the dynamic problem reduces to a sequence of static profit maximizing problems Desmet & Rossi-Hansberg (2014).

As a result the firms' profit maximization problem becomes:

$$\max_{l_{it},\phi_{it}} p_{it} B_{it}\left(\phi_{it}^{\gamma_1}\right) \times \left(l_{it}^{\mu}\right) - w_{it} \underbrace{l_{it}}_{\text{\# of production workers}} - w_{it} \underbrace{\left(\nu\phi_{it}^{\xi}\right)}_{\text{\# of innovation workers}} - r_{it},$$

which has the following first order conditions:

$$\gamma_1 B_{it} p_{it} \phi_{it}^{\gamma_1 - 1} l_{it}^{\mu} = \xi \nu w_{it} \phi_{it}^{\xi - 1}$$
$$\mu B_{it} p_{it} \phi_{it}^{\gamma_1} l_{it}^{\mu - 1} = w_{it},$$

which combine to yield:

$$\frac{\gamma_1}{\mu} l_{it} = \xi \nu \phi_{it}^{\xi} \iff$$
$$\left(\frac{\gamma_1}{\mu \xi \nu} l_{it} \right)^{\frac{1}{\xi}} = \phi_{it}. \tag{B.14}$$

Total employment $\tilde{l}_{it}$ per unit land is equal to the sum of the production workers and the innovation workers:

$$\tilde{l}_{it} = l_{it} + \nu \phi_{it}^{\xi} \iff$$
$$\tilde{l}_{it} = \left(1 + \frac{\gamma_1}{\mu \xi} \right) l_{it}.$$

Rent and income: Equilibrium rent ensures that profits per unit land are equal to zero:

$$r_{it} = B_{it} p_{it} \phi_{it}^{\gamma_1} l_{it}^{\mu} + w_{it} l_{it} + \nu w_{it} \phi_{it}^{\xi} \iff$$
$$r_{it} = \left(\frac{1}{\mu} + 1 + \frac{\gamma_1}{\mu \xi} \right) w_{it} l_{it}.$$

Note that total income per unit labor in a location is:

$$Y_{it} = r_{it} H_{it} + w_{it} \tilde{L}_{it} \iff$$
$$\frac{Y_{it}}{\tilde{L}_{it}} = \left(\frac{\frac{1}{\mu} + 1 + \frac{\gamma_1}{\mu \xi}}{\left(1 + \frac{\gamma_1}{\mu \xi} \right)} + 1 \right) w_{it}.$$

The productivity microfoundation: The output price is:

$$\mu B_{it} p_{it} \phi_{it}^{\gamma_1} L_{it}^{\mu - 1} = w_{it} \iff$$
$$p_{it} = \frac{1}{B_{it}} \left(\frac{1}{\mu} \left(\frac{\xi \nu \mu}{\gamma_1} \right)^{\frac{\gamma_1}{\xi}} \right) w_{it} l_{it}^{1 - \mu - \frac{\gamma_1}{\xi}}$$

total output is:

$$q_{it} = \phi_{it}^{\gamma_1} B_{it} l_{it}^{\mu} \iff$$
$$Q_{it} = \left(\frac{\gamma_1}{\mu \xi \nu} \right)^{\frac{\gamma_1}{\xi}} B_{it} \tilde{L}_{it}^{\mu + \frac{\gamma_1}{\xi}} H_{it}^{1 - \mu - \frac{\gamma_1}{\xi}},$$

where $\tilde{L}_{it}$ is total employment in location i at time t. Combining equations (B.13) and (B.14) yields:

$$B_{it} = \phi_{it-1}^{\tilde{\delta} \gamma_1} \bar{B}_{it} \iff$$
$$B_{it} = \left(\frac{\frac{\gamma_1}{\mu \xi \nu}}{\left(1 + \frac{\gamma_1}{\mu \xi} \right)} \frac{\tilde{L}_{it-1}}{H_{it-1}} \right)^{\tilde{\delta} \frac{\gamma_1}{\xi}} \bar{B}_{it},$$

so that in total we have:

$$
\begin{aligned}
Q_{it} &= \left(\frac{\gamma_1}{\mu\xi\nu}\right)^{\frac{\gamma_1}{\xi}} \left(\left(\frac{\frac{\gamma_1}{\mu\xi\nu}}{\left(1+\frac{\gamma_1}{\mu\xi}\right)} \frac{\tilde{L}_{it-1}}{H_{it-1}} \right)^{\tilde{\delta}\frac{\gamma_1}{\xi}} \bar{B}_{it} \right) \tilde{L}_{it}^{\mu+\frac{\gamma_1}{\xi}} H_{it}^{1-\mu-\frac{\gamma_1}{\xi}} \iff \\
Q_{it} &= \bar{A}_{it} \tilde{L}_{it}^{\alpha_1} \tilde{L}_{it-1}^{\alpha_2} \tilde{L}_{it},
\end{aligned}
$$

where $\bar{A}_{it} \equiv \left(\frac{\gamma_1}{\mu\xi\nu}\right)^{(1+\tilde{\delta})\frac{\gamma_1}{\xi}} \left(1+\frac{\gamma_1}{\mu\xi}\right)^{-\tilde{\delta}\frac{\gamma_1}{\xi}} \bar{B}_{it} H_{it}^{1-\mu-\frac{\gamma_1}{\xi}} H_{it-1}^{-\delta\frac{\gamma_1}{\xi}}$, $\alpha_1 \equiv \frac{\gamma_1}{\xi} - (1-\mu)$, and $\alpha_2 \equiv \tilde{\delta}\frac{\gamma_1}{\xi}$, as required.

B.2 Amenity spillover

We formalize here a possible microfoundation for the amenity spillover function, $u_{it} = \bar{u}_{it} L_{it}^{\beta_1} L_{it-1}^{\beta_2}$.

Demand: Suppose that consumers have Cobb-Douglas preferences over land and a consumption good, so that their indirect utility function can be written as:

$$W_{it} = \frac{(Y_{it}/L_{it})}{(P_{it})^{\lambda}\left(r_{it}^{H}\right)^{1-\lambda}},$$

where r_{it}^{H} is the rental cost of housing. Let H_{it} denote the (equilibrium quantity) of housing and let K_{it} denote the (exogenous) quantity of land in a location, so that $h_{it} \equiv H_{it}/K_{it}$ is the housing density (e.g. square feet of housing per acre of land).

Given the Cobb-Douglas preferences (and, from balanced trade, that income equals expenditure, $Y_{it} = E_{it}$), we have:

$$
\begin{aligned}
r_{it}^{H} H_{it} &= (1-\lambda) Y_{it} \\
w_{it} L_{it} &= \lambda Y_{it},
\end{aligned}
$$

so that we can write the payment to housing as a function of the payment to labor:

$$r_{it}^{H} = \left(\frac{1-\lambda}{\lambda}\right) \frac{1}{H_{it}} w_{it} L_{it}.$$

Note then that we can write:

$$
\begin{aligned}
W_{it} &= \frac{(Y_{it}/L_{it})}{(P_{it})^{\lambda}\left(r_{it}^{H}\right)^{1-\lambda}} \iff \\
\tilde{W}_{it} &= \frac{1}{\lambda(1-\lambda)^{\frac{1-\lambda}{\lambda}}} \frac{w_{it}}{P_{it}} \left(\frac{H_{it}}{L_{it}}\right)^{\frac{1-\lambda}{\lambda}}, \qquad \text{(B.15)}
\end{aligned}
$$

where $\tilde{W}_{it} \equiv W_{it}^{\frac{1}{\lambda}}$ is a positive monotonic transform of W_{it} that hence can serve as our measure of welfare.

Supply: We now determine the equilibrium stock of housing H_{it}. Suppose that each unit of land is owned by a representative developer, who decides how much to upgrade the housing tract. The amount of housing per unit land ($h_{it} \equiv \frac{H_{it}}{K_{it}}$) is a function of the housing stock that has survived from the previous period $\left(h_{it}^{existing} \equiv \frac{H_{it}^{existing}}{K_{it}}\right)$ and the amount of labor that the firm chooses to hire to rebuild it:

$$
\begin{aligned}
h_{it} &= \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} \iff \\
H_{it} &= \left(H_{it}^{existing}\right)^{\mu} \left(L_{it}^{d}\right)^{1-\mu}.
\end{aligned}
$$

In what follows, we assume for simplicity that the existing housing stock from period $t-1$ in period t is some fraction of the development in the previous period:

$$H_{it}^{existing} = \bar{C}_{it} \left(L_{it-1}^{d}\right)^{\rho}, \tag{B.16}$$

where $\bar{C}_{it}$ is an (exogenous) shock.

Profit maximization: A developer solves:

$$\begin{aligned} \max_{l_{it}^{d}} r_{it}^{H} h_{it} - w_{it} l_{it}^{d} - f_{it} &\iff \\ \max_{l_{it}^{d}} r_{it}^{H} \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} - w_{it} l_{it}^{d} - f_{it}, \end{aligned}$$

where f_{it} is a fixed cost (a "permit cost") that is remitted back to local residents and is set via a competitive biding process, ensuring that the firm earns zero profits (and hence the dynamic problem simplifies into a series of static profit maximization problems, as above).

First order conditions are:

$$\begin{aligned} (1-\mu) r_{it}^{H} \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{-\mu} = w_{it} &\iff \\ \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} &= \frac{1}{1-\mu} \frac{1}{r_{it}^{H}} w_{it} l_{it}^{d}. \end{aligned}$$

Note that the fixed "permit costs" are then:

$$\begin{aligned} f_{it} &= r_{it}^{H} \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} - w_{it} l_{it}^{d} \iff \\ f_{it} &= \left(\frac{\mu}{1-\mu}\right) w_{it} l_{it}^{d}, \end{aligned}$$

which recall are remitted to workers and ensure profits are zero.

We can combine this with the rental rate above to calculate the fraction of workers hired in the development of the land:

$$\begin{aligned} h_{it} = \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} &\iff \\ (1-\mu) \left(\frac{1-\lambda}{\lambda}\right) L_{it} = L_{it}^{d}, \end{aligned}$$

so we require as a parametric restriction (so that only a fraction of workers are hired as local developers):

$$(1-\mu) \left(\frac{1-\lambda}{\lambda}\right) < 1.$$

Since a constant fraction of local workers are hired, we can express the housing density solely as a function of the local population, the local land area, and then:

$$\begin{aligned} h_{it} &= \left(h_{it}^{existing}\right)^{\mu} \left(l_{it}^{d}\right)^{1-\mu} \iff \\ H_{it} &= \left((1-\mu)\left(\frac{1-\lambda}{\lambda}\right)\right)^{(1-\mu)+\rho\mu} \bar{C}_{it}^{\mu} \left(L_{it-1}\right)^{\rho\mu} \left(L_{it}\right)^{1-\mu}. \end{aligned} \tag{B.17}$$

The amenity microfoundation: We substitute equation (B.17) for the equilibrium stock of housing into the welfare equation (B.15) to yield:

$$\tilde{W}_{it} = \frac{1}{\lambda\left(1-\lambda\right)^{\frac{1-\lambda}{\lambda}}} \frac{w_{it}}{P_{it}} \left(\frac{H_{it}}{L_{it}}\right)^{\frac{1-\lambda}{\lambda}} \iff$$
$$\tilde{W}_{it} = \frac{w_{it}}{P_{it}} \bar{u}_{it} L_{it}^{\beta_1} L_{it-1}^{\beta_2},$$

where $\bar{u}_{it} \equiv \frac{1}{\lambda(1-\lambda)^{\frac{1-\lambda}{\lambda}}} \left(\left(1-\mu\right)\left(\frac{1-\lambda}{\lambda}\right)\right)^{\frac{1-\lambda}{\lambda}\left(\left(1-\mu\right)+\rho\mu\right)} \bar{C}_{it}^{\frac{1-\lambda}{\lambda}}$, $\beta_1 \equiv -\mu\frac{1-\lambda}{\lambda}$, and $\beta_2 \equiv \rho\mu\frac{1-\lambda}{\lambda}$ as required.

C Online Appendix: Additional tables and figures

This section includes additional tables and figures mentioned in footnotes in the text.

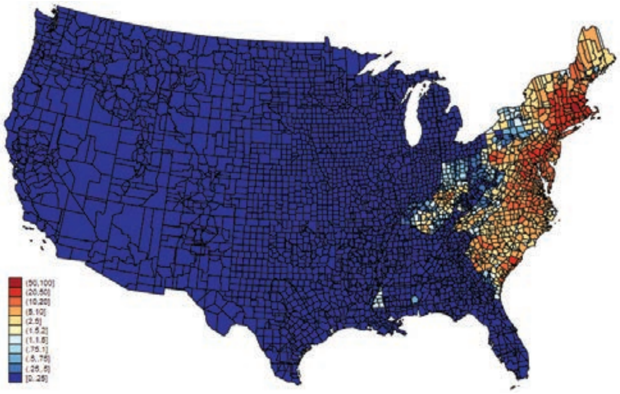
Table C.1: First stage estimates

Endogenous regressors (log):	(1) Pop. (L_{it})	(2) Pop. 50 yrs ago (L_{it-1})
Instruments shifting amenities (used to estimate prod. spillovers):		
Year*Average max. temp. in hottest month (z-score)	-0.271***	-0.052
	(0.034)	(0.067)
Year*Average max. temp. in hottest month (z-score)^2	-0.063***	0.040*
	(0.015)	(0.024)
Year*Average min. temp. in coldest month (z-score)	0.140***	-0.394***
	(0.029)	(0.049)
Year*Average min. temp. in coldest month (z-score)^2	0.170***	0.168***
	(0.018)	(0.029)
Instruments shifting productivities (used to estimate amen. spillovers):		
Year*High - low inten. corn potential yield (z-score)	0.369***	0.583***
	(0.044)	(0.087)
Year*High - low inten. corn potential yield (z-score)^2	0.007	-0.070**
	(0.018)	(0.035)
Year*High inten. soy - low inten. wheat potential yield (z-score)	-0.119***	-0.275***
	(0.046)	(0.090)
Year*High inten. soy - low inten. wheat potential yield (z-score)^2	0.041**	0.032
	(0.016)	(0.028)
Fixed effects:		
Sub-county	Yes	Yes
Region-year	Yes	Yes
F-statistic	22.809	21.968
R-squared	0.890	0.805
Observations	15764	15764

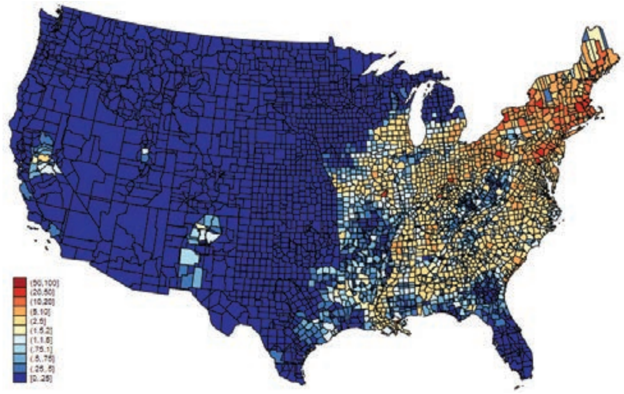
Notes: Ordinary least squares. Each observations is a sub-county from 1850, 1900, 1950 or 2000. Sub-county and region-year fixed effects are included in all specifications. There are 10 regions, constructed using a k-means clustering algorithm based on latitude and longitude. The sample is all sub-counties in all years where geographic instruments and contemporaneous/lagged population values are observed. Standard errors two-way clustered at the sub-county (to allow for serial correlation across time) and county-year level (to allow for data aggregation across sub-counties within year) reported in parentheses. Stars indicate statistical significance: * p<.10 ** p<.05 *** p<.01.

Figure C.1: Spatial distributions of population over time

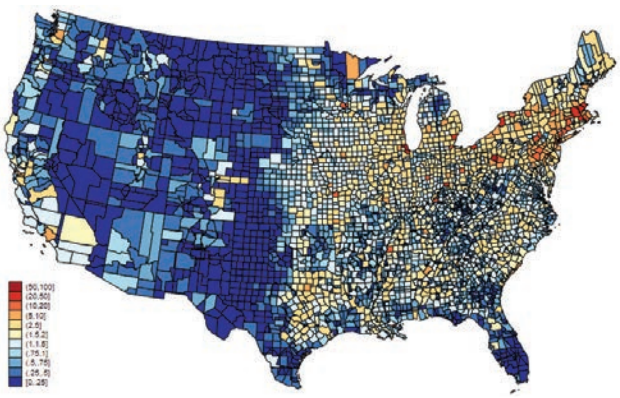
(a) 1800



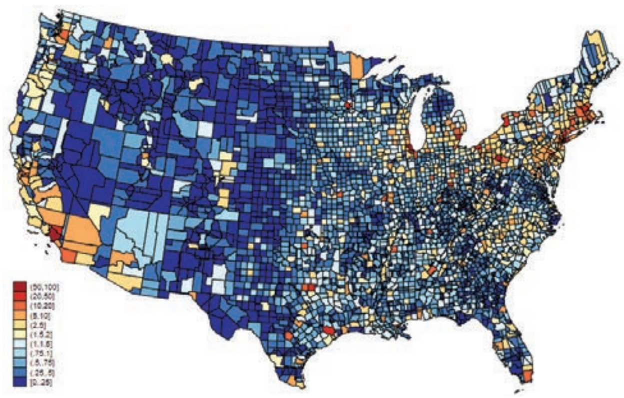
(b) 1850



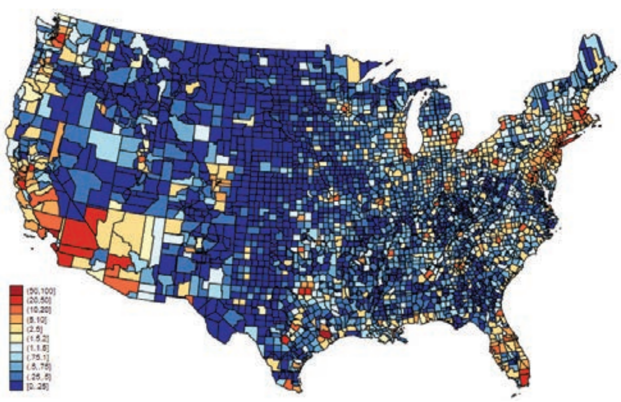
(c) 1900



(d) 1950

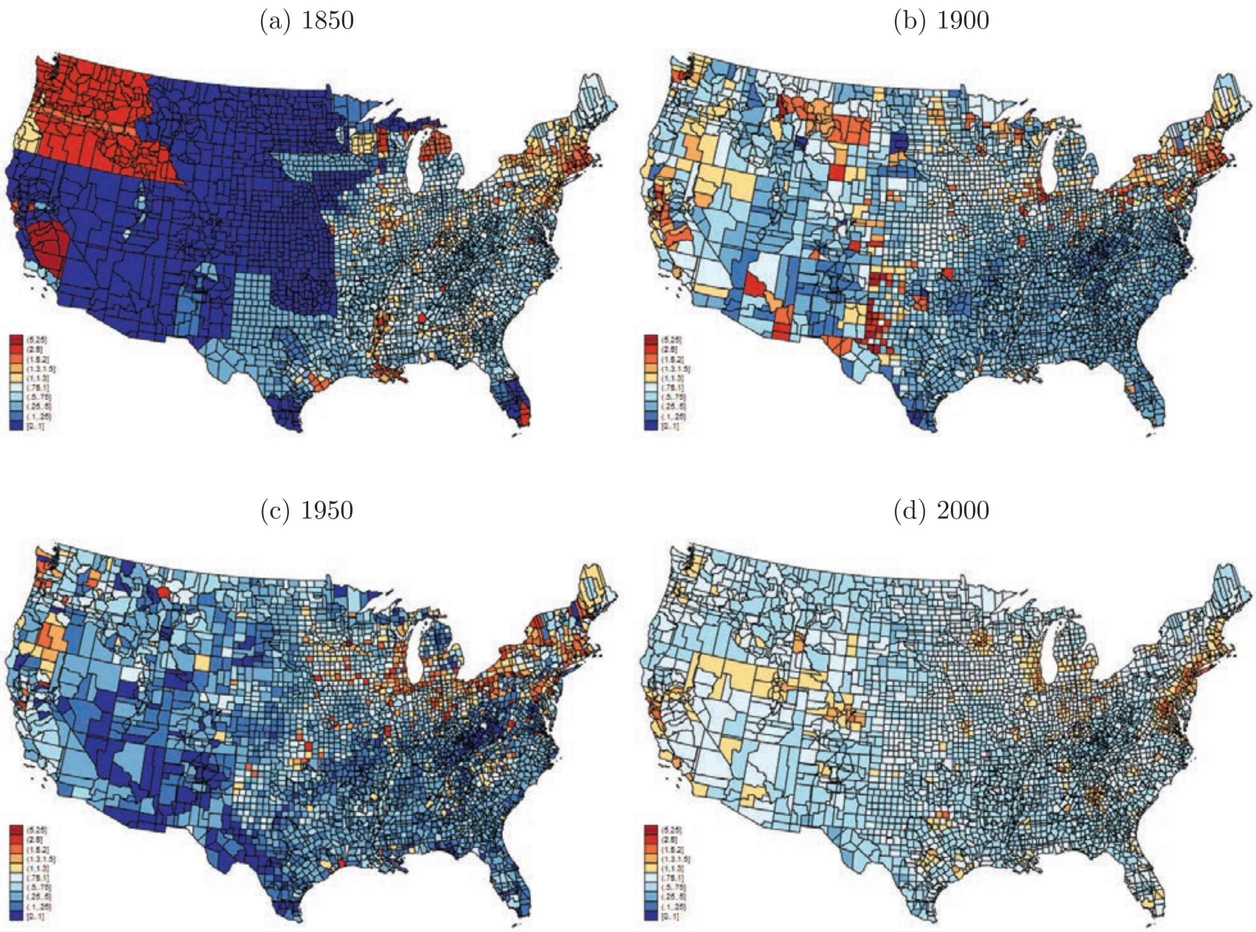


(e) 2000



Notes: This figure illustrates the distribution of population (L_{it}) across all locations from 1800 to 2000. The average population in a location in each year is normalized to one. The colors indicate the value, with red indicating a higher population and blue indicating a lower population.

Figure C.2: Spatial distributions of per capita income over time



Notes: This figure illustrates the distribution of per capita income (w_{it}) in all locations 1850 to 2000. The average value of w_{it} in a location in each year is normalized to one. The colors indicate the value, with red indicating a higher population and blue indicating a lower wage.

Figure C.3: Estimated speed of travel

(a) 1850

(b) 1900

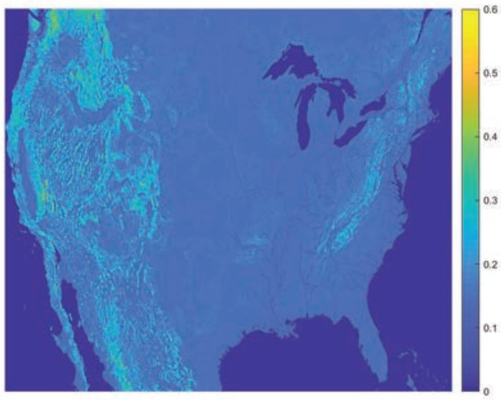
(c) 1950

(d) 2000

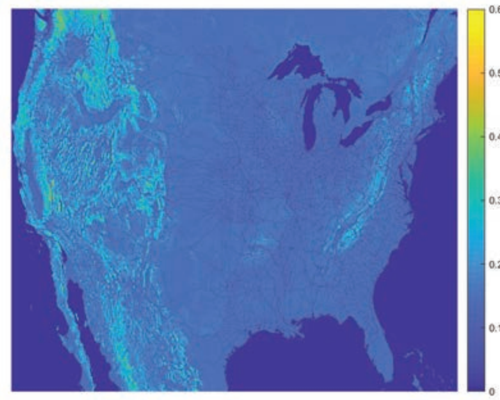
Notes: This figure illustrates the estimated speed of travel (measured in kilometer per hour) for each year. The underlying data is at the 1km×1km square grid, but for readability (and file size concerns), the figures above present the average speed of travel across a disc of 3 kilometer radius. Speed of travel is calculated using the observed local topography, the navigable waterway network, the railroad network, and the road network (by type of road).

Figure C.4: Estimated cost of travel

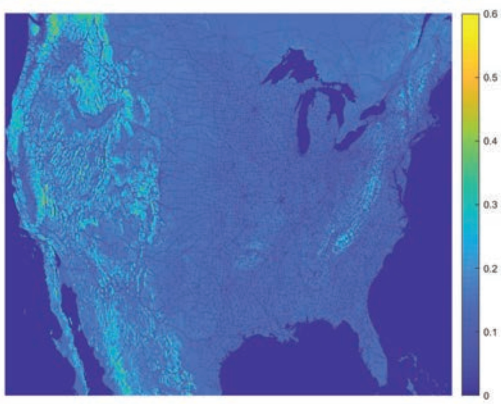
(a) 1850



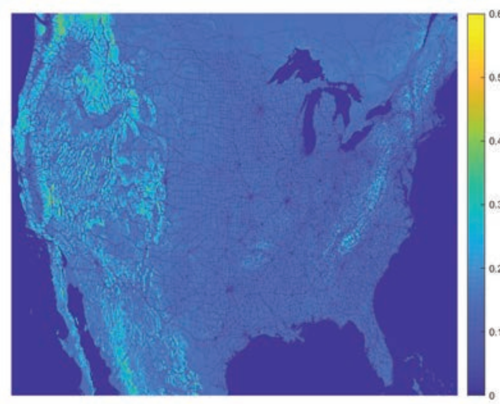
(b) 1900



(c) 1950



(d) 2000



Notes: This figure illustrates the estimated cost of travel (measured in 1850 dollars per ton) for each year. The underlying data is at the 1km×1km square grid, but for readability (and file size concerns), the figures above present the average cost of travel across a disc of 3 kilometer radius. The cost of travel is calculated using the observed local topography, the navigable waterway network, the railroad network, and the road network (by type of road).

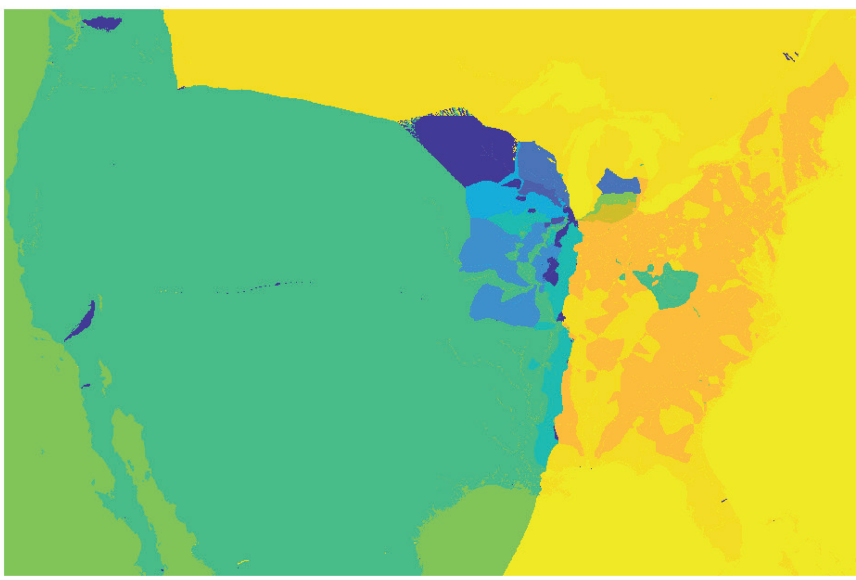
Figure C.5: Estimating trade costs from flows through Chicago

(a) Example data from 1858 Chicago Commerce Report

The following table will show the wheat account for 1858:

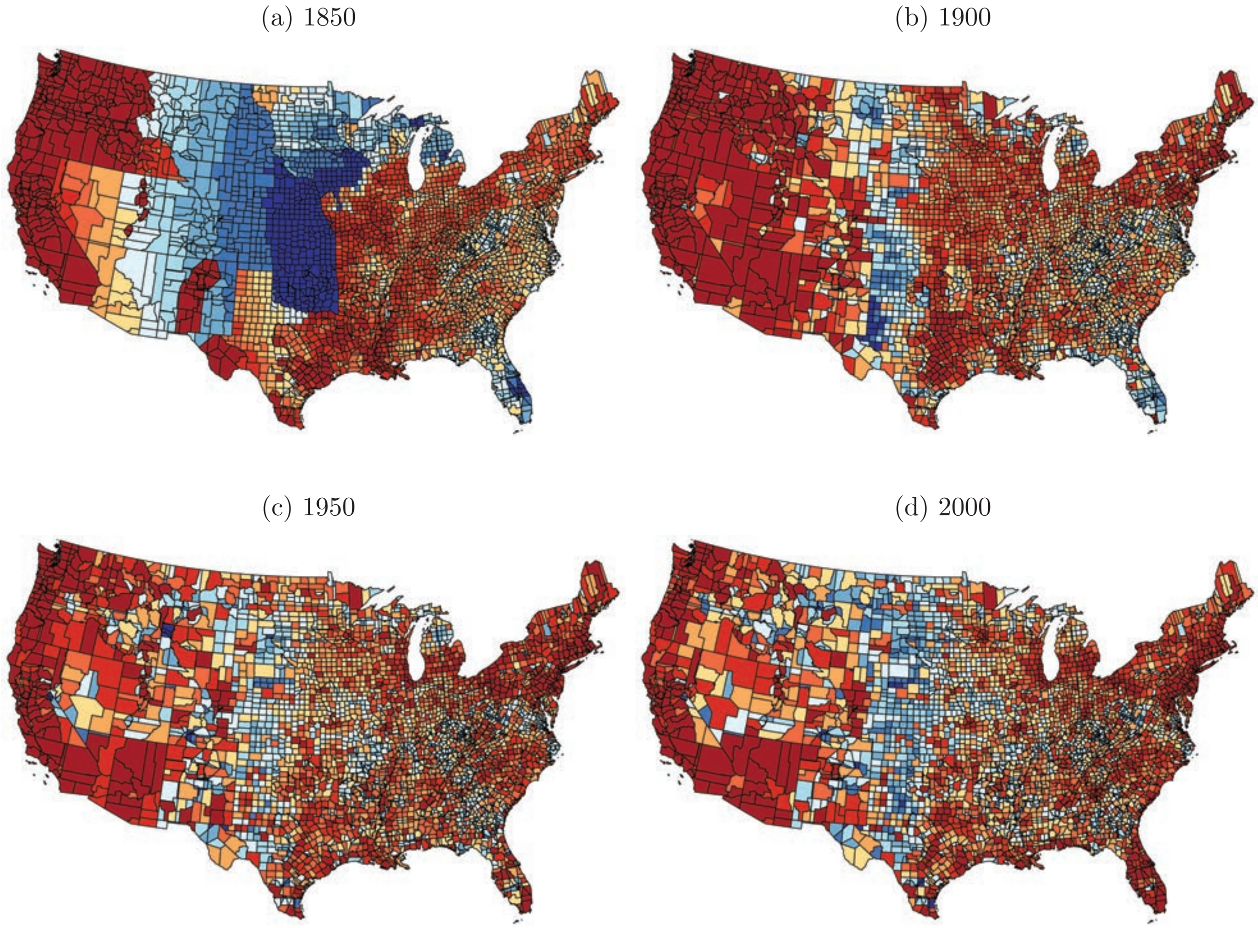
In store, January 2, 1858				307,862
RECEIVED.				
By Lake			4,458	
" Illinois and Michigan Canal			919,754	
" Galena and Chicago Union Railroad			3,880,151	
" Chicago and Rock Island Railroad			523,458	
" Illinois Central Railroad			1,100,482	
" Chicago, Burlington and Quincy Railroad			1,673,075	
" Chicago, St. Paul and Fond du Lac Railroad			342,903	
" Chicago, Alton and St. Louis Railroad			787,575	
" Chicago and Milwaukee Railroad			4,869	
" Michigan Southern Railroad			258,167	
" Michigan Central Railroad			44,722	9,539,614
" Teams, (estimated)				100,000
Total bushels				9,947,476
Disposed of as follows:				
Shipped to Buffalo	4,508,141			
" " Oswego	2,853,862			
" " Ogdensburgh	143,710			
" " other U. S. ports	69,766			
" by Illinois and Michigan Canal	626	7,576,105		
" to Montreal, Canada	439,374			
" " Kingston, Canada	487,536			
" " Collingwood, Canada	26,504			
" " St. Catherines, Canada	172,897			
" " Therold	14,318	1,140,629		
Forwarded by Michigan Southern Railroad	1,829			
" " Michigan Central Railroad	59,700			
" " Illinois Central Railroad	29,219			
" " Chicago, Burlington & Quincy Railr'd	11,288			
" " Pittsburgh, Ft. Wayne & Chicago Railr'd	1,187			
" " Chicago and Milwaukee Railroad	19,781			
" " Chicago and Rock Island Railroad	10,519	133,523	8,850,257	
Floured by City Mills			700,015	
Used by distillers and unaccounted for			122,584	
In store January 1, 1859			274,620	9,947,476

(b) Estimated basins for each mode of travel into Chicago, 1861



Notes: Panel (a) of the figure illustrates the raw data from the 1858 *Annual report of the trade and commerce of Chicago* for wheat. This information (in quantities) is converted to values using the reported prices from the the same publication and combined with similar information from 17 other available commodities to generate the total value of Chicago imports and exports by mode (canal, lake, overland, and the ten different rail lines). Panel (b) of the figure illustrates the mode that we estimate offers the least cost route to Chicago, i.e. the "basin" of each possible mode of transit. For example, the green area to the west of Chicago indicates locations for which the least cost route arrives in/departs from Chicago via canal; the yellow area indicates locations for which the least cost route arrives in/departs from Chicago via Lake Michigan; and the blue and orange regions indicate basins for each respective railroad line (using the 1861 railroad network).

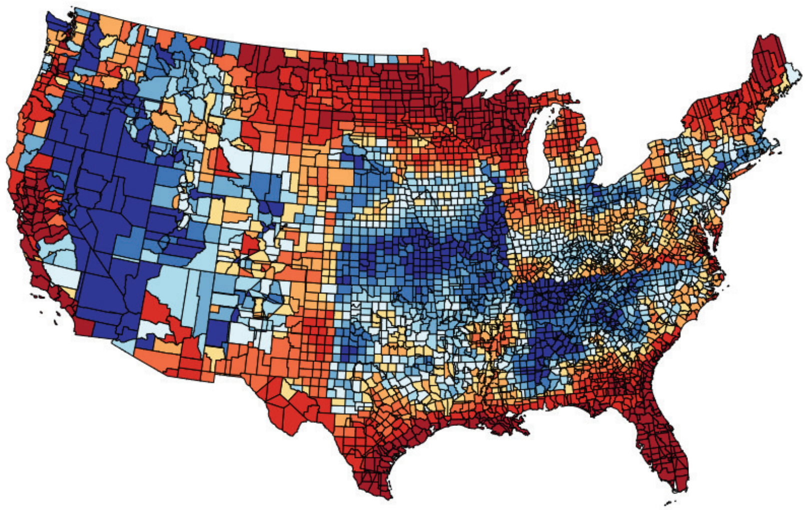
Figure C.6: Spatial distributions of the recovered present discounted value of residence (V_{it}^{θ})



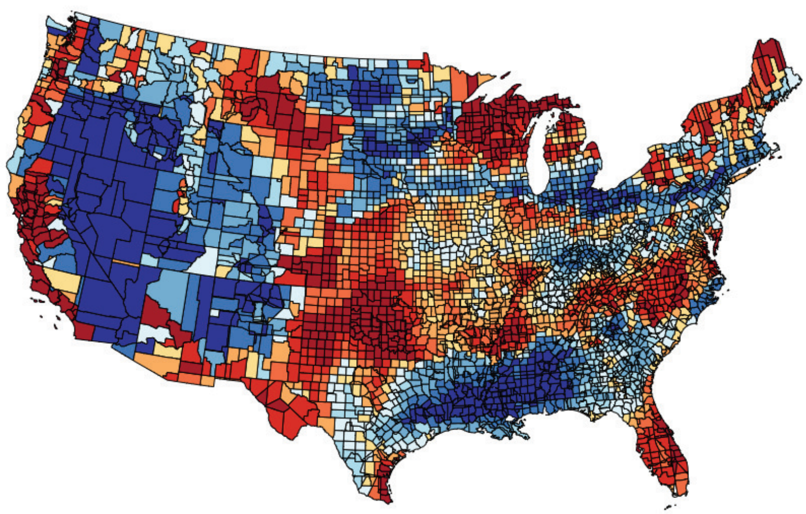
Notes: This figure illustrates the distribution of the present discounted value of residence (to the power of the migration elasticity), V_{it}^{θ}, which can be uniquely recovered given observed populations, outputs, and estimated migration cots; see Proposition 3. The colors indicate the value, with red indicating a higher PDV and blue indicating a lower PDV.

Figure C.7: Estimating productivity and amenity spillovers using plausibly exogenous shifts in labor supply and demand curves over time

(a) Shifts to the supply curve from amenity changes



(b) Shifts to the demand curve from productivity changes



Notes: This figure illustrates the fitted values of the first-stages from the 2SLS regressions in Tables 2 and 3. The left panel shows the predicted change (from 1850-2000) in log population due to plausibly exogenous changes in amenities based on technological improvements which make residing in places with extreme climates of relatively higher amenity value over time. These improvements shift the labor supply curve in each location and can be used to identify the contemporaneous productivity spillover. The right panel shows the predicted trend in log population from plausibly exogenous changes in productivities based on technological improvements and changes in international demand in agricultural production. These improvements shift the labor demand curve in each location and can be used to identify the contemporaneous amenity spillover. Red indicates relatively large values and blue indicates relatively low values.