Regional Science and Urban Economics 94 (2022) 103724

Contents lists available at ScienceDirect

Regional Science and Urban Economics

journal homepage: www.elsevier.com/locate/regec

Persistence and path dependence: A primer[☆]

Treb Allen [a], Dave Donaldson [b,*]

[a] Dartmouth and NBER, USA
[b] MIT and NBER, USA

ARTICLE INFO

Keywords:
Economic geography
Dynamics
Persistence
Path dependence
Agglomeration
Multiplicity

ABSTRACT

How much of the spatial distribution of economic activity today is determined by history rather than by geographic fundamentals? How long should we expect temporary local shocks to persist in their effects on local economic concentration? When will such shocks have permanent (i.e. path-dependent) consequences? This paper develops a simple dynamic model of economic geography—with many heterogeneous locations interacting through trade, migration, agglomeration externalities, and endogenous fertility—that delivers tractable answers to these questions. Our results highlight an important distinction between agglomeration spillovers that endogenously affect productivity (or amenities) contemporaneously and those that do so with a lag.

1. Introduction

Today's urban and regional economists live in historic times. Urbanization increasingly dominates economic life around the globe, and those who study the spatial issues that emerge have never been more blessed with rich datasets and computational power. But we also live in *historical* times—the hypothesis that there runs a strong causal thread from historical to modern spatial conditions has never seen the level of empirical attention that it does today (see Hanlon and Heblich, 2021 and Lin and Rauch, 2021 in this volume for reviews).

This interest in history is natural. Some of the most influential models of economic geography are exactly those that emphasize dynamic behavior in which historical conditions can matter a great deal. These dynamic phenomena could include genuine multiplicity, where the model features indeterminate transitions from period to period. Further, even when uniqueness holds, the model could display *persistence*, a long decay of temporary changes in historical economic conditions, or even *path dependence*, where temporary changes in historical conditions have permanent effects by causing an economy to gravitate towards an alternative steady state. However, these models have been sufficiently rich that explicitly connecting them to empirical estimation and quantification has been a challenge in the sorts of settings (featuring many interacting heterogeneous locations) that typically underpin empirical work.

In this paper we offer a simple economic geography framework that can be used to aid the empirical study of dynamic spatial phenomena such as multiplicity, persistence and path dependence. We allow for an arbitrary number of locations, each of which experiences an unrestricted stream of changes to its exogenous locational characteristics (productivity, amenity value, and spatial frictions). And we work with isoelastic functional forms that emphasize key elasticity parameters that can be estimated in straightforward ways, despite the scope for empirical complexity inherent to an environment with potential equilibrium multiplicity. As we show, the estimated values of such elasticities allow one to assess—irrespective of the underlying exogenous characteristics that may differ across space and time—the model's potential for spatial economic behavior to exhibit a range of dynamic features with rich implications.

In doing so we draw heavily on our earlier work (Allen and Donaldson, 2020). In that model, locations interact via trade (subject to trade costs) and via migration (again subject to frictions), and locations exhibit potential agglomeration externalities in both production and consumption. Crucially, these spillovers depend on both the contemporaneous size of the location and also, potentially, the historical size in each location. A central theme of our analysis, both previously and in the current paper, involves the distinct role played by these contemporaneous and historical spillover functions for determining multiplicity of equilibria, persistence, and the multiplicity of stable steady states (which is necessary for path dependence to arise). Our present analysis

☆ Invited submission for a special issue of *Regional Science and Urban Economics.* We thank the editor and two anonymous referees for helpful comments that have improved this paper. This material is based upon work supported by the National Science Foundation under grant SES-1658838.
* Corresponding author.
E-mail addresses: treb@dartmouth.edu (T. Allen), ddonald@mit.edu (D. Donaldson).

https://doi.org/10.1016/j.regsciurbeco.2021.103724
Received 23 November 2020; Received in revised form 12 August 2021; Accepted 19 August 2021
Available online 30 August 2021

aims to offer a simplified version of this earlier work—a "primer" for those beginning to study spatial persistence and path dependence.

To see the essence of the simplified model studied here, note that when locations interact through both trade and migration, equilibria are described by a system of equations that has a block structure, with each block arising from a particular source of cross-location interaction. Our previous work has four such blocks, but the present paper reduces this to one—a system describing the dynamics of the vector of population sizes in each location i (contemporaneous population, $\{L_{it}\}$, as a function of lagged populations, $\{L_{i,t-1}\}$) and no other additional endogenous variables.

Briefly, this reduction is achieved in two steps. First, we consider the special case in which trade costs are absent and all locations produce a common homogeneous good. This eliminates two of the four blocks (one deriving from a location's exports, which determines its nominal wage; and one deriving from its imports, which determines its cost of living).

Second, we introduce a form of endogenous fertility that (under special circumstances) collapses the two remaining blocks to one. This works as follows. In standard models of costly migration—including ours, in which an overlapping generations set of children are born into a location of their parent's choosing, but then get to choose their own location once they become adults—locations possess two endogenous characteristics: their attractiveness as destinations and their attractiveness as origins at which to be born. Solving for those two characteristics entails two blocks. However, we introduce here a form of endogenous fertility in which parents are more willing to have children if they expect those children to enjoy high lifetime welfare. As a result, there is an offsetting tendency for locations with high origin attractiveness to feature a low probability of any migrant leaving the origin but also more migrants starting out there. Under a particular parameter condition these two forces exactly balance, which removes cross-block interactions and hence allows the system to be written as a single block (which we cast in terms of the endogenous population of each location).

Using this single-block system we emphasize a number of results:

1. Dynamic equilibria are unique and stable when total (i.e. production plus amenity) contemporaneous spillovers are weaker than dispersion forces (the tendency for children to feature idiosyncratic preferences for locations other than their birthplace).
2. In a particular partial equilibrium sense, rates of local persistence (an AR(1) process that relates L_{it} to $L_{i,t-1}$) are higher when either total contemporaneous spillovers are large or total historical spillovers are large.
3. In general equilibrium, rates of economy-wide persistence (an AR(1) process that relates the maximal element-wise change in the vector $\{L_{it}\}$ to that of $\{L_{i,t-1}\}$) are larger when the sum of total contemporaneous and historical spillovers is large.
4. Multiple stable steady states can (and are guaranteed to, for some geographies) arise when the sum of total contemporaneous and historical spillovers is large.

Put together, these results imply that there exists a parameter region in which equilibria are unique and stable, convergence occurs in partial equilibrium, and yet stable steady states can be multiple and path dependence can arise. We provide a brief discussion of analogous results in Allen and Donaldson (2020) that hold in more general environments, including ones that: (i) relax our parameter restriction on fertility preferences; (ii) feature trade costs and differentiated products; (iii) involve parents who live for multiple periods and are forward-looking throughout their lifetime; and (iv) display agglomeration spillovers of more general functional forms than our isoelastic baseline.

We conclude by discussing empirical strategies that can be used to estimate these parameters, as well as a sense of what common parameter estimates from the literature would imply for the aforementioned results.

This work aims to contribute to several literatures. As mentioned above, we hope to extract theoretical insights that can resonate with the largely empirical literature—comprising the work of not only urban and regional economists and economic historians, but also the growing interest among development, labor, macro-, and political economists, for example, in understanding the importance of historical legacies—that has documented numerous forms of empirical persistence in spatial contexts. Prominent studies by, for example, Davis and Weinstein (2002, 2008), Banerjee and Iyer (2005), Glaeser and Gyourko (2005), Bosker et al. (2007), Nunn (2008), Dell (2010), Redding et al. (2011), Bleakley and Lin (2012, 2015), Voigtlander and Voth (2012), Kline and Moretti (2014), Glaeser et al. (2015), Hanlon (2017), Hornbeck and Keniston (2017), Henderson et al. (2018), Michaels and Rauch (2018), and Lee and Lin (2018) have all shaped our understanding of the myriad ways in which temporary historical shocks have (typically, but not always) left long-lived traces on the intra-national, spatial organization of a modern economy. The surveys in Kim and Margo (2014), Nunn (2014), Hanlon and Heblich (2021), Lin and Rauch (2021), and Voth (2020) collect and synthesize much of the evidence on such historical persistence and path dependence.

Second, within the sphere of spatial modeling and quantification, our goal, both here and in Allen and Donaldson (2020), has been to merge lessons from the richly dynamic, but low-dimensional settings of pioneering economic geography models—for example, Krugman (1991), Matsuyama (1991), Rauch (1993), Fujita et al. (1999), Ottaviano (2001), and Ottaviano et al. (2002)—with the more data-driven, high-dimensional, recent tradition of quantitative spatial modeling in both static settings—for example, Roback (1982), Glaeser (2008), Allen and Arkolakis (2014), Ahlfeldt et al. (2015), Redding and Rossi-Hansberg (2017)—and dynamic ones—for example, Desmet et al. (2018), Caliendo et al. (2019), Duranton and Puga (2019), Nagy (2020), and Liu et al. (2021).

2. A simple model of economic geography dynamics

We begin here with a simple model but return to a discussion of certain extensions in Section 3.1. See also Allen and Donaldson (2020) for further discussion and details.

2.1. Setup

Consider a dynamic economic geography model featuring a finite number of locations, indexed by i, and time periods, indexed by t. Agents live for two periods, in an overlapping generations sense. We refer to their first period of life as childhood and their second as adulthood.

2.1.1. Production

Locations are endowed with the ability to produce (under competitive conditions) a nationally homogeneous and freely-traded good produced using local adult labor only. In particular, any producer ϕ in location i at time t produces its output $y_{it}(\phi)$ from its labor input $l_{it}(\phi)$ according to the linear production function $y_{it}(\phi) = A_{it} l_{it}(\phi)$. All producers charge the price $p_{it}(\phi) = 1$ by our choice of numeraire. As such, the nominal (and real) wage in location is given by $w_{it} = A_{it}$.

All producers take their location's productivity A_{it} as given. However, A_{it} is endogenously determined by potential spillovers from the presence of population in a location. In particular, letting $L_{it} \equiv \sum_{\phi} l_{it}(\phi)$ denote the total local adult population in period t, we model productivity as

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

where $\overline{A}_{it}$ represents an arbitrary but exogenous source of productivity. The parameter α_1 captures potential (positive or negative) spillovers, whereby the total local population can affect the productivity of any given producer.[1] Such spillovers, often modeled with the use of precisely such a functional form, are the purview of a large body of work in urban and regional economics (see, for example, Duranton and Puga, 2004 and Redding and Rossi-Hansberg, 2017 for surveys). The representation of spillovers here is admittedly reduced-from, but one that can stand in for a range of underlying mechanisms.[2]

Equation (1) also contains the parameter α_2 as a way of allowing for *lagged* population sizes in a location to potentially affect its current productivity. This is intended to capture—again, in a reduced-form manner—a range of reasons for investments that have been made in the past to remain productive in the present. In Allen and Donaldson (2020) we discuss a number of explicit microfoundations for the form seen in equation (1)—including both its dependence on $L_{it}^{\alpha_1}$ and $L_{i,t-1}^{\alpha_2}$—that have appeared in prior work.[3] As we shall see below, the distinction between contemporaneous (governed by α_1) and historical (by α_2) agglomeration spillovers is crucial for the study of persistence and path dependence in a model such as ours.

An implication of the model so far is that (log, inverse) labor demand in any location is given by

$$\ln w_{it} = \alpha_1 \ln L_{it} + \alpha_2 \ln L_{i,t-1} + \ln \overline{A}_{it}. \tag{2}$$

Notably, in this model the labor demand elasticity will be positive whenever local (contemporaneous) productivity spillovers are positive. As we discuss further in Section 3.2, an equation such as this is commonly used to estimate (contemporaneous) production spillovers, α_1, and standard methods can be extended to obtain estimates of α_2 via equation (2) as well.

2.1.2. Preferences

We turn now to the specification of utility and household decisions. Agents do all their work and consumption during their adulthood time period, and all those residing in location i at t share a systemic component of utility given by

$$W_{it} = w_{it} u_{it}, \tag{3}$$

where u_{it} denotes the amenity value of living in i. Recall that w_{it} is the real (and nominal) wage, so W_{it} reflects the amenity-adjusted real wage. The amenity component u_{it} is itself given as by

$$u_{it} = \overline{u}_{it} L_{it}^{\beta_1} L_{i,t-1}^{\beta_2}, \tag{4}$$

which is hence analogous to the productivity case in equation (1). That is, $\overline{u}_{it}$ embodies exogenous amenity features of location and time, and the second and third components allow amenities to adjust endogenously to both the contemporaneous population (L_{it}) and the historical population ($L_{i,t-1}$). Again, the elasticities β_1 and β_2 represent the strength of (positive or negative) externalities in the contemporaneous and historical senses, respectively. And again, Allen and Donaldson (2020) discusses a range of microfoundations for equation (4).

In addition to the systematic component of utility, W_{it}, each adult also draws an idiosyncratic preference shifter for each location. This draw occurs at the start of adulthood, such that a child born in location j in $t-1$ will choose where to reside as an adult (that is, in period t) after learning her idiosyncratic preferences for all locations. Finally, adults who were born in j in $t-1$ must pay a proportional utility cost in order to "move" to, and hence reside throughout adulthood in, location i at time t. We denote such costs by $\mu_{jit} \geq 1$ and leave them unspecified in what follows, though we note that a natural component of these costs may include distance, transportation facilities, etc.

We further assume that agents' idiosyncratic preferences for each location derive from a multivariate Fréchet distribution with shape parameter $\theta > 1$ and location-specific scale parameters normalized to one without loss of generality. Following standard derivations, this means that the probability that any child born in j at $t-1$ will choose to live as in i as an adult in t is

$$\pi_{jit} = \frac{\left(W_{it}/\mu_{jit}\right)^{\theta}}{\sum_k \left(W_{kt}/\mu_{jkt}\right)^{\theta}}. \tag{5}$$

It also proves useful below to define the numerator of equation (5) explicitly via

$$\Pi_{jt} \equiv \left(\sum_k \left(W_{kt}/\mu_{jkt}\right)^{\theta} \right)^{\frac{1}{\theta}}. \tag{6}$$

Indeed, Π_{jt} is equal to the expected value of welfare, with the expectation taken across idiosyncratic preference differences, for a child who is born in location j at time $t-1$.

We assume that there are a very large number of adults in each location such that this probability is also well approximated by the actual share doing so. This implies that inter-location population flows will be more responsive to differences in systematic utility W_{it} over space when θ is larger (because this corresponds to less dispersion in idiosyncratic preferences). Low values of θ act as a dispersion force in this model because they reduce the extent to which adults seek out locations with greater systematic appeal.

2.1.3. Endogenous fertility

The final ingredient of the model is each adult's decision about fertility. In Allen and Donaldson (2020) we assumed that all adults had an exogenous number of children, but we endogenize this decision here. Every adult in location j at time $t-1$ decides on the number of children $l_{j,t-1}$ to raise. We model this in line with the canonical theory of endogenous fertility, as expressed in, for example, Barro and Becker (1989). In particular, we assume that parents desire to maximize a net benefits function $\Xi_{j,t-1}(l_{j,t-1})$, given by

$$\Xi_{j,t-1} \equiv \max_{l_{j,t-1}} \Pi_{j,t} l_{j,t-1} - \frac{1}{\lambda} c_{j,t-1} l_{j,t-1}^{\lambda}. \tag{7}$$

This function posits both a perceived benefit (the first term) and a perceived cost (the second term) that parents experience when they have $l_{j,t-1}$ children. In particular, we assume that the per-child benefits scale with $\Pi_{j,t}$, the expected lifetime welfare of a child born in location j at time t; further, we assume that the costs are determined by a child-rearing disutility function that we assume is convex (i.e. $\lambda > 1$) for some exogenous constant $c_{j,t-1}$.[4]

It is important to note the extent to which agents have rational and forward-looking behavior in this model. Each young adult makes his

[1] Equation (1) suggests a form of *local* spillover whose scope may depend on the spatial scale in question.

[2] We refer to the parameter α_1 (and those analogous to it) as spillovers but it is important to note that not all potential microfoundations for equation (1) would involve true agglomeration externalities. For example, a natural reason to suspect $\alpha_1 < 0$ would derive from our omission of capital or land.

[3] Briefly, such microfoundations could include: (i) a model (a la Deneckere and Judd, 1992) with innovating firms that enjoy patent protection in period t, face competition without such protection in $t+1$, and suffer from product obsolescence in $t+2$; or (ii) a model (a la Desmet and Rossi-Hansberg, 2014) of endogenous innovation in which there are contemporary scale effects and spatial spillovers that occur with a lag. Richer historical processes—for example, immobile capital that decays slowly over several periods—would require equation (1) to be augmented to include further lags, such as $L_{i,t-2}$, etc. Such additional effects could be incorporated into our analysis in a straightforward manner provided that they enter in an analogously isoelastic form.

[4] Like Barro and Becker (1989), the program in equation (7) proposes that parents derive utility from the complementarity between the utility that each child will themselves enjoy and the number of such children the parent has. In practice, Barro and Becker (1989) assume that parental altruism is diminishing (rather than linear in 7), whereas child-rearing costs are linear (rather than convex as in 7) but this alternative amounts to a simple change in variables in (7).

migration decision in a forward-looking manner, and he fully understands the welfare W_{it} that he will achieve in equilibrium upon choosing any destination i. The fact that adulthood lasts just one period means that this young adult correctly understands that there is no need to evaluate payoffs beyond period t. However, this young adult bases his location decision on the basis of where his own consumption W_{it} will be highest, rather than on where the net benefits of child-rearing Ξ_{it} may be highest. That is, parents in this model are selfish (since they do not forego their own consumption for the sake of their future child's consumption), even though when they choose a fertility level they do incorporate the utility that they know their child will attain (via equation (7)).

Returning to the parent's fertility decision, equation (7) implies that the number of children per adult born in location j will satisfy

$$l_{j,t-1} = \left(\frac{\Pi_{jt}}{c_{j,t-1}}\right)^{\frac{1}{\lambda-1}}. \tag{8}$$

This means that the total number of adults migrating from j to i at the start of period t will be $L_{jit} = \pi_{jit} \times l_{j,t-1} \times L_{j,t-1}$, which stems from the combination of endogenous fertility $l_{j,t-1}$ per adult at location j, the number of adults $L_{j,t-1}$ at location j, and the probability π_{jit} of an adult's child deciding to move from j to i. Combining equations (5) and (8), and imposing a scale normalization that is without loss of generality,[5] we then have

$$L_{jit} = \mu_{jit}^{-\theta} W_{it}^{\theta} \left(\Pi_{jt}\right)^{\frac{1}{\lambda-1}-\theta} L_{j,t-1}. \tag{9}$$

At this stage we are positioned to state the labor supply relationship that prevails in each location and time period. Substituting equations (3) and (4) into (9), summing across destinations, and rearranging yields the (inverse) labor supply equation

$$\ln w_{it} = \left(\frac{1}{\theta} - \beta_1\right) \ln L_{it} - \beta_2 \ln L_{i,t-1} - \frac{1}{\theta} \ln IMMA_{it} - \ln \bar{u}_{it} \tag{10}$$

where $IMMA_{it}$ refers to the "inward migration market access" of location i at time t, which in turn is given by

$$IMMA_{it} \equiv \left(\sum_j \mu_{jit}^{-\theta} \left(\Pi_{jt}\right)^{\frac{1}{\lambda-1}-\theta} L_{j,t-1}\right). \tag{11}$$

Similarly to the labor demand case, the elasticity of labor supply can take either sign depending on the strength and sign of contemporaneous spillovers β_1. Equation (10) can form the basis of parameter estimation that can be used to measure the strength of contemporaneous spillovers β_1 (which, net of the dispersion effects from θ, govern the elasticity of labor supply) as well as historical spillovers β_2. We return to this point in Section 3.2.

2.1.4. A simplifying parameter restriction

The presence of the expected welfare Π_{jt} in expression (9) reflects two forces. The first, governed by the elasticity $-\theta$, reflects the (standard) tendency for location j to feature relatively less out-migration (to any given destination i) if the expected welfare of all migration options from the perspective of location j is high. The second, governed by the elasticity $\frac{1}{\lambda-1}$, reflects the tendency for a location that offers high expected welfare to have higher fertility and hence a greater number of potential out-migrants to any given i. These two channels each add complex, but offsetting, general equilibrium features to a model that (as we shall see below) already features important general equilibrium interactions across locations. So in the spirit of simplicity we restrict attention in what follows to the special case in which the preference parameters θ and λ satisfy the joint restriction that

$$\frac{1}{\lambda-1} = \theta. \tag{12}$$

Under this restriction, the Π_{jt} term plays no further role in our analysis because its two potential roles in equation (9) exactly balance each other. Section 3.1 discusses the implications of departing from the parameter restriction in equation (12).

We now use equation (9) to derive a simple expression for the dynamics of total adult populations L_{it} in each location. Adding up mandates that $L_{it} = \sum_j L_{jit}$ for each i and t, which then implies (imposing the restriction 12)

$$L_{it} = W_{it}^{\theta} \sum_j \mu_{ijt}^{-\theta} L_{j,t-1}. \tag{13}$$

This expression relates populations at t to those at $t-1$, but also depends on the endogenous utility component W_{it}. However, equations (1)–(4) together imply that W_{it} can be written as a function of population, via

$$W_{it} = \bar{A}_{it} \bar{u}_{it} L_{it}^{\alpha_1+\beta_1} L_{i,t-1}^{\alpha_2+\beta_2}. \tag{14}$$

That is, the systematic component of welfare W_{it} enjoyed at any location is a function of exogenous productivity and amenity components, $\bar{A}_{it}\bar{u}_{it}$, as well as an endogenous contribution from the contemporaneous population, driven by the net amount of contemporaneous spillovers $\alpha_1 + \beta_1$, and a similar contribution from historical population governed by net historical spillovers, $\alpha_2 + \beta_2$.

Using (14) to eliminate W_{it} in (13), we hence arrive at

$$L_{it}^{1-\theta(\alpha_1+\beta_1)} = \left(\bar{A}_{it}\bar{u}_{it}\right)^{\theta} \times L_{i,t-1}^{\theta(\alpha_2+\beta_2)} \times \left(\sum_j \mu_{jit}^{-\theta} L_{j,t-1}\right). \tag{15}$$

Equation (15) is the key equilibrium relation in this model. For any path of the exogenous fundamentals $\left\{\bar{A}_{it}, \bar{u}_{it}, \mu_{ijt}\right\}$ and any initial conditions $\{L_{i0}\}$, this equation relates the population in any location L_{it} in a period to the vector of populations $\{L_{j,t-1}\}$ in all locations in the previous period.

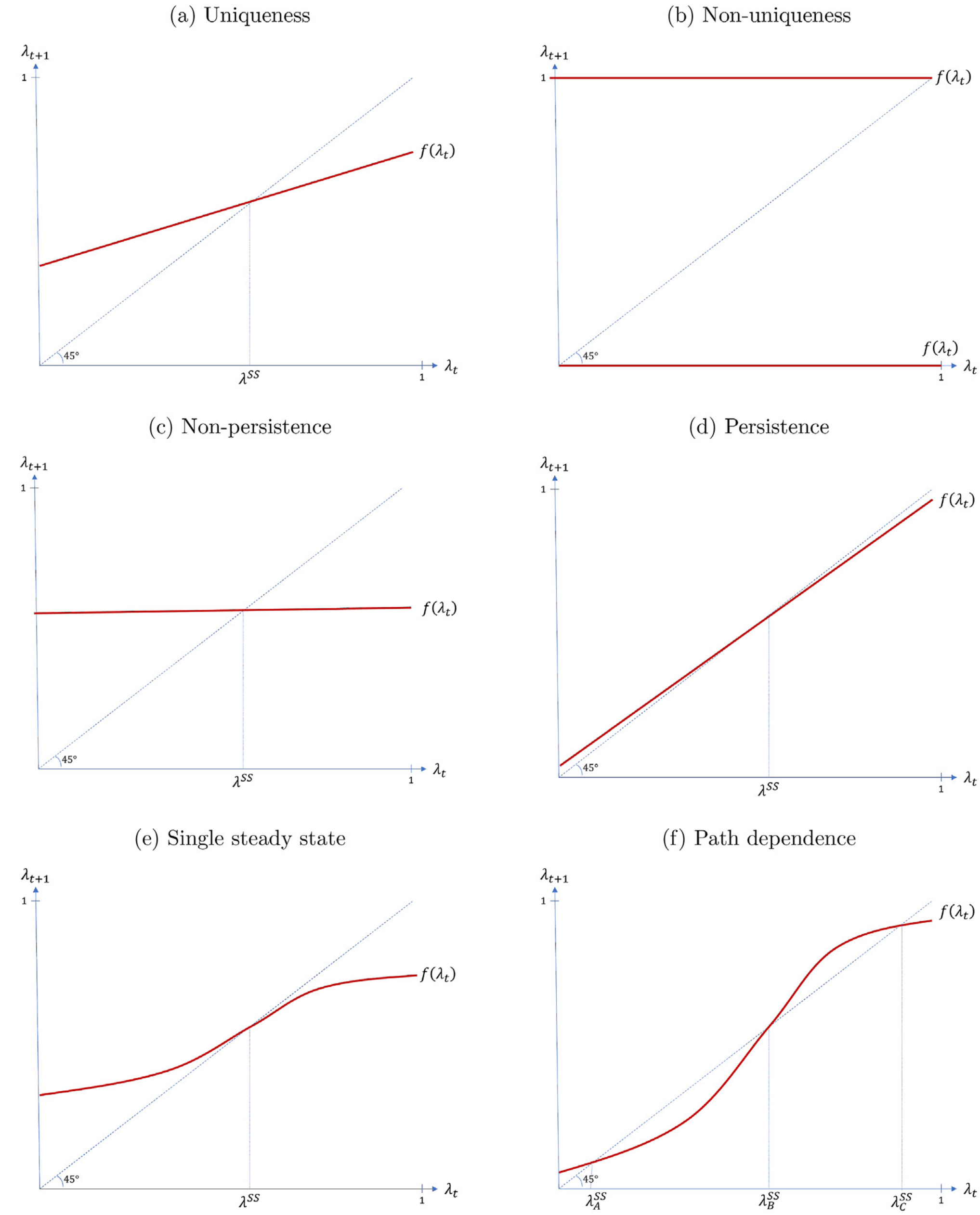
What follows below consists entirely of exploring the equilibrium implications of this system. We stress two features at this point. First, L_{it} on the left-hand side is raised to the power of $1-\theta(\alpha_1+\beta_1)$, which reflects the joint role of contemporaneous net agglomeration forces (given by the spillover parameters $\alpha_1 + \beta_1$) and dispersion forces (given inversely by θ). As we shall see, much rides on the question of whether agglomeration forces are low enough, or dispersion forces are high enough, such that the combined agglomeration and dispersion forces lie in the range where $\theta(\alpha_1+\beta_1) < 1$. Second, historical populations turn up in two respects: (i) through the direct effect of location i's historical population on its contemporaneous productivity (if $\alpha_2 + \beta_2 \neq 0$) and (ii) through the extent to which location i can be reached via low migration costs from the distribution of historical populations in all locations (the term $\sum_j \mu_{jit}^{-\theta} L_{j,t-1}$). These two forms of historical dependence play distinct roles, as we discuss below.

2.2. Qualitative implications

The simple dynamic model of Section 2.1 has a number of implications that we now explore. These relate to qualitative features such as whether the model has unique (and stable) equilibria, to the speed of convergence towards a steady state that can be expected, and to the question of whether steady states will be unique.

Before considering such features in the full model, it is useful to define concepts in the context of a simple case. To that end, consider a generic two-location model whose endogenous outcomes

[5] This normalization sets $\left(c_{j,t-1}\right)^{\frac{1}{1-\lambda}} = 1$ for all locations j and time periods t. It is without loss because location-specific child-rearing costs $c_{j,t-1}$ are isomorphic in this model to location-specific amenities and productivities.



Notes: This figure provides example dynamic systems in a two location world to illustrate the concepts of uniqueness, persistence, and path dependence. The distribution of the spatial economy is summarized by the share of the world population in location 1 in time t, $\lambda_t \in [0, 1]$; the (possibly multivalued) function f defines the dynamic evolution of the economy, where $\lambda_{t+1} = f(\lambda_t)$. See the text for a detailed discussion.

Fig. 1. Uniqueness, persistence, and path dependence.

can be studied by tracking the share of total population in location "one" at time *t*, which we denote by λ_t.[6] Suppose this model's dynamic system can be summarized by the scalar mapping $\lambda_{t+1} = f(\lambda_t; \left\{\overline{A}_{i,t+1}\right\}, \left\{\overline{u}_{i,t+1}\right\}, \left\{\mu_{ij,t+1}\right\}, \Theta)$, where $\Theta \equiv (\alpha_1, \alpha_2, \beta_1, \beta_2, \theta)$. This mapping is analogous to the dynamic vector system in equation (15). We suppress the dependence of the mapping on exogenous components and parameters and write $\lambda_{t+1} = f(\lambda_t)$ for short in what follows.

Fig. 1 illustrates six possible forms that $f(\cdot)$ may take. First, panels (a) and (b) demonstrate the matter of whether the model exhibits *uniqueness* of equilibria for any initial condition λ_t, or not. This hinges on whether $f(\cdot)$ is a single-valued function—leading to uniqueness for all λ_t, as in panel (a)—or a multi-valued function—leading to non-uniqueness, as in panel (b). Proposition 1 below discusses this case.

Second, panels (c) and (d) describe the concept of *persistence*. In both cases, $f(\cdot)$ is strictly upward-sloping and single-valued (so we have

[6] In the model of Section 2.1, the economy's total population is endogenous so it is not necessarily the case that (holding fundamentals fixed) two equilibria with identical population shares will have the same population level. The discussion surrounding Fig. 1 abstracts from this further element of indeterminacy but the general results in Propositions 1-3 do not.

uniqueness), but with a slope less than one and crossing the 45-degree line from above. This implies that there is a unique and stable steady state, denoted λ^{SS}. In panel (c) the slope of $f(\cdot)$ is comparatively flat, indicating very little persistence in this economy—that is, regardless of initial conditions λ_t, the system will converge quickly to λ^{SS}. By contrast, in figure (d) the function $f(\cdot)$ has a slope that approaches one from below, indicating a great deal of persistence. One would see a stark trace of a shock to λ_t in this economy for many periods. Proposition 2 below discusses this matter.

Finally, panels (e) and (f) illustrate the notion of *path dependence* that we emphasize in this paper. Panel (e) shows an economy with a single stable steady state, whereas (f) shows an economy with two stable steady states (λ_A^{SS} and λ_C^{SS}) and one unstable one (λ_B^{SS}). We say that economy (f) could exhibit path dependence but that economy (e) could not. This is because in the case of (e) the eventual steady state reached does not depend on initial conditions λ_t—nor, on the "path" of values of λ that have occurred in the past as long as $f(\cdot)$ is constant. By contrast, in panel (f) it is clear that if the economy were to start with $\lambda_t < \lambda_B^{SS}$ then it would eventually settle at λ_A^{SS}; similarly, if it started out with $\lambda_t > \lambda_B^{SS}$ it would settle at λ_C^{SS}. This implies that a temporary shock to λ_t, if it were to move this state variable across the boundary at λ_B^{SS}, would lead to a permanent change in the long-run outcome of the system. That is, a necessary condition for path dependence to occur is that steady states are multiple. Proposition 3 below describes conditions under which this can (and cannot) arise in our model.

2.2.1. Uniqueness

Given the externalities embodied in equations (1) and (4), a natural question to ask is whether the equilibrium of this dynamic economic geography model will be uniquely pinned down by exogenous parameters. Given any initial population distribution $\{L_{i0}\}$ and geography $\{\overline{A}_{it} > 0, \overline{u}_{it} > 0, \mu_{ijt} \geq 1\}$, and as long as $\theta(\alpha_1 + \beta_1) \neq 1$, equation (15) can be applied iteratively from period $t = 1$ onward to determine the full evolution of the economy. As a result, the following proposition is immediate:

Proposition 1. *If $\theta(\alpha_1 + \beta_1) \neq 1$, then, for any initial population distribution $\{L_{i0}\}$ and geography $\{\overline{A}_{it} > 0, \overline{u}_{it} > 0, \mu_{ijt} > 0\}$, there exists a unique equilibrium.*

To see the intuition behind this result, note that as $\theta(\alpha_1 + \beta_1)$ approaches one from below, L_{it} in equation (15) becomes infinitely large and infinitely responsive to the combination of exogenous characteristics and predetermined populations on the right-hand side of (15). That is, in such a scenario, all population would agglomerate around the location with the best fundamentals. Yet the same argument could be made for other locations too. These extreme equilibria of perfect specialization, sometimes referred to as "black hole" equilibria, are the root of potential multiplicity in this model. If $\theta(\alpha_1 + \beta_1) > 1$, there also exists a unique equilibrium, but it is one in which worse locations have greater populations so that their larger agglomeration forces exactly offset their inferior geographic fundamentals. Because such an equilibrium is inherently unstable (in the sense that any perturbation that moved individuals from locations with inferior geographic fundamentals would be welfare-improving for those individuals), in what follows we focus on parameter constellations that satisfy the following condition:

$$\theta(\alpha_1 + \beta_1) < 1, \tag{16}$$

which guarantees the existence and uniqueness of a stable equilibrium.

Condition (16) depends, as is intuitive, on the conflict—anticipated above—between the model's dispersion and agglomeration forces that are at work *contemporaneously*. Dispersion forces are strong when children have stronger desires to spread out from their birthplace locations due to the idiosyncratic preferences (which scale inversely with θ). Contemporaneous agglomeration forces are governed, on net, simply by $\alpha_1 + \beta_1$, which can take any sign. Recalling that $\theta > 1$, we see that if net contemporaneous agglomeration forces are negative ($\alpha_1 + \beta_1 < 0$) then uniqueness is guaranteed. Otherwise, when agglomeration forces are present ($\alpha_1 + \beta_1 > 0$) the question of uniqueness and stability hinges on whether these agglomeration spillovers are offset by dispersion forces (as condition 16 requires) or not.

A final important point about condition (16) is that it does not depend on historical agglomeration forces, α_2 or β_2. This is because forward-looking agents take all past values of the state of their economy (summarized by the lagged population distribution, $\{L_{i,t-1}\}$) as given.

2.2.2. Persistence (in partial equilibrium)

Uniqueness, as discussed in Proposition 1, guarantees only that from a given set of exogenous conditions the path of the model's endogenous population flows and spatial agglomerations can be pinned down with certainty. This of course says nothing about the *nature* of the mapping from exogenous conditions to endogenous outcomes. We now explore this mapping further.

We begin by examining a partial equilibrium notion of persistence in this model economy. Taking logs of equation (15) we have

$$\ln L_{it} = \frac{\theta}{1-\theta(\alpha_1+\beta_1)} \ln\left(\overline{A}_{it}\overline{u}_{it}\right) + \frac{\theta(\alpha_2+\beta_2)}{1-\theta(\alpha_1+\beta_1)} \ln L_{i,t-1} + \frac{1}{1-\theta(\alpha_1+\beta_1)} \ln \sum_j \mu_{jit}^{-\theta} L_{j,t-1}, \tag{17}$$

which is of course a simple AR(1) process for the log of the population at each location i. That is, the exogenous component of the autoregressive process is given by the first term, reflecting $\ln\left(\overline{A}_{it}\overline{u}_{it}\right)$. General equilibrium, cross-region interactions are governed by the third term, which takes a migration "market access" form. The notion of partial equilibrium that we use here consists of holding this market access term constant. This is a coherent exercise in the case of a model with many small regions, such that the mechanical contribution of location i to its own market access term is negligible. When this is the case, we can discuss the effect on the current population of a *ceteris paribus* increase in the log population in the previous period, $\ln\left(L_{i,t-1}\right)$, captured by the second term of (17). Such an effect is then given by the *partial equilibrium AR(1) coefficient*, ρ^{PE}, which we define as

$$\rho^{PE} \equiv \frac{\theta(\alpha_2+\beta_2)}{1-\theta(\alpha_1+\beta_1)}. \tag{18}$$

This coefficient will govern fully, at least in partial equilibrium, the persistence of the population in any location i in this model economy.[7] Interpreting this coefficient, it is helpful to begin by restricting attention to the case where condition (16) is satisfied and there exists a unique stable equilibrium, which means that the denominator of expression (18) is positive. However, we see immediately that, even in this uniqueness region, if contemporaneous spillovers were to increase, such that $\theta(\alpha_1 + \beta_1)$ approached one from below, then ρ^{PE} would grow without bound. That is, in a model with uniqueness, the closer the parameters get to the non-uniqueness threshold of condition (16), the greater the persistence, all else equal.

By contrast, the numerator of equation (18) emphasizes the role of α_2 and β_2, the historical spillover parameter values. If the sum of these historical effects is positive, then ρ^{PE} will also be positive. But again, the larger such spillovers become, the greater the persistence. Indeed, there

[7] In principal, one could imagine estimating the partial equilibrium persistence ρ^{PE} using a regression of the form of equation (17). Such a procedure must contend not only with the need to suitably control for the general equilibrium market access and potentially unobserved productivities and amenities highlighted here, but also that (a) L_{it} and L_{it-1} are negatively correlated in the presence of measurement error; and (b) such a regression may have different interpretations in alternative frameworks (see e.g. Duranton and Puga (2014)).

is nothing to rule out a scenario in which ρ^{PE} is actually larger than one, which implies *divergence* of this partial equilibrium dynamics system. We note that such divergence can be ruled out, in partial equilibrium, whenever $\rho^{PE} < 1$, or

$$\theta\left(\alpha_1+\beta_1+\alpha_2+\beta_2\right)<1. \tag{19}$$

The key lesson from equation (18) is that persistence will be greater (even resulting in divergence, if condition 19 is violated) whenever net contemporaneous spillovers ($\alpha_1 + \beta_1$) are large, whenever net historical spillovers ($\alpha_2 + \beta_2$) are large, and whenever dispersion effects (θ) are small.

2.2.3. Persistence (in general equilibrium)

We can extend the previous partial equilibrium logic into full general equilibrium by simply incorporating an analysis of how the market access term in equation (15) itself evolves endogenously. To consider the cases of $1-\theta\left(\alpha_1+\beta_1\right) \lesseqgtr 0$ simultaneously, we define $x_{it} \equiv L_{it}^{1-\theta(\alpha_1+\beta_1)}$ such that equation (15) can be written in proportional changes (defining $\hat{x}_{it} \equiv x_{it}/x_{i,t-1}$) as

$$\hat{x}_{it}=\left(\hat{\overline{A}}_{it}\hat{\overline{u}}_{it}\right)^{\theta}\times\hat{x}_{i,t-1}^{\frac{\theta(\alpha_2+\beta_2)}{1-\theta(\alpha_1+\beta_1)}}\times\sum_j\left(\frac{\mu_{jit}^{-\theta}x_{j,t-2}^{\frac{1}{1-\theta(\alpha_1+\beta_1)}}}{\sum_j\mu_{ji,t-1}^{-\theta}x_{j,t-2}^{\frac{1}{1-\theta(\alpha_1+\beta_1)}}}\right)\hat{x}_{j,t-1}^{\frac{1}{1-\theta(\alpha_1+\beta_1)}}.$$

We then define the maximum and minimum (across locations) proportional changes in x_{it} as $M_t \equiv \max_j \hat{x}_{jt}$ and $m_t \equiv \min_j \hat{x}_{jt}$, respectively. Letting χ_t denote the ratio of these maximum and minimum changes in x_{it} we then have

$$\begin{aligned}\ln\chi_t \leq{} & \theta\ln\left(\frac{\max{}_j(\hat{\overline{A}}_{jt}\hat{\overline{u}}_{jt})}{\min{}_j(\hat{\overline{A}}_{jt}\hat{\overline{u}}_{jt})}\right)\\ &+\left(\left|\frac{\theta(\alpha_2+\beta_2)}{1-\theta(\alpha_1+\beta_1)}\right|+\left|\frac{1}{1-\theta(\alpha_1+\beta_1)}\right|\right)\ln\chi_{t-1}.\end{aligned} \tag{20}$$

This expression bounds the maximal differences in speeds of change over space, χ_t, as a function of the maximal difference in speeds of change in productivities and amenities, and the value of χ_{t-1} in the previous period. It is akin to an AR(1) process for the variable χ_t, but holding as a lower bound rather than an equality. The first term in equation (20) comprises exogenous conditions that will introduce spatial churn, but holding this constant the coefficient on $\ln\chi_{t-1}$ describes a bound on the speed of aggregate convergence in this system. We summarize this result in the following proposition:

Proposition 2. *Consider any initial population* $\{L_{i0}\}$ *and time-invariant geography* $\{\overline{A}_i>0,\overline{u}_i>0,\mu_{ij}>0\}$. *Suppose that* $\theta\left(\alpha_1+\beta_1\right)<1$ *so that from Proposition* 1 *the dynamic equilibrium is unique. Further suppose that* $\alpha_2+\beta_2 > 0$ *so that historical spillovers are net agglomerative. Then the following relationship holds:*

$$\ln\chi_t \leq \rho^{GE}\ln\chi_{t-1}, \tag{21}$$

where

$$\rho^{GE}\equiv\frac{1+\theta\left(\alpha_2+\beta_2\right)}{1-\theta\left(\alpha_1+\beta_1\right)}$$

is the general equilibrium AR(1) *coefficient*.

This result clarifies the conditions under which the model's dynamic system is guaranteed to display *uniform convergence*: where any temporary shocks anywhere in the system would eventually die out, even in full general equilibrium.[8] This occurs when $\rho^{GE} < 1$, or

$$\theta\left(\alpha_1+\beta_1+\alpha_2+\beta_2\right)<0 \tag{22}$$

holds, or (since $\theta > 0$) whenever the model's total net agglomeration forces (i.e. $\alpha_1 + \beta_1 + \alpha_2 + \beta_2$) are non-positive. That is, on any dynamic transition path, it will always be the case that the model economy is strictly converging in the global sense that the difference between the rate of population change in the most rapidly changing location, relative to that in the least rapidly changing location, is getting smaller with each time step.

As we might expect, the general equilibrium AR(1) coefficient is greater than the partial equilibrium AR(1) coefficient, i.e. $\rho^{GE} > \rho^{PE}$. This means that the economy as a whole may exhibit greater persistence, in the sense described above, than the (partial equilibrium) evolution of the population in any given location.

These results offer guidance concerning when and where we should expect slow persistence in geographic economies. In settings that mimic the partial equilibrium assumptions of Section 2.2.2—that is, where shocks to historical conditions in a location i have only minimal bearing on the location's own (relative) market access to other locations—speeds of convergence are given by expression (18). However, in the more general case of regions that interact in a meaningful way the expression in Proposition 2 becomes the place to turn. Comparing the two expressions, one can see the hazards of partial equilibrium thinking, since one could easily have a setting in which the partial equilibrium AR(1) coefficient is below one, indicating convergence, and yet the general equilibrium AR(1) rate bound is larger than one, indicating the possibility of divergence. Such a tension could in principal be the source of discrepancy across studies in the speeds of convergence estimated, since some studies may approximate well the partial equilibrium ideal of Section 2.2.2 and others may not.

2.2.4. Path dependence

Finally, we turn to the concept of path dependence. As discussed above, we define this term, as in prior work, as a concept entirely twinned with that of multiple steady states. That is, a path-dependent effect of some temporary shock in the past occurs when the shock has permanent effects on the outcome of the system. This is only possible if the shock caused the economy to end up in a distinct steady state.

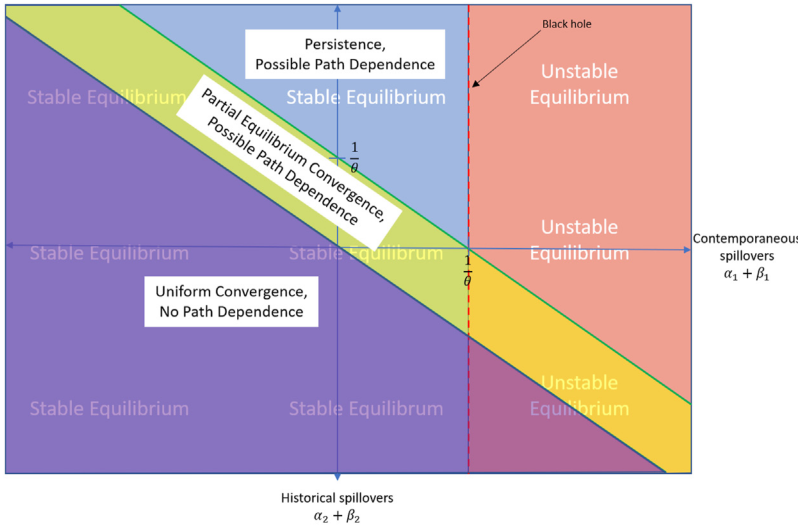
In order to discuss steady states and their potential multiplicity we consider the case in which all exogenous conditions in the model are constant at some arbitrary values, $\left\{\overline{A}_i>0,\overline{u}_i>0,\mu_{ij}>0\right\}$. Using this notation, and searching for constant values of the endogenous population variables L_{it} in equation (15), we find that steady-state equilibria will satisfy the expression

$$L_i^{1-\theta(\alpha_1+\beta_1+\alpha_2+\beta_2)}=\sum_j\mu_{ji}^{-\theta}\left(\overline{A}_i\overline{u}_i\right)^{\theta}L_j$$

for every location i. This is a particular example of the system of equations $z_i=\sum_j K_{ij}z_j^{\gamma}$ (where $z_i \equiv L_i^{1-\theta(\alpha_1+\beta_1+\alpha_2+\beta_2)}$ and $\gamma \equiv 1/\left(1-\theta\left(\alpha_1+\beta_1+\alpha_2+\beta_2\right)\right)$), for which it is well known (see e.g. Karlin and Nirenberg (1967); Zabreyko, Koshelev, Krasnosel'skii, Mikhlin, Rakovshchik, and Stetsenko (1975); Allen and Arkolakis (2014)) that there exists a unique solution if $|\gamma|<1$, resulting in the following proposition:

Proposition 3. *For any time-invariant geography* $\left\{\overline{A}_i>0,\overline{u}_i>0,\mu_{ij}>0\right\}$, *there exists a steady-state equilibrium and*

[8] We note that Proposition 2 is derived for time-invariant fundamentals such that the first term in (20) is zero. It is therefore suitable for applications in which, from the vantage point of period t, there can be an arbitrary path of changes in earlier fundamentals but such changes are temporary because the geography is constant from period t onwards. Alternatively, an augmented version of Proposition 2 for ongoing shocks could simply incorporate assumptions about the first term in (20).



Notes: This figure depicts the behavior of the equilibrium system as a function of the strength of the contemporaneous spillovers (summarized by $\alpha_1+\beta_1$) on the x-axis and the strength of the historical spillovers (summarized by $\alpha_2+\beta_2$) on the y-axis. See the text for a discussion of each area of the parameter space.

Fig. 2. When should we expect historical persistence and path dependence?.

that equilibrium is unique if

$$\left|\frac{1}{1-\theta(\alpha_1+\beta_1+\alpha_2+\beta_2)}\right| < 1. \tag{23}$$

Moreover, if instead $\left|\frac{1}{1-\theta(\alpha_1+\beta_1+\alpha_2+\beta_2)}\right| > 1$, *then there exist many geographies for which there are multiple steady states at each geography.*[9]

To interpret the condition in Proposition 3 it is useful to restrict attention to the case in which net total agglomeration forces are not so extreme as to imply $\theta(\alpha_1+\beta_1+\alpha_2+\beta_2) > 1$. This region also corresponds to that in which partial equilibrium convergence is guaranteed as in equation (19). In such a case, the condition for unique steady states in Proposition 3 simplifies to the requirement that (since $\theta > 0$) net total agglomeration forces are negative, or

$$\alpha_1+\beta_1+\alpha_2+\beta_2 < 0. \tag{24}$$

Three comments on this result are in order. First, as one might expect, the conditions for a unique steady state in Proposition 3 are more demanding than those for a unique dynamic equilibrium in Proposition 1. Unique equilibria obtain when the net contemporaneous spillovers $\alpha_1+\beta_1$ are small enough that $\alpha_1+\beta_1 < 1/\theta$. Unique steady states arise when the sum of both contemporaneous and historical externalities, $\alpha_1+\beta_1+\alpha_2+\beta_2$, is negative. This makes intuitive sense because while historical spillovers do not *per se* affect the payoffs of forward-looking agents who take the past as given, their strength does make it possible that the history-driven attractions of a location, even though taken as given by every generation of forward-looking migrants, are grounds enough for migrants to move there and this sort of behavior can become a self-fulfilling resting-place for the economy.

Second, the condition for a unique steady state in equation (24) is the same as that for uniform convergence in equation (22). This need not have been the case in general, as there are dynamic systems which converge uniformly to one of several steady states and dynamic systems with a unique steady state where convergence is not uniform. In this dynamic system, however, that the total agglomeration forces are net dispersive is sufficient to guarantee both a single steady state and that the economy will converge to that steady state uniformly for all initial conditions.

Third, the goal of Proposition 3 is to state a sufficient condition that holds for any exogenous features of the economy (its distribution of initial conditions and its geography). Clearly this could not be a necessary condition under such generality. However, Proposition 3 does provide assurance that it is the weakest such general sufficient condition because there do always exist geographical conditions under which steady states are indeed multiple when condition (23) is violated.[10]

2.3. Summary

Stepping back, we offer here a summary of the results of the previous section. These findings highlight the distinction between historical spillovers and contemporaneous spillovers. To illustrate this, Fig. 2 summarizes the results of the previous subsections in the two-dimensional slice of the model's parameter space described by the strength of contemporaneous $(\alpha_1+\beta_1)$ on the x-axis and historical $(\alpha_2+\beta_2)$ spillovers on the y-axis.

First, as per Proposition 1, multiplicity happens in this model only when contemporaneous spillovers $\alpha_1+\beta_1$ equal the dispersion parameter $1/\theta$. Second, when $\alpha_1+\beta_1 > 1/\theta$, equilibria are unique but unstable. We therefore think of a plausible range of parameter values as one where contemporaneous spillovers may exist, but are not so strong as to exceed the dispersion parameter—that is, where $\alpha_1+\beta_1 < 1/\theta$. This is the region to the left of the dashed vertical line in Fig. 2.

Third, as per equation (18), in partial equilibrium settings the rate of population persistence will be shaped by the AR(1) parameter $\rho^{PE} = (\alpha_2+\beta_2)/[\frac{1}{\theta}-(\alpha_1+\beta_1)]$, which hence combines the effect of historical spillovers $\alpha_2+\beta_2$ scaled up by the extent to which contemporaneous spillovers approach the stable/uniqueness threshold of

[9] In the case $\left|\frac{1}{1-\theta(\alpha_1+\beta_1+\alpha_2+\beta_2)}\right| = 1$, there is at most one (up to scale distinct) steady-state equilibrium.

[10] It is straightforward to find examples of multiplicity with as few as two locations when the condition of Proposition 3 does not hold, see Karlin and Nirenberg (1967); Allen and Donaldson (2020) show there are a continuum of such examples of multiplicity even in the presence of trade costs with as few as four locations.

$\alpha_1 + \beta_1 < 1/\theta$ from below. So we expect slower persistence whenever either contemporaneous or historical spillovers are large. Fourth, while persistence in general equilibrium settings is more complicated, Proposition 2 shows how a particular system-wide AR(1) process is guaranteed to converge whenever $\alpha_1 + \beta_1 + \alpha_2 + \beta_2 < 0$. That is, when total (contemporaneous plus historical) spillovers are negative, the economy will exhibit uniform convergence. Combining these observations, the region in purple in Fig. 2 denotes that with general equilibrium convergence, whereas that in green highlights the wider region in which only partial equilibrium convergence would be guaranteed.

Finally, as per Proposition 3, this same condition guarantees the existence of a unique, stable steady state; but when it fails the possibility of multiple steady states arises (and is actually guaranteed to arise under a range of geographic conditions). This latter scenario opens up the possibility of temporary events having permanent consequences if they happen to tip the economy's spatial orientation from one steady state to another—a case often referred to as path dependence. Returning to Fig. 2, this possibility arises in either the blue or green regions.

Combining these observations, we see that a particularly rich and yet tractable set of dynamic phenomena can occur in this model when contemporaneous spillovers $\alpha_1 + \beta_1$ are negative, and yet historical spillovers $\alpha_2 + \beta_2$ are large enough that the sum of both sets of spillovers $\alpha_1 + \beta_1 + \alpha_2 + \beta_2$ is positive. This would guarantee that the economy's dynamic equilibrium is unique and stable, and yet that convergence is not guaranteed to be uniform and that multiple steady states may exist. An intriguing possibility within this space is where the total spillovers also satisfy $\alpha_1 + \beta_1 + \alpha_2 + \beta_2 < 1/\theta$, as in the green region of Fig. 2. This parameter range would exhibit partial equilibrium convergence and yet could still feature steady-state multiplicity and path dependence.

3. Extensions and empirical quantification

The model in Section 2 contained a number of simplifications—a particular restriction on taste parameters, no trade costs, agents who are effectively myopic, and agglomeration spillovers that were restricted to an isoelastic form. We begin in Section 3.1 by discussing more realistic versions of each of these aspects of the model. Section 3.2 then outlines a method that can be used to estimate the model's parameters.

3.1. Extensions to the simple model

3.1.1. Departing from parameter restriction (12)

Returning to the model of endogenous fertility in Section 2.1.3, suppose that the idiosyncratic taste dispersion parameter θ and the child-rearing costs elasticity λ do not satisfy the restriction $\frac{1}{\lambda-1} = \theta$.[11] This results in a system of two "blocks" of equations (with one equation for each of the N locations i and T time periods t) that needs to be solved for endogenous variables in each period t (taking lagged values of these variables as given). In particular, this system can be written as follows:

$$L_{it} W_{it}^{-\theta} = \sum_j \mu_{jit}^{-\theta} \left(\Pi_{jt}\right)^{\frac{1}{\lambda-1}-\theta} L_{j,t-1}, \tag{25}$$

$$\Pi_{it}^{\theta} \equiv \sum_j \mu_{ijt}^{-\theta} W_{jt}^{\theta}, \tag{26}$$

and where, recall, $W_{it} = \overline{A}_{it}\overline{u}_{it} L_{it}^{\alpha_1+\beta_1} L_{i,t-1}^{\alpha_2+\beta_2}$. When $\frac{1}{\lambda-1} = \theta$ this collapses to the single block of equations (i.e. equation (13)); but otherwise there will be two interacting blocks, since the block expressed by (25) contains the endogenous populations $\{L_{it}\}$ and expected utility terms $\{\Pi_{it}\}$, as does the block expressed by (26).[12] As detailed in Allen and Donaldson (2020), the essence of the results in Section 2 continue to go through in this multi-block case, albeit in a form that includes the more involved cross-block interactions. For example, the analogs of Propositions 1 and 3 would now emphasize sufficient conditions involving λ, as well as θ, α_1, and β_1. And the general equilibrium notion of convergence expressed in Proposition 2 would place a bound on the rate of convergence of a vector system of not just χ_{Lt}, but also of $\chi_{\Pi t}$ (i.e. the ratio of the maximum change in Π_{it} across space to the minimum such change). However, if we were to extend the notion of partial equilibrium discussed in Section 2.2.2 to mean, in this extended case, that Π_{it} is additionally held constant, then equation (18) would continue to describe the rate of partial equilibrium convergence.

3.1.2. Costly trade and differentiated products

A second extension to the simple model of Section 2 would be to allow for locations that make differentiated products that are potentially sold to any other location at a variable cost τ_{ijt}. One version of this is the case of Armington differentiation, in which adults have CES preferences (with elasticity of substitution $\sigma > 1$) over the goods produced in all locations, though this has well-known microfoundations that are richer at the micro-level but equivalent for our (aggregate) purposes. This introduces two complications. First, the nominal wage w_{it} at a location is now a function of not just the location's productivity but also the endogenous demand for this location's product derived from consumption at all other locations. Second, the cost of living (a CES price index that we denote by P_{it}) now differs across locations because of trade costs, and so the systematic component of utility—the amenity-adjusted real wage—now satisfies $W_{it} = u_{it} w_{it}/P_{it}$. As Allen and Donaldson (2020) show, one way to summarize the equilibrium (assuming goods market clearing and balanced trade every period) co-determination of w_{it} and W_{it} is via two additional blocks of equations, given by

$$w_{it}^{\sigma} L_{it}^{1-\alpha_1(\sigma-1)} = \sum_j K_{ijt} L_{jt}^{\beta_1(\sigma-1)} W_{jt}^{1-\sigma} w_{jt}^{\sigma} L_{jt}, \tag{27}$$

$$w_{it}^{1-\sigma} L_{it}^{\beta_1(1-\sigma)} W_{it}^{\sigma-1} = \sum_j K_{jit} L_{jt}^{\alpha_1(\sigma-1)} w_{jt}^{1-\sigma}, \tag{28}$$

where $K_{ijt} \equiv (\tau_{ijt} \overline{A}_{it}^{-1} \overline{u}_{jt}^{-1} L_{i,t-1}^{-\alpha_2} L_{j,t-1}^{-\beta_2})^{1-\sigma}$. Equations (25) and (26) continue to hold, despite the introduction of trade costs, so the model becomes a four-block system. However, under the additional restriction that trade costs are symmetric (i.e. $\tau_{ijt} = \tau_{jit}$ holds for all locations and time periods), as may be plausible in some applications, blocks (27) and (28) collapse into one (e.g. the dependence on nominal wages w_{it} can be eliminated). Allen and Donaldson (2020) report analogous results to Propositions 1-3 for this extended model, as a function of σ, θ, and the spillover parameters.[13] Unlike the single block system of equations above, with multiple blocks, uniqueness conditions now depend on the spectral radius of a matrix of coefficients, where the size of the matrix corresponds to the number of blocks in the system.

3.1.3. Forward-looking behavior

The model in Section 2 features agents who make their location decisions in adulthood, but where adulthood lasts, by assumption, just one period. They therefore have no reason to be forward-looking. However,

[11] This case is considered in Allen and Donaldson (2020), in which fertility is exogenous and hence $\lambda \to \infty$.

[12] Formally, one could consider equations (25) and (26), together with the expression $W_{it} = \overline{A}_{it}\overline{u}_{it} L_{it}^{\alpha_1+\beta_1} L_{i,t-1}^{\alpha_2+\beta_2}$, as representing a three-block system in the period t endogenous variables, $\{L_{it}\}$, $\{\Pi_{it}\}$, and $\{W_{it}\}$. However, since the relation between W_{it} and L_{it} involves no interactions across regions it is a trivial block, just as it was in the simplified model of Section 2.

[13] Allen and Donaldson (2020) consider the case of exogenous fertility, but introducing finite λ (as in the model of Section 2) would modify the conditions stated in that paper in a straightforward manner.

in Allen and Donaldson (2020) we describe an extension to this setup in which adults live forever—from their birth period until the economy's end after a finite number, T, of periods—and are fully forward-looking throughout that lifetime. This amounts to (following microfoundations outlined in, for example, Artuç et al., 2010, Caliendo et al., 2019, and Balboni, 2019) replacing the dependence of migration decisions on the single-period payoff W_{it} in equation (5) with a forward-looking version V_{it}, defined as the value of (remaining) life when based in location i at time t. In particular, V_{it} in this extended model is given by

$$V_{it} = W_{it}\Pi_{i,t+1}^{\delta}, \tag{29}$$

which combines the period payoff W_{it} in t with the expected value, discounted by $\delta < 1$, of starting out in i at the beginning of period $t + 1$. Technically, this introduces dependencies across "blocks" of equations defined by their time periods—so this introduces T additional blocks. But since the dependencies across blocks are limited in nature (the t block in equation 29 interacts with the $t + 1$ block only via the dependence on $\Pi_{i,t+1}$, not any other variable in $t + 1$ or any other variable in periods more than one ahead), tractability is preserved. As is intuitive, the result of this addition is that raising δ (from zero, the value it implicitly takes in Section 2) will shift inwards both the purple and blue regions of Fig. 2; intuitively, this shift inward arises from the fact that different expectations of future spatial configurations can become self-fulfilling, despite weaker spillovers, if future periods matter more to the agents in the model (a force highlighted by e.g. Krugman, 1991 and Matsuyama, 1991).

3.1.4. Spillovers with varying elasticities

A final extension that we consider involves the possibility that agglomeration spillovers in production and/or consumption do not take the isoelastic forms suggested in equations (1) and (4). This is potentially important for several reasons. First, the above functional form assumptions are clearly restrictive and it is valuable to understand what happens under alternative scenarios. Second, a long tradition in urban modeling considers cities whose size is finite because increasing returns to scale in local production are eventually overcome by increasing congestion externalities—unlike in our model where these two elasticities are constant and locational sizes are typically interior (even with $\alpha_1 + \beta_1 > 0$) because of the dispersion force of preference heterogeneity (captured by θ).

To consider the case of varying elasticities, suppose for simplicity that $\alpha_2 = 0$, and amend equation (1) such that it takes the generalized form of $A_{it} = \overline{A}_{it} f_i(L_{it})$. Notably, the function $f_i(\cdot)$ is now not only unrestricted but is also free to vary in arbitrary ways across locations. Applying the tools in Allen et al. (2020) one can show that the results in Section 2 go through if we replace their dependence on the elasticity α_1 with, instead, $\overline{\alpha}_1 \equiv \max_{i,L_{it}} \frac{d \ln f_i(L_{it})}{d \ln L_{it}}$. That is, for example, the sufficient condition for uniqueness of steady states in Proposition 1 involves the maximum value that production elasticities can take (across all locations and all possible levels of production). This basic principle applies to our full baseline model as well as to the extensions of it discussed above.

3.2. Empirical quantification

What the results summarized in Fig. 2 highlight is that in order to assess the possibilities of uniqueness, of slow persistence (in both the partial and general equilibrium senses), and of multiple steady states and hence the possibility of path dependence, one needs to know the strength of contemporaneous (α_1, β_1) and historical (α_2, β_2) spillovers separately. While the contemporaneous spillovers (especially α_1) have been the focus of a large literature (Duranton and Puga, 2004; Combes and Gobillon, 2015; Rosenthal and Strange, 2004), much less attention has been devoted to the historical spillovers (α_2 and β_2), especially over the sort of time lag (a generation) that is natural in the model above.

We briefly describe here a simple method—an extension of the Rosen-Roback spatial equilibrium estimation tradition (Rosen, 1979; Roback, 1982; Glaeser, 2008)—that can be used to estimate these parameters. As anticipated, this draws on the (inverse) labor demand and supply equations (2) and (10), which we repeat here for convenience:

$$\ln w_{it} = \alpha_1 \ln L_{it} + \alpha_2 \ln L_{i,t-1} + \ln \overline{A}_{it}, \tag{30}$$

$$\ln w_{it} = \left(\frac{1}{\theta} - \beta_1\right) \ln L_{it} - \beta_2 \ln L_{i,t-1} + \frac{1}{\theta} \ln IMMA_{it} - \ln \overline{u}_{it}. \tag{31}$$

Starting with equation (30), we think of this as an estimating equation in which appropriate data on w_{it}, L_{it}, and $L_{i,t-1}$ are available for a group of locations i and in at least one time period t. However, the productivity term $\overline{A}_{it}$ is unobserved. Because prices and quantities are codetermined in the system of equations (30) and (31), simultaneity bias would generically afflict OLS estimates of α_1 (and hence, generically, α_2 as well) even if the error term (comprising $\overline{A}_{it}$) were purely exogenous. The standard solution is to seek an instrumental variable (IV) that enters equation (31) but is excluded from (30). One natural source of such IVs would consist of observed locational characteristics that plausibly affect amenities $\overline{u}_{it}$ but not productivity $\overline{A}_{it}$.[14] But other options can derive from observable components of migration frictions μ_{ijt} in location i and/or elsewhere, and both contemporaneous and lagged values of productivity and amenities in locations other than i. Finally, an analogous discussion applies to the supply equation (31), where $\overline{u}_{it}$ is unobserved and where observed components of contemporaneous productivity shifters $\overline{A}_{it}$ can serve, in principle, as valid instruments. One distinction in this case, however, is that (log) migration market access $\ln IMMA_{it}$ must be controlled for, though doing so offers the possibility of identifying θ, as is necessary to identify β_1 from the coefficient on $\ln L_{it}$.[15] Another approach to estimating θ would draw on the bilateral migration equation (5).

We are unaware of attempts to estimate equations (30) and (31) in the forms given here. However, an important literature—surveyed in Rosenthal and Strange (2004) and Combes and Gobillon (2015)—has devoted much attention to the estimation of production spillovers, where a typical range of estimates might span 0.03–0.08. Such estimates would correspond to α_1 in cases where $L_{i,t-1}$ is either controlled for or orthogonal to L_{it}, but they would correspond to $\alpha_1 + \alpha_2$ in cases where an economy is approximately in steady state (and hence no opportunities to isolate separate effects of L_{it} and $L_{i,t-1}$ are available). Regarding the amenity spillovers, β_1 might be thought to stand in—admittedly, in a highly reduced-form manner—for the effects of unmodeled immobile local goods that are in fixed supply (such as land). Under this interpretation one would expect a negative value for β_1. The parameter β_2, however, would then be positive if investments in local un-modeled factors made in the past are still durable a generation later. Finally, regarding θ, one recent intra-national estimate of such a migration elasticity parameter (though admittedly not necessarily of the inter-generational sort in the model here) is that from Monte et al. (2018), who obtain $\theta = 3.30$.[16]

These values provide a back-of-the-envelope sense for where we could expect estimated versions of typical economies to lie in Fig. 2.

[14] One important caveat with such an approach is that the attraction of more workers to high amenity places may increase land prices, potentially causing firms to substitute toward labor and away from land in production, thereby affecting labor productivity; see Combes and Gobillon (2015) for a discussion and potential remedies.

[15] Allen and Donaldson (2020) provides further discussion concerning strategies for controlling for migration market access.

[16] A closely related literature estimates the speed of adjustment of location-specific labor supply (including via migration) to local wage changes (as well as changes in local housing prices and unemployment rates). See, for example, Topel (1986), Blanchard and Katz (1992), or Beaudry et al. (2014).

For example, suppose that: (i) estimates of productivity spillovers take the value of 0.08 and this corresponds to the value of $\alpha_1 + \alpha_2$ (i.e. is estimated from economies that are close to steady state); (ii) contemporaneous amenity spillovers derive from un-modeled housing, which accounts for about one-third of expenditure (and hence to the value $\beta_1 = -0.33$); and (iii) historical amenity spillovers also derive from such housing, which is approximately fully durable over a generation (so $\beta_2 = 0.33$). At these hypothetical parameter values we have $\alpha_1 + \alpha_2 + \beta_1 + \beta_2 > 0$. If we further believe that contemporaneous and historical productivity spillovers are both non-negative (or simply that contemporaneous spillovers are bounded by $\alpha_1 < 0.41$) then we can expect the model economy to lie in the green region of Fig. 2—with partial equilibrium convergence but the scope for possible path dependence—as long as θ takes any value below 12.5.

4. Concluding remarks

Our goal in this paper has been to offer a simple take on the complicated dynamic phenomena that can arise in economic geography models—often these models' very *raison d'être*—yet to do so in settings that feature enough spatial granularity, heterogeneity and frictions that they could form the basis for empirical quantification. We have taken several shortcuts along the way, assuming, *inter alia*: free trade in a homogeneous good, a particular relationship between preferences for locational diversity and preferences for child-rearing, and a particular form of myopia in dynastic planning. However, the benefit of these restrictions is that a guide to the system's properties is summarized by three parameters (as in Fig. 2): equilibria are unique and stable when contemporaneous spillovers are weaker than dispersion forces; convergence occurs when the sum of both contemporaneous and historical spillovers is weaker than dispersion forces (in partial equilibrium) and weaker than zero (in general equilibrium); and the multiplicity of steady states that is at the root of path dependence can (and often will) arise when the sum of all spillovers is positive. The extended tools described in Allen and Donaldson (2020) provide ways to obtain analogous results while relaxing these restrictions.

Our focus here has been on the positive dynamic properties of spatial models. Turning to normative properties, it is certainly expected, given heterogeneity in locational fundamentals, that when steady states are multiple they generate different levels of aggregate welfare. However, it also seems likely that exactly the settings where fundamental heterogeneity is large (and hence aggregate welfare could differ most across steady states) are those for which such heterogeneity shuts down the possibility of multiple steady states in the first place; indeed, Lee and Lin (2018) document precisely such a phenomenon in the neighborhoods of US cities. Such an intuition is an expression of Rauch's (1993) question—"Does history matter only when it matters little?"—which strikes us as one of the most important open questions in historical urban and regional economics.

Our earlier work (Allen and Donaldson, 2020) provides one window into this question by deriving bounds on the welfare gaps that could exist across all possible steady states in a given setting in order to delimit the possibilities, as well as simulations in which relatively minor spatial perturbations do lead to large aggregate welfare consequences. The findings of Michaels and Rauch (2018) also highlight a setting where a region's historically-driven city configuration becomes substantially suboptimal ex-post relative to a counterfactual region that had opportunities (over many centuries) to start from (relative) scratch. But a full understanding of when and where spatial persistence and path dependence are consequential—as well as the policy implications that would naturally follow—is a central goal for the historical work of the future.

Author statement

The authors declare that they contributed equally to the development of the manuscript.

Declaration of competing interest

The authors declare that they have no conflicts of interest to report.

References

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. Q. J. Econ. 129 (3), 1085–1140.
Allen, T., Arkolakis, C., Li, X., 2020. On the Equilibrium Properties of Network Models with Heterogeneous Agents. NBER working paper no. 27837.
Allen, T., Donaldson, D., 2020. Persistence and Path Dependence in the Spatial Economy. NBER working paper no. 28059.
Artuç, E., Chaudhuri, S., McLaren, J., 2010. Trade shocks and labor adjustment: a structural empirical approach. Am. Econ. Rev. 100 (3), 1008–1045.
Balboni, C.A., 2019. In Harm's Way? Infrastructure Investments and the Persistence of Coastal Cities. MIT working paper.
Banerjee, A., Iyer, L., 2005. History, institutions, and economic performance: the legacy of colonial land tenure systems in India. Am. Econ. Rev. 95 (4), 1190–1213.
Barro, R.J., Becker, G.S., 1989. Fertility choice in a model of economic growth. Econometrica 57 (2), 481–501.
Beaudry, P., Green, D.A., Sand, B.M., 2014. Spatial equilibrium with unemployment and wage bargaining: theory and estimation. J. Urban Econ. 79, 2–19.
Blanchard, O.J., Katz, L.F., 1992. "Regional Evolutions," Brookings Papers on Economic Activity, vol. 1.
Bleakley, H., Lin, J., 2012. Portage and path dependence. Q. J. Econ. 127 (2), 587–644.
Bleakley, H., Lin, J., 2015. History and the sizes of cities. Am. Econ. Rev.: Papers & Proceedings 105 (5), 558–563.
Bosker, M., Brakman, S., Garretsen, H., Schramm, M., 2007. Looking for multiple equilibria when geography matters: German city growth and the WWII shock. J. Urban Econ. 61 (1), 152–169.
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.
Combes, P.-P., Gobillon, L., 2015. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), "The Empirics of Agglomeration Economies," in Handbook of Regional and Urban Economics, vol. 5. Elsevier, pp. 247–348.
Davis, D., Weinstein, D., 2002. Bones, bombs, and break points: the geography of economic activity. Am. Econ. Rev. 92 (5), 1269–1289.
Davis, D., Weinstein, D., 2008. A search for multiple equilibria in urban industrial structure. J. Reg. Sci. 48 (1), 29–65.
Dell, M., 2010. "The persistent effects of Peru's mining mita. Econometrica 78 (6), 1863–1903.
Deneckere, R.J., Judd, K.L., 1992. In: Benhabib, J. (Ed.), "Cyclical and Chaotic Behavior in a Dynamic Equilibrium Model, with Implications for Fiscal Policy," in Cycles and Chaos in Economic Equilibrium. Princeton University Press.
Desmet, K., Nagy, D.K., Rossi-Hansberg, E., 2018. The geography of development. J. Polit. Econ. 126 (3), 903–983.
Desmet, K., Rossi-Hansberg, E., 2014. Spatial development. Am. Econ. Rev. 104 (4), 1211–1243.
Duranton, G., Puga, D., 2004. Microfoundations of urban agglomeration economies. In: Handbook of Regional and Urban Economics, vol. 4. North-Hol.
Duranton, G., Puga, D., 2014. The growth of cities. Handb. Econ. Growth 2, 781–853.
Duranton, G., Puga, D., 2019. Urban Growth and its Aggregate Implications. NBER working paper no. 26591.
Fujita, M., Krugman, P., Venables, A.J., 1999. The Spatial Economy: Cities, Regions, and International Trade. MIT Press.
Glaeser, E.L., 2008. Cities, Agglomeration and Spatial Equilibrium. Oxford University Press.
Glaeser, E.L., Gyourko, J., 2005. Urban decline and durable housing. J. Polit. Econ. 113 (2), 345–375.
Glaeser, E.L., Kerr, S.P., Kerr, W.R., 2015. Entrepreneurship and urban growth: an empirical assessment with historical mines. Rev. Econ. Stat. 97 (2), 498–520.
Hanlon, W., 2017. Temporary shocks and persistent effects in urban economies: evidence from British cities after the U.S. Civil war. Rev. Econ. Stat. 99 (1), 67–79.
Hanlon, W., Heblich, S., 2021. History and urban economics. Reg. Sci. Urban Econ. (forthcoming).
Henderson, V., Squires, T., Storeygard, A., Weil, D.N., 2018. The global spatial distribution of economic activity: nature, history, and the role of trade. Q. J. Econ. 133 (1), 357–406.
Hornbeck, R., Keniston, D., 2017. Creative destruction: barriers to urban growth and the great boston fire of 1872. Am. Econ. Rev. 107 (6), 1365–1398.
Karlin, S., Nirenberg, L., 1967. On a theorem of P. Nowosad,. J. Math. Anal. Appl. 17 (1), 61–67.

Kim, S., Margo, R.A., 2014. In: Henderson, J.V., Thisse, J.-F. (Eds.), "Historical Perspectives on US Economic Geography," in Handbook of Regional and Urban Economics, vol. 4. Elsevier, pp. 2981–3019.
Kline, P., Moretti, E., 2014. Local economic development, agglomeration economies and the big push: 100 years of evidence from the Tennessee valley authority. Q. J. Econ. 129, 275–331.
Krugman, P., 1991. History versus expectations. Q. J. Econ. 106 (2), 651–667.
Lee, S., Lin, J., 2018. Natural amenities, neighborhood dynamics, and persistence in the spatial distribution of income. Rev. Econ. Stud. 85 (1), 663–694.
Lin, J., Rauch, F., 2021. What future for history dependence in spatial economics? Reg. Sci. Urban Econ. 103628 (forthcoming).
Liu, E., Redding, S.J., Kleinman, B., 2021. Dynamic Spatial General Equilibrium. NBER working paper no. 29101.
Matsuyama, K., 1991. Increasing returns, industrialization, and indeterminacy of equilibrium. Q. J. Econ. 106 (2), 617–650.
Michaels, G., Rauch, F., 2018. Resetting the urban network: 117-2012. Econ. J. 128 (608), 378–412.
Monte, F., Redding, S.J., Rossi-Hansberg, E., 2018. Commuting, migration and local employment elasticities. Am. Econ. Rev. 108 (12), 3855–3890.
Nagy, D.K., 2020. Hinterlands, City Formation and Growth: Evidence from the US Westward Expansion. UPF working paper.
Nunn, N., 2008. Slavery, inequality, and economic development in the americas: an examination of the engerman-sokoloff hypothesis. In: Helpman, E. (Ed.), Institutions and Economic Performance. Harvard University Press, pp. 148–180.
Nunn, N., 2014. In: Aghion, P., Durlauf, S. (Eds.), "Historical Development," in Handbook of Economic Growth, vol. 2. Elsevier, pp. 347–402.
Ottaviano, G.I., 2001. Monopolistic competition, trade, and endogenous spatial fluctuations. Reg. Sci. Urban Econ. 31 (1), 51–77.
Ottaviano, G.I., Tabuchi, T., Thisse, J.-F., 2002. Agglomeration and trade revisited. Int. Econ. Rev. 43 (2), 409–436.
Rauch, J.E., 1993. Does history matter only when it matters little? The case of city-indu try location. Q. J. Econ. 108 (3), 843–867.
Redding, S., Sturm, D., Wolf, N., 2011. History and industrial location: evidence from German airports. Rev. Econ. Stat. 93 (3), 814–831.
Redding, S.J., Rossi-Hansberg, E.A., 2017. Quantitative spatial economics. Annual Review of Economics 9 (1).
Roback, J., 1982. Wages, rents, and the quality of life. J. Polit. Econ. 1257–1278.
Rosen, S., 1979. In: Mieszkowski, P., Straszheim, M. (Eds.), "Wage-Based Indexes of Urban Quality of Life," in Current Issues in Urban Economics. Johns Hopkins University Press, pp. 74–104.
Rosenthal, S.S., Strange, W.C., 2004. In: Henderson, J.V., Thisse, J. (Eds.), "Evidence on the Nature and Sources of Agglomeration Economies," in Handbook of Urban and Regional Economics, vol. 4. Elsevier, pp. 2119–2172.
Topel, R.H., 1986. Local labor markets. J. Polit. Econ. 94 (3), 111–143.
Voigtlander, N., Voth, H.-J., 2012. Persecution perpetuated: the medieval origins of anti-semitic violence in nazi Germany. Q. J. Econ. 127 (3), 1339–1392.
Voth, H.-J., 2020. "Persistence – Myth and Mystery". CEPR working paper no. 15417.
Zabreyko, P., Koshelev, A., Krasnosel'skii, M., Mikhlin, S., Rakovshchik, L., Stetsenko, V., 1975. Integral Equations: A Reference Text. Noordhoff International Publishing Leyden.