The unintended consequences of increasing returns to scale in geographical economics

Steven Bond-Smith[†]

May 2021

Abstract

Increasing returns to scale is the basis for many powerful results in economics and economic geography. But the limitations of assumptions about returns to scale in economic growth theories are often ignored when applied to geography. This leads to an unintentional bias favoring scale and mistaken conclusions about geography, scale and growth. Alternatively, this bias is used as a convenient modelling trick by urban economists to describe agglomeration economies for innovation without examining the spatial mechanisms that actually create agglomeration economies. I discuss techniques to focus on the distinctly geographic mechanisms that define returns to scale at appropriate spatial scales.

JEL codes: O41, R11

Keywords: Increasing returns to scale, endogenous growth, scale effects, knowledge spillovers.

1 Introduction

The varying *scale* of economic activity is a striking characteristic of economic geography. People and businesses cluster in ever larger cities. Innovation, the engine of economic growth, is even more concentrated (Audretsch and Feldman, 1996; Bettencourt et al., 2007). It is unmistakable that scale has a fundamental role explaining local differences in productivity, innovation and growth. Understanding exactly *how* scale characteristics lead to regional differences requires explaining the explicit mechanisms that generate *returns* to scale. Returns to scale is the basis for many powerful results in economics and economic geography. *Increasing* returns to scale recognizes that production of outputs increases in *greater* proportion than required inputs. In economic geography and urban economics, returns to scale translate into forces of concentration, dispersion and spatial sorting that balance to determine the spatial organization and scale characteristics of economies (Proost and Thisse, 2019). Increasing returns to scale is also fundamental to endogenizing growth. Firstly, it means that productivity depends on the stock of knowledge rather than its division between people (Romer, 1990). And secondly, endogenizing growth requires returns to some *unit* scale in the ideas production function, so that innovation is affected by effort (Jones, 1999).

[†]*Curtin Business School, Bankwest Curtin Economics Centre, Curtin University, GPO Box U1987, Perth WA 6845, Australia; E-mail address: steven.bond-smith@curtin.edu.au*

For many helpful comments and suggestions I am very grateful to the editors Frédéric Robert-Nicoud and Harald Bathelt, two anonymous referees, Philip McCann, Jakob Madsen, Katsufumi Fukuda and seminar participants at the University of Western Australia. Your insights have clarified any misunderstandings and resulted in a much improved paper.

But an approximation that knowledge has a single dimension leads to an implicit assumption that ideas production has returns to the *aggregate* scale, otherwise known as the "scale effect". This is a well-known *limitation* of first generation endogenous growth theories most publicised by Jones (1995b). It implies that an increase in any rival factor endowment in the economy results in a higher growth rate without any microfoundation for such an increase. If any rival factor is growing, it implies an *ever-increasing* growth rate and explosive output in finite time. Time-series data on the growth of inputs to R&D for the United States is also not consistent with the functional form of ideas production in these first generation models of endogenous growth (Jones, 1995b). Returns to scale are indeed characteristic of innovation and endogenous growth, but the scale effect is widely recognised as *an error in aggregation* (Laincz and Peretto, 2006).

While the scale effect has been overwhelmingly rejected by scholars of (aspatial[1]) endogenous growth (Bond-Smith, 2019), scale is widely recognized by scholars of innovation and geography as a key characteristic of innovation and growth (Audretsch, 2003). *This is not a paradox.* Instead the mechanisms that generate increasing and decreasing returns to scale in geographical economics, which translate into spatial forces, determine a distribution of activity *within* the spatial economy but not necessarily an effect on the *aggregate* economy. In order to reconcile local, global and aggregate scale characteristics, researchers should carefully consider the specific appropriate spatial and economic scales in which returns to scale occurs. Avoiding the impact of the scale effect in these first generation models models requires a very limiting assumption of *no population growth* to avoid explosive growth rates. Unfortunately, growth models from (aspatial) economics have been used in economic geography without recognising the spatial constraints of this limiting assumption, with *unintended* consequences.

Unsatisfied with the limitations of assuming no population growth, a dispute between researchers of (aspatial) economic growth has continued over three decades about the appropriate modelling technique to negate the scale effect (Bond-Smith, 2019). This dispute implies two types of solutions. Firstly, Jones (1995a); Kortum (1997); Segerstrom (1998) suppose that ideas become more difficult to find as the simplest ideas are found first. This eliminates the implicit assumption of returns to the aggregate scale *in the long run*, but allows returns to scale in the short run to endogenize growth in response to research effort. Unfortunately in spatial models, the implicit returns to scale assumption that causes the scale effect and the diminishing returns to ideas assumption that is supposed to neutralize it, unintentionally translate into uneven forces of concentration and dispersion respectively—as spatial models are supposed to do—but *without any spatial foundation.* This is because the limiting approximation that knowledge has a single dimension sustains the aggregate scale effect in the short run where spatial forces operate. Alternatively, Young (1998); Peretto (1998); Dinopoulos and Thompson (1998); Howitt (1999) suppose that the important factor for endogenizing innovation is returns to *relative* scale in the product line. To achieve this, it is assumed that knowledge has two dimensions: variety and quality. As a result, increases in aggregate rival factors of production fragment across new product lines, removing the implicit limitation of increasing returns in the aggregate scale in both the short and long run without affecting returns to relative scale within product lines as required to endogenise growth. Hence I describe these models as "scale-neutral". In spatial models, this approach still eliminates the scale effect because returns to relative scale in the product line has no aggregate spatial implications. This allows deliberate spatial mechanisms included in the model to explicitly and intentionally define the spatial scales of returns to scale that directly define spatial forces in the spatial economy. Of course knowledge is far more complex than these simplistic one and two-dimensional approximations, but the approximation of knowledge must be sufficiently sophisticated that returns to scale occurs at accurate spatial and economic scales.

While spatial economists attempt to describe regional growth with over-arching spatial equilibrium models, Michael Storper (2010) describes in this journal that "*much work from economic geographers is to go in the opposite direction from the spatial economists, emphasizing that spatial economic development is the result of unique, context-driven, place-specific combinations of forces that, as a consequence, can neither be*

modelled nor even subject to large-scale causal inquiry. At best, they can sometimes be captured in descriptive typologies." There is an opportunity for a middle-ground, in which spatial growth models incorporate deliberate mechanisms that capture some elements of the context-driven and place-specific focus of geographers. Such models require otherwise scale-neutral models of growth that do not contain implicit biases regarding scale. Urbanization, MAR and Jacobs' externalities determine returns to scale mechanisms for a firm's R&D facilities (Ciftci and Cready, 2011), the aggregate scale of a city's R&D facilities (Bettencourt et al., 2007; Bettencourt et al., 2007), the aggregate scale of a country's R&D investment (Luintel and Khan, 2017; Curtis et al., 2020) and the scale local or national portfolios of industries (Henderson, 1997; de Groot et al., 2016). Such externalities also vary by industry and over time (Neffke et al., 2011; Diodato et al., 2018).

In order to explore this, I use a series of toy models to understand the role of assumptions about scale and growth in geographical economics. These "toy" models are the most parsimonious specifications that generalise both endogenous growth and spatial growth theories. The models cited are of course far more nuanced and sophisticated than is articulated in this analysis, but the parsimonious model most clearly shows how *implicit returns to scale assumptions about growth have unintended consequences for economic geography.*[2]

This article serves a number of purposes. Firstly, I highlight a fundamental misunderstanding of increasing returns to scale and the scale effect in endogenous growth theory when applied in a spatial context. I emphasize the importance of mechanisms that generate returns to scale in geographic research that translate into forces for concentration, diffusion and spatial sorting, such that spatial economists must be very careful about the scales in which increasing returns applies. In a series of toy models, I explore the spatial consequences of assumptions about scale effects, and techniques to remove scale effects. I review the broad economic geography literature for examples of the use and misuse of the implicit assumptions about returns to scale that are borrowed from aspatial endogenous growth but overlook its spatial limitations. I aim to reduce the risk that economic geographers or spatial economists inappropriately apply aspatial economic theory to spatial phenomena by offering techniques to avoid unintended spatial consequences. Lastly, I highlight that this article is but one example of the mistaken use of theoretical models and I encourage a best practice approach to economic theory that justifies all of its assumptions and notes its limitations.

2 Returns to scale assumptions and approximations

Assumptions matter. All theoretical models use assumptions to approximate relationships and avoid the complexity of a truly realistic but intractable model. Economists use theoretical mathematical models to organise their assumptions about the world and draw hypotheses about economic phenomena. Empirical research assumes an implicit underlying theory, even when a theory is not explicitly defined. Some assumptions are intended to approximate true facts about the world, but many are entirely unrealistic and used purely for simplicity. Assumptions eliminate specific variables so research can focus on what matters. Many assumptions don't matter. Some assumptions eliminate negligible factors or factors that are of no interest. And some assumptions are simply "modelling tricks" to ensure the model is tractable, internally consistent or behaves in a way that reflects stylized facts. For example, the Constant Elasticity of Substitution (CES) function is often required to achieve tractable results in models of geography and trade or a balanced growth path in models of endogenous growth. As a result, the CES function has become a workhorse tool for models of growth, trade and geography. For all of these types of assumptions, realism is secondary to parsimony. Conclusions drawn from these assumptions are indirect because they need to account for the real world context that is absent in the model. Theorists and empiricists alike, must understand the limitations of their assumptions to avoid drawing hypotheses and conclusions based on unintended effects. *When assumptions are used to directly draw findings*, those assumptions are crucial and should at least represent a realistic foundation on which to draw strong conclusions. In economic geography, such a foundation is also spatial.

In the most extreme case, the instrumentalist approach (e.g. Milton Friedman (1953)) values a theory by its predictive accuracy only, rather than whether its assumptions also correspond to the real world. Economic theory is often criticized that its assumptions are unrealistic (Martin and Sunley, 1998; Romer, 2015), but opposition to instrumentalism is not the main argument in this article. Instead, this article describes a contagion from implicit assumptions about growth that interact with geographical economics—in ways that were never intended—to mistakenly draw false conclusions in both theoretical and empirical research about the spatial nature of economic growth. In this way, even the instrumentalist argument is defeated by its own standards when theories are applied in other contexts, such as the spatial application of endogenous growth theory. While the focus in this article is scale and geography, it is part of a broader narrative about the unintentional effects of overlooking the limitations of modelling assumptions to draw mistaken conclusions.

In its simplest form increasing returns to scale is caused by a fixed cost of production or scale economies that are internal to the firm, but increasing returns can also create external economies. That is, returns to scale apply to specific (e.g. internal and external) scales such as the scale of a firm, industry, city or aggregate economy. Mechanisms that imply returns to scale are powerful tools to approximate real world economic and spatial relationships. Like Blaug (1980) and Boland (1992), this is partly an argument for more realism in economic theory—specifically for these *critical* assumptions of returns to scale in geographical economics—in order to understand the distinct mechanisms that actually generate returns to scale and translate into forces in the spatial economy. While simplifying assumptions are necessary to approximate stylized facts or enhance tractability, they imply limitations that must be accounted for. In this way, if implicit assumptions about returns to scale and growth are used, it's direct conclusions can be appropriately qualified, rather than used to define strong conclusions about relationships with scale. Before examining spatial growth models explicitly, I first explain the scales at which returns to scale assumptions apply to growth; geography and trade; and their combination.

2.1 Returns to scale in endogenous growth theory

Returns to scale is essential for understanding growth. Specifically, endogenous growth theory emphasized the role of non-rival knowledge spillovers for endogenous investment in innovation in return for temporary monopoly profits (Romer, 1990). Non-rivalry implies increasing returns to *firm scale* production because an idea can be used multiple times to produce outputs. A business can double production by duplicating its factory, but it can more than double production if its second factory uses a new design. The firm has constant returns to scale in its rival factors of production, but it has increasing returns to scale in the rival and nonrival factors combined. This is crucial for growth, because per capita productivity depends on the stock of ideas, rather than its division between people. But for endogenous growth, returns to scale shows up twice: the scale of production of output *and* the scale of production of ideas. The ideas production function is explicitly assumed to have returns to *firm scale* research effort, which is crucial to *endogenizing* growth because it means that innovations are the result of deliberate research effort by firms to find new ideas. In this way, each business can make greater improvements by increasing its effort or investment in R&D.

Returns to scale is particularly unique in the ideas production function because ideas are both an output and non-rival input of innovation such that ideas multiply (Bond-Smith, 2019). While the ideas production function explicitly assumes constant or diminishing returns to firm scale rival production inputs, the non-rival factor is increasing endogenously over time. In first generation growth models (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992), this means that growth in the scale of any rival input also makes use of the same endogenously growing pool of non-rival ideas, implying increasing returns for innovation to the *growing aggregate scale of any rival input.* In other words, first generation models add an additional *implicit* assumption that innovations are produced with increasing returns to both the *firm scale* and the *aggregate scale* of any of the rival inputs.

Returns to the aggregate scale of the economy is problematic because population growth implies an *ever-increasing* growth rate to explosive levels of output in finite time. In first-generation endogenous growth models this is known as the scale effect: that any rival factor increasing in scale results in an *ever-increasing growth rate* because the non-rival factor is also increasing endogenously through this circularity of ideas. This illogical result has been widely recognised as a limitation of first-generation endogenous growth models that requires an assumption of no population growth. The scale effect also implies that a larger economy grows faster than a smaller economy but involves no spatial or micro-foundation for why this would be true: it is only a result of this implicit assumption that biases growth in favor of scale. Economic geographers and urban economists can strongly argue that such a result may well be true, due to agglomeration economies for innovation, but agglomeration externalities are the result of well-known microfounded mechanisms rather than an implicit assumption. This implies a further limitation of not examining scale differences if the underlying growth model contains the scale effect because it would mistakenly apply the benefits of agglomeration economies to scale differences rather than to their causal mechanisms. To be clear, returns to *firm scale* rival inputs to innovation is an approximation that is essential for endogenous growth, but requires appropriate limiting assumptions in the *aggregate scale* to avoid scale differences or scale growth. Clearly, this has implications for its use in economic geography.

The scale effect is eliminated in subsequent generations of semi-endogenous (Jones, 1995a; Kortum, 1997; Segerstrom, 1998) and Schumpeterian (Young, 1998; Peretto, 1998; Dinopoulos and Thompson, 1998; Howitt, 1999) growth models (See Bond-Smith (2019) for a summary). Semi-endogenous growth models counteract its effect by assuming diminishing returns to cumulative ideas. For innovation, this still implies returns to scale in the short run so that ideas are produced endogenously, but eliminates returns to scale in the *long run aggregate scale* where returns to scale and the ever-increasing flow of non-rival ideas cause the scale effect. In the balanced growth path the net effect implies constant returns because the scale effect is matched by diminishing returns to cumulative ideas. Schumpeterian models of endogenous growth without scale effects allow innovation to expand in two dimensions (Young, 1998; Peretto, 1998; Dinopoulos and Thompson, 1998; Howitt, 1999) such that aggregate increasing returns are tempered by the additional dimension. This implies constant (or increasing (Peretto, 2018)) returns for innovation at the *relative scale of the firm* so that idea production is endogenous. But it eliminates the scale effect from aggregate productivity growth because non-rivalry in the vertical dimension of innovation no longer interacts with aggregate scale in the horizontal dimension of the ideas production function. Instead, non-rivalry only interacts with the *relative scale of research effort* as required to endogenize growth. There is increasing returns to the scale of the firm, but there is now no aggregate scale relationship as greater scale is spread accross additional varieties. By clearly understanding the scales at which returns to scale should apply, Schumpeterian models enable increasing returns to the scale of the firm or industry to achieve stylized facts about endogenous innovation while eliminating returns to scale in aggregate. Rather than eliminating increasing returns to scale, these techniques are additional assumptions that change the scale in which returns to scale appears in the ideas production function.

2.2 Returns to scale in trade and geography

New trade theory and new economic geography recognize that increasing and decreasing returns to scale occurs over both aspatial scales (firm; industry; market; aggregate) and *spatial* scales (local; regional; urban; international; global). Krugman (1979) and Helpman and Krugman (1985) showed how this tool could be embedded into a general equilibrium model of trade that influenced the scale and location of production in order to explain that trade of similar goods between countries was the result of increasing returns to scale by differentiated producers. Krugman (1991) extended his trade model to economic geography by enabling the mobility of workers such that increasing returns to scale also had implications for the location of people

and firms.

In these models of trade or geography, a fixed cost and otherwise constant returns to scale in production, implies increasing returns to the scale of the firm's *local* production. That is, the firm's local output increases by more than its required inputs simply by increasing the scale of production without having to incur additional fixed costs. With trade, this enables a larger factory in each location to produce differentiated varieties, rather than requiring multiple factories in each location. In this way, increasing returns to the scale of the *local factory* has implications for the scale and spatial distribution of economic activity. Transport costs are a pecuniary externality related to distance that result in decreasing returns to the scale of the *local factory* as any increase in the scale of production must be exported to customers further away. For an exporting firm, trade costs also imply increasing returns to the scale of the *local market*, as less production needs to be exported since local sales are more profitable. Competition also leads to decreasing returns to the scale of *aggregate local and global export market* that balances increasing returns to the scale of the local factory. While transport costs imply decreasing returns at the local scale of the firm, with vertical linkages it also implies increasing returns to the scale of local relative aggregate production as the firm makes greater use of local suppliers. Venables (1999) disaggregated production such that industries divide into clusters based on these vertical linkages, implying increasing returns to the scale of *local supply chains*. Other mechanisms related to city scale such as public goods (Andersson and Forslid, 2003), amenities (Zhang, 2007) and congestion costs (Ottaviano et al., 2002; Fujishima, 2013) also generate increasing or decreasing returns to scale that imply additional concentration and dispersion forces.[3]

In these models, increasing returns to aggregate industry or economy scales implies productivity benefits for density or concentration at those scales while decreasing returns to aggregate scales implies a congestion cost and dispersion force. Notably, all of these instances of returns to scale in theories of trade and geographical economics (excluding growth) are the result of deliberately assumed (and empirically supported) mechanisms that imply these returns to scale relationships at very deliberate spatial scales. Carefully applied to the correct scales, mechanisms that lead to increasing and decreasing returns to appropriate scales offer powerful insights on the spatial economy.

2.3 Returns to scale in geography and growth

The spatial nature of knowledge spillovers (Audretsch and Feldman, 1996) and similarities in modelling techniques implied a suitable marriage between endogenous growth and new economic geography (Bond-Smith and McCann, 2014). The continued development of endogenous growth theories over four generations (Bond-Smith, 2019) is repeatedly followed by spatial versions to understand the geographic implications of endogenous growth. By understanding how returns to scale applies in either growth or geographical economics, assumptions from endogenous growth theory must be carefully examined when they are applied in spatial models. Models that combine growth and geography contain a labyrinth of implicit and explicit assumptions about increasing and decreasing returns to scale.

The limitations of first generation endogenous growth models discussed in Section 2.1 imply limiting assumptions of no population growth and no examination of scale differences. These limitations still apply to spatial applications but are typically overlooked by geographical economists. Specifically, the ideas production function in which ideas have a single dimension, is applied directly to a spatial model, without any consideration for its appropriate spatial scale. The key modification to the ideas production function is typically to assume that non-rival knowledge spillovers do not transfer perfectly between locations to determine the spatial distribution of knowledge inputs to research. If there is an implicit assumption of returns to the aggregate scale in the ideas production function (as in the scale effect in first generation and in the short run in semi-endogenous growth models), then these imperfect spatial externalities also unintentionally determine the degree of *aggregate* returns to scale for innovation. But there is no explicit micro-foundation

to believe that the degree of aggregate returns to scale in the ideas production function is specifically related to the spatial nature of knowledge spillovers: it is assumed implicitly by overlooking the approximation in first-generation growth models that knowledge has a single dimension only and its corresponding limitations. *As a result, the assumption of returns to the aggregate scale unintentionally interacts with spatial assumptions about knowledge spillovers.* This is in addition to the intended direct impacts of the imperfect spatial externality mechanism for knowledge spillovers.

When implicit assumptions about returns to scale from aspatial growth theories are mistakenly applied to inappropriate spatial scales in geographical economics, they unintentionally imply consequences for the spatial economy. In spatial models, an ideas production function with vertical returns to the aggregate scale unintentionally implies greater local innovation productivity for larger regions which amplifies any spatial factor that affects innovation productivity as a larger group of researchers make greater use of available local knowledge. This means spatial models which contain a first generation or semi-endogenous growth model imply agglomeration economies for innovation without specifying a causal mechanism. Second generation semi-endogenous growth models eliminate increasing returns to the aggregate scale in the long run only, but still sustain aggregate returns to scale in the short run and therefore unintentionally imply agglomeration economies for innovation in the short run. In the long run, this translates to an unintended level effect on incomes in larger scale cities. However, there are additional unintended consequences. These models negate the scale effect by assuming decreasing returns to the scale of cumulative ideas. Although this leads to the desired result of constant growth in an aspatial growth model, despite a growing population, in geographic space the assumption of decreasing returns to the cumulative scale of innovation *unintentionally* assumes decreasing returns to the local scale of ideas by interacting with non-rivalry while not changing the assumption of increasing returns to the local scale of innovation. Instead of eliminating the scale effect, in geographic models this implies a balance of concentrating research effort due to the scale effect implying a level effect on incomes and diminishing innovation implying ideas congestion from cumulative local knowledge, *without any spatial foundation* for such forces.

The assumptions about increasing returns to the aggregate scale *amplify*—and assumptions about decreasing returns to remove the scale effect *unequally diminish*—any spatial mechanisms in the model because they are applied to inappropriate spatial scales without a spatial- or micro-foundation, leading to mistaken conclusions about the spatial nature of growth. If such scale relationships can be logically justified then this implies that such mechanisms could be modelled explicitly, rather than implicitly embedded in the ideas production function.

Schumpeterian growth models without scale effects have the potential for a scale-neutral approach (Bond-Smith et al., 2018; Bond-Smith and McCann, 2019) that enables only the intentionally-included spatial mechanisms to affect the spatial equilibrium rather than implicit assumptions about growth and returns to scale. These models allow the inclusion of mechanisms that enable returns to scale for innovation at various scales, but require deliberate mechanisms to determine the spatial scales of returns to scale. For example, imperfect knowledge spillovers only determine the location of knowledge and its spatial dispersion as intended, rather than the degree of returns to local and global returns to scale.

2.4 Reterns to scale mechanisms

Although the scale effect has been overwhelmingly rejected by scholars of (aspatial) endogenous growth (Bond-Smith, 2019), scale is increasingly recognized by scholars of innovation and geography as a key characteristic of innovation and growth (Audretsch, 2003). While I argue that underlying growth models must be scale-neutral in geographical economics, I am not dismissing scale as a characteristic of geography, innovation or growth. There is strong evidence that scale is an important factor in the spatial distribution of economic activity, and innovation in particular. Indeed, the World Intellectual Property Organization

(2019) cites a number of economic forces that explain the concentration of innovation in urban hotspots related to pools of skilled labor, market scale and knowledge spillovers, as well as weaker dispersion forces. Scale has crucial but varied roles in economic geography (Krugman, 1991; McCann and Acs, 2011), firm size or industry structure (Acs and Audretsch, 1988; Tether et al., 1997). These roles are determined by many factors and mechanisms that economists and geographers both seek to understand. The extent that scale might also explain aggregate relationships at a country level, such as the recent experience of economic growth in China, or urban and regional level, such as the concentration of innovation in metropolitan areas (WIPO, 2019), would also invoke spatial mechanisms at appropriate nation or supra-regional scales that facilitate scale differences. It is these mechanisms which explain scale phenomena in economic geography and represent the opportunity in geographical economics if underlying growth models are scale neutral. To understand these mechanisms spatial research should avoid implicit assumptions embedded in the economic growth model by translating these known forces into distinctive spatial mechanisms that determine returns to scale at appropriate spatial scales which translate into concentration, dispersion and spatial sorting.

Implicit assumptions about ideas production with increasing returns to the aggregate scale of the local economy are often rationalized as "agglomeration economies". However, agglomeration economies are not magic. Agglomeration economies are the result of positive externalities related to scale exceeding negative externalities related to scale. Such positive externalities relate to specialization, diversity and competition (Glaeser et al., 1992; de Groot et al., 2016) as well as the amenities or infrastructure that can be provided with economies of scale. When negative externalities of scale exceed the positive externalities there are agglomeration diseconomies. Such negative externalities relate to land rents from competition for space, congestion, pollution, commuting time and overcrowding. For geographers and economists to truly understand appropriate spatial scales for returns to scale, these mechanisms should be assumed explicitly. Implicit assumptions about returns to local scales are not an explanation of agglomeration economies, but a control variable for unknown or unobserved agglomeration externalities at those scales.

Agglomeration externalities for innovation are typically categorised by Marshall-Arrow-Romer (MAR) (Glaeser et al., 1992; Henderson et al., 1995), Jacobs' (Jacobs, 1969) and urbanization externalities. Urbanization externalities refer to city-wide internal and external economies such as spatial sorting of highly educated workers (Eeckhout et al., 2014; Redding, 2020; Verginer and Riccaboni, 2021), centers of R&D (Bettencourt et al., 2007), national and international connectivity (Simmie, 2003) and infrastructure (Blind and Grupp, 1999; Aghion et al., 2013) that all facilitate innovative activity. MAR or Jabobs' externalities refer to internal and external economies within cities based on their industrial structure. Specifically, MAR externalities refer to the benefits associated with *specialized* cities such as labor market pooling to reduce matching costs (Duranton and Puga, 2004), vertical linkages (Venables, 1996, 1999) and knowledge spillovers (Combes and Duranton, 2006) within industries or along supply chains (Isaksson et al., 2016). Businesses located close to their suppliers can benefit from shared knowledge through frequent interactions or from knowledge spillovers through imitation and skills transfers between competitors. Jacobs' externalities refer to the benefits of industrial *diversification* or the benefits from combining knowledge across different industries. Serendipitous interactions between workers with different knowledge give rise to new product combinations and stable diversified demand reduces risk from the volatility of inputs and output prices. Similarly, the concept of relatedness (Hidalgo et al., 2018) captures that diversification is a spectrum of related and unrelated varieties (Frenken et al., 2007; Boschma and Frenken, 2009). Furthermore, complex economic activities also concentrate in cities (Balland et al., 2020). These examples of mechanisms and externalities that generate scale characteristics are not inconsistent with the rejection of scale effects by growth theorists. Instead this multitude of mechanisms ultimately imply a spatial distribution *within* the economy but not an effect across the *aggregate* economy.

3 Two-region endogenous growth models

Endogenous growth theories can be characterized into generations by the scale effect (Jones, 1999; Bond-Smith, 2019). Each generation is followed by its application to core-periphery or trade models to understand the spatial implications of growth. Using toy models, this section disentangles the unintended spatial consequences of assumptions about returns to scale and growth: (i) in first generation endogenous growth models; (ii) in semi-endogenous growth models; and (iii) in Schumpterian models without scale effects. Micro-details vary, but this analysis distils the essential elements. These toy models are the most parsimonious specifications that generalise both endogenous growth and spatial growth theories in order to clearly show the relationship between assumptions about returns to scale for innovation and conclusions in economic geography.

3.1 A simple growth framework

For each of the following models I first present the standard (aspatial) endogenous growth framework. I then add a spatial dimension by extending the model to two regions. For simplicity, I add only one spatial mechanism, *localized knowledge spillovers*, that weight knowledge spillovers by its original location. Other spatial mechanisms affecting the production of knowledge can be easily included and realistic models would obviously require multiple mechanisms. To keep the scope of the analysis focused, I only specify the portions of these models that are necessary to the main issue of concern.

In these models, final output is a composite good $Y_t = A_t^{\alpha} L_{Yt}$ where L_{Yt} is labor used in production, A_t represents technology or the stock of ideas and α represents the preference for technology improvement or the degree of increasing returns to scale in production. This function is usually the result of a CES aggregator of multiple intermediate varieties that is required to generate "balanced growth" in which all sectors grow at the same rate, but other aggregators are also possible. The CES aggregator is useful in these toy models, but not essential to the argument about unintended spatial forces in the model. A variable elasticity of substitution (VES) function would generate unbalanced growth paths in which some sectors or factors grow faster than others as an additional source of growth (Palivos and Karagiannis, 2010). But the argument regarding unintended spatial consequences would still apply. Physical and human capital are left aside in these toy models in order to focus on technological change.

New ideas are the result of research effort that builds on the stock of ideas to increase productivity. There is free entry for entrepreneurs to develop new ideas. Productivity growth from the flow of new ideas is given by the function $\dot{A}_t = f(L_{At}, A_t)$ where the dot indicates change in technology over time. This ideas production function assumes that new ideas are a function of the existing stock of ideas A_t (i.e. intertemporal knowledge spillovers) and the research effort to discover new ideas L_{At}. Along the balanced growth path, a constant share of labor ($s < 1$) is employed in research such that $L_{At} = sL_t$ where $L_t = L_{Yt} + L_{At}$.

In order to capture a geographical element in these toy models, analogous equations apply to two regions (referenced by home and foreign) with foreign variables denoted by a tilde. To implement the spatial mechanism of localized knowledge spillovers in two region models, the intertemporal knowledge spillover is also adjusted by the location of knowledge using the spillover parameter $\lambda \leq 1$ such that knowledge transfers imperfectly between researchers in different locations:

$$\dot{A}_t = f\left(L_{At}, A_t + \lambda \tilde{A}_t\right). \tag{1}$$

These toy models are used to understand the effect of assumptions about returns to scale and mechanisms that generate returns to local scale on the spatial forces affecting growth and the location of economic activity. Migration and transport costs are not specified, but the toy models here are flexible enough to accomodate many alternatives.

Dividing final output by population to find per capita output, $y_t = Y_t/L = (1-s)\,A_t^{\alpha}$. Taking logs and time derivatives of per capita output, equilibrium implies a per capita growth rate of

$$g_y = \alpha \frac{\dot{A}}{A_i} = \alpha g_A. \tag{2}$$

where g represents the growth rate of the variable in the subscript. Based on this consistent framework, the analysis in the following sections focuses on specifying the innovation production function $\dot{A}$ and the resulting growth rate of technology g_A.

3.2 First Generation endogenous growth

3.2.1 The aspatial model

In first generation models of endogenous growth (Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992), the function

$$\dot{A}_t = \gamma L_{At} A_t \tag{3}$$

describes the flow of productivity improvements where $\gamma > 0$ is a parameter for calibration. In this model, productivity Acan be interpreted as product variety (Romer, 1990) or quality (Grossman and Helpman, 1991; Aghion and Howitt, 1992). Output per capita is proportional to the stock of knowledge $y_t = \frac{Y_t}{L_t} = A_t\,(1-s)$. Taking the time derivative of output and rearranging, growth of per capita output is $g_y = \frac{\dot{y}_t}{y_t} = \gamma L_{At} = \frac{\dot{A}_t}{A_t} = \gamma s L$. The growth rate of technology $g_A = \frac{\dot{A}_t}{A_t}$ and per capita output are proportional to scale:

$$g_{At} = \gamma s L_t \tag{3a}$$

and $g_{yt} = \alpha\gamma s L_t$. The scale factor L in Equation 3a is the well-known '*scale effect*' limitation in first generation endogenous growth models, where per capita growth and technology growth are an increasing function of the scale of the economy L_t such that a growing population implies an ever-increasing growth rate.[4] Models with a scale effect *require* a limiting assumption that population is constant to generate a constant growth rate.

3.2.2 The regional model

There are many examples of first generation endogenous growth models extended to two regions or countries (for example Walz (1997); Martin and Ottaviano (1999, 2001); Baldwin and Forslid (2000a,b); Baldwin et al. (2001); Yamamoto (2003); Baldwin and Martin (2004); Davis (2009)). These models typically assume that there is a constraint on knowledge spillovers between firms that are geographically separated. Consider the innovation function in a two region model based on the aspatial growth model above:

$$\dot{A}_t = \gamma s L_t\,(A_{Wt})\,(n + (1-n)\,\lambda) \tag{4}$$

where L_t now represents the home-region workforce, $\lambda \leq 1$ describes how easily knowledge transfers between places, A_{Wt} represents global technology and n represents the share of the knowledge stock that was developed in the home region with an analogous equation describing innovation in the foreign region. Final output is a composite good made from traded intermediates so local output per capita is proportional to the global stock of non-rival knowledge $y_t = \frac{Y_t}{L_t} = A_{Wt}\,(1-s)$ where s and L_t now refer to local characteristics. Taking the time derivative and rearranging $g_y = \frac{\dot{y}_t}{y_t} = \gamma\left(sL_t\,(n + (1-n)\,\lambda) + \tilde{s}\tilde{L}_t\,(n\lambda + 1 - n)\right) = \frac{\dot{A_{Wt}}}{A_{Wt}} = g_{AW}$. As a result, the growth rates of the two regions equalise in the steady state to the global technology growth

rate,

$$g_{AW} = \frac{\dot{A_{Wt}}}{A_{Wt}} = \gamma \left(sL_t \left(n + (1-n)\,\lambda \right) + \tilde{s}\tilde{L}_t \left(n\lambda + 1 - n \right) \right). \tag{4a}$$

While the spatial knowledge spillover diminishes the use of spatially dispersed knowledge as intended, the scale effect unintentionally amplifies the impact of the spatial knowledge spillover.

To examine this closely, reconsider Equation 4a. The growth function is made up of two types of factors: (i) spillover factors

$$n + (1-n)\,\lambda \quad and \quad n\lambda + 1 - n \tag{5}$$

and (ii) scale factors

$$sL_t \quad and \quad \tilde{s}\tilde{L}_t. \tag{6}$$

Spillover factors result from assumptions about the economic geography of knowledge spillovers and scale factors result from assumptions about growth. It is easy to see that an increase in scale (L_t and $\tilde{L}_t$), holding all else constant, results in a higher growth rate. This is the original scale effect that a growing population results in an ever-increasing growth rate and a well-acknowledged limitation of first generation endogenous growth models. Now consider an increase in scale in only the home region (L_t). This causes three types of changes in the growth rate. Firstly, there is an increase in growth in both regions due to it's effect over time on increasing spillovers in the home region via the distribution of economic activity (n) which is itself a dynamic function of the relative scale of research. Secondly, there is a smaller decrease in growth due to the effect of decreasing spillovers in the foreign region over time also via the distribution of economic activity. The third effect is the local scale effect in which increasing home region scale increases home region innovation by greater than its proportional increase in population due to the implicit assumption about increasing returns to the aggregate scale that is now applied at some arbitrary local level. The first two changes are the intentional result of assumptions about economic geography when the distribution of research effort varies between regions. But the third is the *unintentional* result of the assumption about increasing returns to the aggregate scale applied at a local level without a spatial foundation. The scale effect untentionally *amplifies* local knowledge spillovers as a larger group of researchers make use of available knowledge.

To be clear, this means that implicit assumptions about returns to the aggregate scale for innovation in first generation models of endogenous growth *unintentionally* bias conclusions by exaggerating agglomeration economies for innovation without any foundation in economic geography. This is in addition to any specific spatial mechanisms that actually create agglomeration economies such as increasing returns to the scale of production, transport costs or the mechanism shown here: spatial externalities on the transfer of knowledge.

To avoid the scale effect, the aspatial model requires an assumption of no population growth. But in the spatial model the impact of the scale effect is not eliminated by assuming no population growth if scale differences occur at the regional level. Holding the distribution of technology (n) fixed and differentiating with respect to time, setting to zero and rearranging implies that constant technology growth requires no population growth *and no change in scale between regions.*

The other solution is to remove the spatial externality from the knowledge spillover so that the scale effect only acts at an aggregate scale where it is eliminated by assuming no population growth. That is, if there are global spillovers ($\lambda = 1$) firms do not unintentionally benefit from locating in the larger region. The unintentional scale effect in the Romerian growth model now applies only at the global level and can be ignored by assuming no population growth:

$$g_{AW} = \frac{\dot{A_{Wt}}}{A_{Wt}} = \gamma \left(sL_t + \tilde{s}\tilde{L}_t \right). \tag{7}$$

But the model is also no longer capable of examining the intentional specific channel of spatial externalities

on knowledge spillovers that create regional differences for innovation.

The scale effect is often disguised and its effect is mistaken for the spatial knowledge spillover because the distribution of n is correlated with relative population (i.e. setting L_t to $\frac{L_t}{L_t+\tilde{L}_t}$) in the long run. If s represents the share of the local population involved in research and firms are symmetric then the scale effect seems to disappear by replacing population with its equivalent in terms of numbers of firms:

$$g_{AW} = \frac{\dot{A_{Wt}}}{A_{Wt}} = \gamma\left(sn\left(n + (1-n)\lambda\right) + \tilde{s}\left(1-n\right)\left(n\lambda + 1 - n\right)\right) \tag{8}$$

because g_{AW} is not a function of scale. Yet the scale effect is still there, it is merely hidden by carefully selected parameters for local population (effectively $\frac{L_t}{L_t+\tilde{L}_t} = n$) that is correlated with the local knowledge stock so it can be easy to mistakenly claim that all spatial outcomes are the result of spatial knowledge spillovers. Even with restrictive assumptions, the scale effect re-emerges whenever a spatial mechamism is applied to innovation (such as $\lambda < 1$) because it exagerates differences in the local scale of rival factors that make use of knowledge spillovers. This *unintentionally* implies spatial forces as a result of implicit assumpitons about returns to scale, rather than assumptions founded in geographic space.

That is, to avoid the scale effect in the spatial model implies an even more restrictive assumption that any rival factor endownment growth is zero *in each region.* This would prevent standard spatial mechanisms such as footloose labor or capital from being included in the model. While increases in effort (s and $\tilde{s}$) increase growth as intended by endogenous growth models, shifts in any rival factor (an increase in one region and an equal decrease in the other) also still imply the unintentional scale effect at a local level, despite the restrictive assumptions. So a further restrictive assumption is required that s and $\tilde{s}$ are equal. This assumption is so restrictive that the model is now unable to examine relationships between growth and economic geography or trade. I have been unable to find a paper with a Romerian growth model that does not violate this restriction as soon as spatial mechanisms are included. The only option would be to clearly acknowledge that the scale effect is also a limitation on the conclusions that can be drawn regarding the spatial innovation economy and that this additional impact from scale should not be expected *because it is an unintentional consequence of modelling assumptions.* It can still be concluded that local scale might affect innovation and growth *but only as a result of specific mechanisms to facilitate scale relationships.* Such a weakness in these models is very concerning.

3.3 Semi-endogenous growth without scale effects

3.3.1 The aspatial model

Jones (1995b) showed that growth remained constant despite an increasing population and increasing research effort, refuting the predicted scale effect. To resolve this apparent paradox, Jones (1995a); Kortum (1997) and Segerstrom (1998) developed second-generation theories of endogenous growth without scale effects by diminishing innovation productivity for developing cumulative improvements. Technology A_t can be interpreted as aggregate product variety (Jones, 1995a), variety within sectors (Kortum, 1997) or quality (Segerstrom, 1998). In these models, the function

$$\dot{A}_t = \gamma L_{At} A_t^{\beta} \tag{9}$$

describes how productivity improvements diminish with cumulative discoveries. The parameter $\beta < 1$ implies that it is increasingly difficult to discover additional new ideas. As above, output per capita is proportional to the global stock of knowledge $y_t = \frac{Y_t}{L_t} = A_t\left(1-s\right)$ and growth in output per capita is found by taking the

time derivative and rearranging, $g_y = \frac{\dot{y}_t}{y_t} = \gamma L_{At} A_t^{\beta-1}$, which is again equal to the growth rate of technology:

$$g_A = \frac{\dot{A}_t}{A_t} = \gamma s L A_t^{\beta-1}. \tag{9a}$$

In this equation, a larger population implies a scale effect, as above, but the rate of growth is diminished by the increasing difficulty of cumulative technological advancement. A balanced growth path implies that technology growth is constant. Differentiating with respect to time finds the constant long run growth is

$$g_A = \frac{g_p}{\gamma (1-\beta)} \tag{9b}$$

where g_p is the population growth rate.That is, in the balanced growth path the scale effect of a growing population is eliminated exactly by diminishing innovation productivity with cumulative ideas to reach a constant growth rate that is proportional to population growth. Research effort has short run impacts on growth, but no impact on the long run growth rate because it is eventually matched by diminishing innovation productivity. While research effort is still an endogenous investment decision in return for temporary monopoly profits, these models attract the label "semi-endogenous" growth because research effort has no long run impact on growth. However, the scale effect on effort is not removed from semi-endogenous models, it is simply equal to the impact of decreasing returns to the scale of cumulative knowledge in the balanced growth path.

3.3.2 The regional model

The semi-endogenous growth model is again extended to two regions by assuming spatial externalities for knowledge spillovers (For example see Minniti and Parello (2011) or Fukuda (2017)). Consider the innovation function in the home region in a two region growth model:

$$\dot{A}_t = \gamma L_{At} \left(A_{Wt} \left(n + (1-n)\lambda \right) \right)^{\beta} \tag{10}$$

with an analogous equation for the foreign region. Taking the time derivative of per capita output and rearranging $g_y = \frac{\dot{y}_t}{y_t} = \gamma s L_t A_{Wt}^{\beta-1} \left(n + (1-n)\lambda \right)^{\beta} + \gamma \tilde{s} \tilde{L}_t A_{Wt}^{\beta-1} \left(n\lambda + 1 - n \right)^{\beta} = \frac{\dot{A}_{Wt}}{A_{Wt}} = g_{AW}$. As a result, the growth rates of the two regions equalise in the steady state to the global technology growth rate,

$$g_{AW} = \frac{\dot{A_{Wt}}}{A_{Wt}} = \gamma \left(s L_t A_{Wt}^{\beta-1} \left(n + (1-n)\lambda \right)^{\beta} + \tilde{s} \tilde{L}_t A_{Wt}^{\beta-1} \left(n\lambda + 1 - n \right)^{\beta} \right). \tag{10a}$$

As with first generation models, the scale effect re-emerges when spatial mechanisms are applied to innovation because it amplifies the impact of spatial factors. Examine the growth rate in Equation 10a. The function is now made up of (i) spillover factors

$$n + (1-n)\lambda \quad and \quad n\lambda + 1 - n, \tag{11}$$

and (ii) scale factors

$$sL_t \quad and \quad \tilde{s}\tilde{L}_t \ , \tag{12}$$

as in first generation growth models, as well as cumulative idea congestion at both the (iii) global

$$A_{Wt}^{\beta-1} \quad and \quad A_{Wt}^{\beta-1}, \tag{13}$$

and (iv) local scales.

$$(X)^{\beta} \quad and \quad \left(\tilde{X}\right)^{\beta}, \tag{14}$$

where X is the local spillover factor. Spillover factors result from assumptions about the economic geography of knowledge spillovers while scale and idea congestion factors result only from assumptions about growth. As in first generation models, the scale effect still applies in exactly the same way in the short run. It is easy to see that an increase in scale (L_t and $\tilde{L}_t$), holding all else constant, increases the growth rate. In the balanced growth path, the scale effect is matched by global idea congestion in the aspatial growth model, leading to constant growth, but this is not necessarily the case in the two-region model if there are differences in scale between regions. Specifically, idea congestion applies to the global scale of ideas with a parallel congestion factor on local spillovers to absorb those ideas, while scale factors apply only to the local scales of regions. As a result, semi-endogenous assumptions about growth imply counteracting spatial forces of agglomeration economies for innovation due to the scale effect (as described above) and diseconomies for innovation due to idea congestion. Indeed, these unequal forces affect the rate of invention of new varieties in each region during the transition path to the steady state in Minniti and Parello (2011) and Fukuda (2017).

As above, this implies strict limitations on the conclusions that can be drawn from the model that should be clearly acknowledged. If there are global spillovers, there is be no benefit from locating in the larger region. Setting $\lambda = 1$ leads to the standard growth rate where growth is proportional to global population and inversely proportional to cumulative knowledge since $0 < \beta < 1$:

$$g_{AW} = \frac{\dot{A_{Wt}}}{A_{Wt}} = \gamma A_{Wt}^{\beta-1}\left(sL_t + \tilde{s}\tilde{L}_t\right). \tag{15}$$

Assuming many of the same restrictive assumptions discussed in Section 3.2.2 above finally eliminates the impact of the scale effect at local scales. Population growth in aggregate is now required to sustain growth, but must apply equally to both regions. Unfortunately, these restrictions now also prevent the model from being a useful description of economic geography.

3.4 Schumpeterian endogenous growth without scale effects

3.4.1 The aspatial model

'Schumpeterian' models of endogenous growth allow ideas to expand in two dimensions: new varieties and variety-specific quality improvements. These models recognized that population growth leads to an increase in the variety of products whereas productivity relates to the quality of individual products. This avoids scale in the ideas production function by sharing a larger population across additional varieties.

In these models equations apply at the firm level. Production per firm is now given by $Y_{it} = A_{it}L_{Yit}$. If $F = \eta L$ represents the number of varieties at time $t$and θ describes preference for variety, then per capita production is given by $y_t = \frac{1}{\eta}F_t^{\theta-1}\overline{A}_{it}^{\alpha}(1-s)$ where

$$\overline{A}_{it} = \int_0^{F_t} \frac{A_{jt}}{F_t} dj \tag{16}$$

represents economy-wide *average* quality (or productivity) per variety.[5] As a result, variety growth is $\dot{F} = \eta\dot{L}$, and per capita output growth is now $g_y = (\theta-1)g_F + \alpha g_{\overline{A}} = (\theta-1)\eta g_L + \alpha g_{\overline{A}}$. If θ is set to one such that there is no preference for variety then growth is only due to changes in productivity such that $g_y = \alpha g_{\overline{A}} = \alpha g_{\overline{A}}$ as in the models above but for *average* technology rather than aggregate technology. The function

$$\dot{A_{it}} = \gamma L_{Ait}\overline{A}_{it} \tag{17}$$

describes the flow of quality improving ideas for each individual firm i where $\overline{A}_{it}$ also represents knowledge spillovers to firms. In this model growth is dependent on research effort at the firm or sector level and not aggregate scale. Rearranging the firm's quality improvement production function finds that the growth rate of average technology is

$$g_{At} = \frac{\dot{\bar{A}}_t}{\bar{A}_t} = \gamma s \frac{L_t}{F_t} = \gamma \frac{s}{\eta}. \tag{17a}$$

Technology growth is dependent on the share of labor devoted to research rather than the scale of factor components. The variety dimension of technology eliminates the implicit assumption of aggregate increasing returns to scale for innovation (the scale effect). Including a love of variety ($\theta > 1$) implies a so-called 'weak' scale effect on income levels but not a 'strong' scale effect on growth rates (Jones, 1999). Since variety is proportional to population, g_F is proportional to g_L which implies that a portion of growth is a result of population growth but not population size or scale. This weak scale effect is equivalent to the income benefits from greater division of labor. As a result, increased research effort by firms or sectors increases the growth rate of average technology, but increases in the scale of population only increase the number of varieties eliminating the scale effect from first generation models. Theoretical arguments (Peretto, 2018; Bond-Smith, 2019)[6] and the weight of empirical research (Zachariadis, 2003; Laincz and Peretto, 2006; Ha and Howitt, 2007; Ulku, 2007; Madsen, 2008; Madsen et al., 2010; Ang and Madsen, 2011; Venturini, 2012; Greasley et al., 2013)[7] now strongly support the Schumpeterian approach to modelling endogenous growth without scale effects.

3.4.2 The regional model

In the two region model knowledge spillovers to each firm are the spatially-weighted average of global productivity that is locally observed by the firm (See Davis and Hashimoto (2014))

$$\overline{A}_{it} = (n + (1-n)\,\lambda) \int_0^{F_t} \frac{A_{jt}}{F_t} dj, \tag{18}$$

where F_t is global variety but the function defining the flow of productivity improvements for the average firm is otherwise unchanged:

$$\dot{A_{it}} = \gamma L_{Ait} \overline{A}_{it}. \tag{19}$$

Per capita output in the home region reduces to $y_t = \frac{Y_t}{L_t} = F_t \bar{A}_{Wt}\,(1-s)$, where $\bar{A}_{Wt} = \frac{1}{F_t}\int_0^{F_t} nA_{it} + (1-n)\,\tilde{A}_{it} di$ now represents global productivity averages because final local production (or consumption) is a global composite. Differentiating with respect to time and rearranging, the growth of per capita output is $g_y = \frac{\dot{y}_t}{y_t} = \frac{\dot{\bar{A}}_{Wt}}{\bar{A}_{Wt}}$ and growth rates equalise between regions. Note that growth refers to change in the global productivity average and differs from local knowledge spillovers by a factor of $(n + (1-n)\,\lambda)$. Rearranging the quality improvement production function for home region firms and assuming symmetry finds that the growth rate of global technology is

$$g_{\overline{A}W} = \frac{\dot{\bar{A}}_{Wt}}{\bar{A}_{Wt}} = n\gamma \frac{s}{\eta}\,(n + (1-n)\,\lambda) + (1-n)\,\gamma \frac{\tilde{s}}{\eta}\,(n\lambda + 1 - n)\,. \tag{19a}$$

Setting $\lambda = 1$ leads to the standard aspatial technology growth rate

$$g_{\overline{A}W} = \frac{\dot{\bar{A}}_{Wt}}{\bar{A}_{Wt}} = \gamma \frac{s_W}{\eta} \tag{19b}$$

where $s_W = ns + (1-n)\,\tilde{s}$ represents the global share of workers employed in research. Growth is unaffected by the scale effect because technology production is neutral to scale. Similarly, in this toy model there is no dispersion effect from diminishing innovation productivity. As a result, the assumptions about growth have no effect on the spatial equilibrium. Instead, the spatial equilibrium is determined precisely by the spatial mechanisms that could be included in the model such as imperfect spillovers ($\lambda < 1$), transport costs, congestion costs, rental costs or any other direct mechanism rather than by assumptions about growth and returns to scale.

The growth function in Equation 19a.is now made up of two factors: (i) spillover factors:

$$n + (1-n)\,\lambda \quad and \quad n\lambda + 1 - n, \tag{20}$$

and (ii) effort factors:

$$s \quad and \quad \tilde{s} \tag{21}$$

Spatial implications result directly from assumptions about the economic geography of knowledge spillovers. Effort factors result from assumptions about growth but are now neutral to scale. As intended by endogenous growth theories, increased effort (s and $\tilde{s}$) increases the rate of productivity growth by increasing the rate of quality improvement per firm but changes in scale (L_t and $\tilde{L}_t$) *have no effect on growth* since any increases in L are matched by increases in F, represented by the calibration parameter η. Changes in local scale affect local shares leading to standard changes in the number of firms and associated spillover factors ($(n + (1-n)\,\lambda)\,and\,(n\lambda + 1 - n)$) as intended from the assumptions about the geography of knowledge spillovers but do not imply any other changes in the growth rate. Effort factors differ from scale factors (and relative scale factors) in first generation and semi-endogenous models in that *there is no additional effect from scale because there is no implicit assumptions about increasing returns to the aggregate scale.* The relationships between growth and scale in the model such as agglomeration economies or diseconomies for innovation are now only a result of the geographic mechanisms deliberately included in the model, *not a result of implicit returns to scale assumptions about growth.* This is an important distinction: *In the scale-neutral model, all spatial implications resulting in scale differences are the result of deliberate, micro-founded, assumptions about economic geography, whereas in first-generation and semi-endogenous models there are unintended spatial impacts from implicit assumptions about returns to scale and growth.* As a result, the Schumpeterian branch of endogenous growth theory enables research in economic geography without the many strict restrictions that apply to earlier generations of growth theory.

3.4.3 Implicit assumptions in the Schumpeterian model

However, even Schumpeterian models can be susceptable to unintentional spatial consequences if the knowledge production function contains *implicit* assumptions about returns to scale. In some versions of the Schumpeterian model, it is assumed that the number of firms is limited by overall research effort (rather than total labor) such that $F = \eta L_A$ (Young, 1998; Howitt, 1999). As a result, $g_{\overline{A}} = \gamma s \frac{L}{F} = \frac{\gamma}{\eta}$ so proportional R&D subsidies have no effect on the research effort of individual firms as induced research effort is used only to expand the number of varieties. While Young (1998) implies R&D subsidies should be based on the *intensity* of research effort, Young acknowledges and accepts this as a known limitation. Peretto (1998) and Dinopoulos and Thompson (1998) (and the model above) assign research effort only to quality improvement to retain the stylized fact that long run growth is influenced by proportional support for research effort. Alternatively, Howitt (1999) uses diminishing returns as a modelling trick by assuming that it becomes progressively more difficult to develop additional varieties for larger populations, which also retains the stylised fact that ideas are becoming harder to find, as in the semi-endogenous models of Jones (1995b); Bloom et al. (2020), now as a result of a larger population. Specifically, quality improving innovations

are subject to constant returns to scale but inventing new varieties is subject to decreasing returns. As a result, economy-wide productivity is not neutral to scale because a growing population implies a slowing rate of aggregate innovation. This idea can be shown in the above toy model by modifying the number of varieties such that $F = \eta L^{\beta}$ where $0 < \beta < 1$ but the ideas production function at the firm level remains Schumpeterian as in Equation 18. Average productivity growth at the firm level is now

$$\frac{\dot{\overline{A}}_{it}}{\overline{A}_{it}} = \gamma s \frac{L}{\eta \left(sL\right)^{\beta}} = \gamma s^{1-\beta} \frac{L^{1-\beta}}{\eta} \tag{22}$$

such that growth is influenced by the proportional R&D subsidy but slows as population increases because new varieties are increasingly difficult to find reducing the knowledge spillover for larger populations. This 'inverse' scale effect is the result of an assumption of decreasing returns to scale as a modelling trick to reflect two stylized facts only rather than a micro-founded spatial mechanism. Variety growth is constantly proportional to population growth $g_F = \beta g_L$ and also affects per capita growth if there is a preference for variety. Unlike the semi-endogenous models, growth remains Schumpeterian overall and positive, even in absence of population growth.

In the two region variant of Howitt (1999) (Davis and Hashimoto, 2015)

$$g_{At} = \frac{\dot{\overline{A}}_{Wt}}{\overline{A}_{Wt}} = \gamma \frac{s}{\eta} \left(\left(n + \left(1 - n\right)\lambda\right) L\right)^{1-\beta} + \gamma \frac{\tilde{s}}{\eta} \left(\left(n\lambda + 1 - n\right) \tilde{L}\right)^{1-\beta} \tag{23}$$

with analogous equations for the foreign region. Effort factors are now modified by the power $1-\beta$ implying that there is now a third factor in the growth function: (iii) local congestion of ideas:

$$\left(L\right)^{1-\beta} \quad and \quad \left(\tilde{L}\right)^{1-\beta}. \tag{24}$$

where L and $\tilde{L}$ are local scale factors, but Howitt's assumption was only intended to apply to the global scale to reflect particular stylized facts. This mistaken application of diminishing returns to the local scale rather than global scale unintentionally implies a new local ideas congestion factor due to assumptions about diminishing returns and growth. As a result, the spatial equilibrium is now a balance between spatial forces cause by the intentional spillover factors and the *unintentional* congestion factor.

Davis and Hashimoto (2015) argue that product development costs may be higher in larger markets to justify these spatial implications. But spatial models are also the appropriate tool to directly model the spatial mechanisms that lead to differences in product development costs rather than rely on implicit assumptions in the innovation production function. Such a mechanism would specify the details about exactly how costs are affected. Taking the explanations proposed by Davis and Hashimoto (2015) for differences in product development costs, the cost of rent, wages and congestion or any other spatial product development cost difference can form specific spatial mechanisms that generate diminishing returns to scale for innovation without implicit assumptions. Each mechanism translates into different policy implications to optimize the innovation system. The explicit inclusion of spatial mechanisms leads to conclusions that are directly tied to the spatial mechanisms that actually cause spatial phenomena in the model. In such a model the stylized facts targeted by Howitt (1999) become a result of spatial externalities. .

4 Discussion

The unintended consequences of implicit assumptions about returns to scale and growth extend beyond the parsimonius two-region toy models explained in Section 3. Any spatial models that rely on growth theories

with implicit assumptions about aggregate returns to scale, also derive unintended effects for spatial dynamics that result in mistaken conclusions misinterpreting the relationship between scale and growth. City-scale may well be a predictor of growth dynamics in various contexts, but urban economists and economic geographers limit their findings if implicit assumptions are simply assumed to facilitate or amplify the spatial mechanisms that lead to scale economies. In this section I explore how these implicit assumptions about scale effects and growth affect the broader economic geography literature and how research practices can be improved.

4.1 Broader impact of implicit assumptions about returns to scale on spatial research

To focus on trade, a number of *trade models* restrict migration and make other assumptions that limit the impact of the scale effect (Davis, 1998; Baldwin et al., 2001; Baldwin and Forslid, 2000b, 2010; Minerva and Ottaviano, 2010; Baldwin and Harrigan, 2011; Breinlich et al., 2014), but such limiting assumptions are not always distinctly defined or even sufficient. With labor mobility between sectors, economies of scale for innovation in the ideas production function still affects the spatial economy by amplifying any spatial mechanism for innovation when regions specialise.

Glaeser (2003) defines the subfield, the *New Economics of Urban and Regional Growth* focusing on empirical urban economics with spatial equilibrium equalising utility across space. While linearising an endogenous growth model implies that city scale is a determinant of productivity growth, spatial equilibrium assumes that the benefits of scale or density are offset by something else such as housing prices that would imply no role for scale in urban growth. Much of this research finds that urban growth of big cities tends to be at the same as many small cities, but the result stems from an assumption of spatial equilibrium in which spatial forces have equalised, while retaining an implicit assumption of increasing returns to scale for innovation. As a result, assumed increasing returns to scale for innovation are immediately extracted by dispersion forces. Dispersion forces are modelled distinctly, but the concentration forces that determine the benefits of agglomeration economies for innovation are based on *implicit* assumptions. As a result, these models have no insight on the determinants of agglomeration economies for innovation at all.

Urban and regional economics models of growth typically combine first generation endogenous growth theories into models of cities (for examples see Duranton (2006, 2007)), in part because the implicit assumption of increasing returns to the aggregate scale of innovation (the scale effect) provides a useful modelling trick for assuming agglomeration economies for innovation. But first generation models *inadvertently* predict that innovation (and productivity growth) increases with the aggregate scale of city-wide research effort without modelling any specific mechanisms for why such agglomeration economies occur. In this way, growth models in urban economics deny many mechanisms that support agglomeration economies for innovation such as sectoral clustering (Duranton and Puga, 2005; Bond-Smith and McCann, 2019) or skill based sorting (Behrens et al., 2014; Davis and Dingel, 2019). To be clear, the scale effect is explicitly assumed to be the source of external economies associated with an agglomeration mechanism (as in Black and Henderson (1999)), without actually modelling these urban mechanisms. As a result, these models draw strong conclusions about the benefits of agglomeration without understanding its causes. This implicit bias continues in very recent research. For example, Duranton and Puga (2019) use a Romerian ideas production function at the city scale to conclude that massive increases in the scale of major American cities, and declines in rural and regional city populations, would result in greater productivity growth. Of course this is the conclusion when they assume it to be true but include no mechanism for why. Similarly, Arkolakis et al. (2020) model the role of immigrants in American growth by assuming a semi-endogenous model of growth applied at city scales, concluding that immigration restrictions hindered the benefits of scale effects. Indeed, this result confirms the finding above in Section 3.3.2 that scale effects in the spatial context are not eliminated by the semi-endogenous model yet they draw strong spatial conclusions based on implicit assumptions about

increasing and diminishing returns in the growth model that have no foundation in the spatial economy at all.

Two recent articles are much more careful about the scales at which increasing and decreasing returns apply to innovation (Desmet et al., 2018; Aloi et al., forthcoming). Desmet et al. (2018) avoid the impact of the scale effect by not implementing any spatial externalities for knowledge spillovers and assuming no global population growth. While these are appropriate limiting assumptions, it means that the model is no longer useful for understanding many mechanisms from economic geography research on innovation. Aloi et al. (forthcoming) instead define a scale-neutral Schumpeterian growth model so that the spatial mechanisms can be explicitly defined without any implicit scale bias, as I recommend in this article.

City scale could also still be predictive of innovation or growth in particular contexts. I am not arguing against its use as a model parameter, but its interpretation. City population may be used as a control variable for unobserved spatial factors that generate returns to scale at urban scales which are unrelated to the variable of interest, but it is of limited use for drawing *strong* conclusions about specific agglomeration mechanisms or regional innovation production functions. Unfortunately, these strong conclusions are seen in the empirical urban economics literature wherever agglomeration economies for innovation are merely assumed with city size as a regressor (for example, Glaeser and Gottlieb (2009)). Such conclusions are the consequences of implicit assumptions about increasing returns to scale in the growth model, which sidelines the externalities that cause agglomeration economies for innovation. Models in urban economics are the ideal tool for examining these causes, but only if appropriate assumptions are used as a basis for research.

Quantitative spatial economics as defined by Stephen Redding and Esteban Rossi-Hansberg (2017) offer a series of recent articles using dimensional space (i.e. along a line or a plane). While most of these models are static, and avoid modelling growth, a subset of this research led by Klaus Desmet and Esteban Rossi-Hansberg utilize a first generation engine of endogenous growth (See for example Desmet and Rossi-Hansberg (2009, 2010, 2012, 2014); Desmet et al. (2018); Nagy (2020)). The authors acknowledge that the scale effect is a limitation in terms of world population growth, but increasing returns to the aggregate scale for innovation is also an implicit agglomeration factor that amplifies spatial mechanisms at locations in space, rather than modelling the causes of agglomeration economies directly. Desmet et al. (2018) suggest that one solution is defining the cost of innovation as an increasing function of world population, which acts as a modelling trick so that the model is neutral to global scale, but does not neutralise implicit returns to *local* scale where the ideas production function applies. The scale effect would remain an implicitly assumed mechanism for agglomeration without modelling the externalities that cause increasing returns to agglomeration.

Lastly, the *empirical regional science* literature uses spatial econometric models to understand spillovers between reigons (Lesage and Fischer, 2008). By adding spillovers the interpretation of "scale" is somewhat differentbut could lead to misleading conclusions about the role of proximity to neighbours. In order to estimate the contribution of various factors to growth, a log transformation of first generation endogenous growth models implies a bias that favors the size of a region and its neighbours combined as a significant explanatory variable without explaining a cause. Empirical analysis that favours this approach may either leave out or over-emphasise key predictors that are correlated with scale. Taking this approach, Izushi (2008) tests a spatial Romerian model that includes the stock of local R&D workers as an explanatory variable. As I would expect, Izushi finds that this is likely to be a mis-specification because the local stock of R&D workers is ultimately a result of how spatial units are defined. Instead Izushi finds in favour of a scale-neutral spatial version of Lucas (1988).

4.2 Avoiding unintended consequences

Economic geographers and spatial economists often seek to examine the explicit spatial mechanisms that create internal and external economies for innovation as a key determinant of the geography of economic

activity. Therefore it is essential that the unintended internal and external economies caused by implicit assumptions about returns to scale are eliminated from underlying models or, as a minimum, are appropriately qualified as a limitation of the analysis.

Unintended spatial consequences can be avoided by carefully applying assumptions that cause increasing and decreasing returns to the appropriate scales. In this way, aspatial aspects can be modelled in a manner which is neutral to space such that spatial consequences are only a result of intentional spatial mechanisms. Similarly, empirical work can be interpreted correctly in terms of intentional spatial mechanisms rather than implicit scale assumptions. The spatial or aspatial nature and analytical purpose of assumptions should be clearly defined. Assumptions used for analytical convenience should be examined closely to have no unintended impact on direct conclusions. Assumptions that do not meet these strict requirements should be avoided, and if they cannot be avoided, clear limitations must be specified regarding the conclusions that can be drawn.

Implementing a scale-neutral approach requires a careful examination of assumptions. Aspatial mechanisms (e.g. growth) must be carefully checked for their unintentional spatial implications. In this way, the conclusions drawn can be explicit about the source of spatial phenomena. Researchers should be explicit about the spatial mechanisms in the model and their micro- and spatial-foundations. Any microfounded argument to assume a scale or inverse-scale relationship should be modelled directly as a microfounded spatial mechanism and should not be implicitly assumed in the knowledge production function.

Spatial conclusions should be directly connected with their causal spatial mechanisms. A requirement to identify the spatial mechanism(s) highlights any spatial consequence that is otherwise unintentionally caused by implicit assumptions. This provides an opportunity to revise the underlying model when unintended spatial consequences are identified. If such a revision is not possible, models with unintended spatial consequences do not have to be disregarded entirely, but affected conclusions can be treated appropriately as limitations of the particular model that is otherwise used for analytical convenience. This allows a focus on the intentional spatial conclusions and their implications.

5 Concluding remarks

Scale and growth interact through very specific spatial channels, but currently, research on growth and geography, is limited by *implicit* assumptions about returns to the aggregate scale for innovation that bias findings in favor of scale. Crucially, theorists and empiricists can use deliberate mechanisms and channels to carefully understand the different scales in which increasing and decreasing returns to scale applyand capture the specific externalities that cause spatial phenomena. There is a particular opportunity for interdisciplinary research to build from the nuanced descriptive typologies in the human economic geography literature by incorporating the many specific mechanisms that contribute to the uniquely local characteristics of innovation in the spatial economy.

While many of these mechanisms imply agglomeration economies, the disentanglement of the various mechanistic causes may overturn the findings of existing research supporting agglomeration. For example, it enables the possibility of greater productivity growth in many specialized cities and reductions in inter-city transport costs that would substantially alter conclusions in Duranton and Puga (2019); Arkolakis et al. (2020). Rather than the recommended urbanization policy targeting settlement in only superstar cities and expanding immigration, this would promote a national and regional transport and connectivity infrastructure programme and specialised regional priorities for innovation (Escobari et al., 2019; Daboín et al., 2019)..

While the main argument in this article is about the foundations of spatial economic theory, perhaps the more important conclusion is with respect to empirical research. There may well still be a scale relationship where scale interacts with innovation in some way that implies growing regional disparities based on scale differences. But economic geography and urban economics are the perfect vehicles for examining the spatial

mechanisms driving such scale relationships. Empirical research should justify assumptions about scale and test different assumptions about scale as a robustness check. In doing so, findings regarding agglomeration economies are far more nuanced than simply 'bigger is better'. For example, the mechanisms that enable related industry clustering in peripheral regions (Bond-Smith and McCann, 2019) can also form key drivers of agglomeration economies and the clustering of complex economic activity (Balland et al., 2020) with nuanced implications for regional growth policy in both cities and the periphery (McCann and Ortega-Argilés, 2015; Balland et al., 2018).

The article's conclusions are also broader than geographical economics and economic geography studies of innovation or growth. Overlooking the limitations of any approximations and assumptions in economic and spatial analysis can lead to unintended effects and mistaken conclusions. A best practice approach should require that no variable, model type and model form (linear, exponential or otherwise) be used without justification and/or acknowledging the limitations and potential unintended effects.[8]

Notes

[1]The term aspatial is used in the sense that these theories assume no role for space or location. There is still an implicit geography defined by the scope of the economy and an implicit assumption that collaboration and transactions work perfectly within this scope.

[2]The term *unintended consequences* typically refers to Robert Merton's (1936) paper describing that for any policy intervention it is inevitable that some outcomes are accidental. In this article, the proverb applies twice: modelling assumpitions can lead to (i) accidental conclusions about spatial-economic phenomena that imply incorrect policies which also result in (ii) outcomes that are accidental. Despite this, I do not agree with Merton's conclusion that this implies there should be no intervention. An intervention that is less (or more) successful than expected, but ultimately improves outcomes is still better than no intervention at all. The important factor is the distribution of risk around the estimates of costs and benefits. There may also be equity reasons to justify an intervention.

[3]See Baldwin et al. (2003) for a thorough overview of the relevant geographical economics and trade literature.

[4]In these toy models labor is the only factor of production, so corresponds directly with scale. In models with additional factors of production these factors are also required in the definition of scale. In such models diminishing returns to labor induces investment in other factors such as physical and human capital. A growing population induces growth in all other rival factors, also implying that per capita growth is an increasing function of the scale of inputs.

[5]In this toy model, F is proportional to L. This assumption is not controversial since Laincz and Peretto (2006) show that this is a natural equilibrium outcome. In any case, assuming proportionality is also not a critical requirement. See Peretto (2018).

[6]See also the Appendix to Peretto (2018).

[7]This is not an exhaustive list.

[8]I thank Harald Bathelt for noting that this article is part of a broader effort to improve modelling practices. He may recognise his words in this concluding paragraph.

References

Acs, Z. J. and D. B. Audretsch (1988). Innovation in large and small firms: An empirical analysis. *The American Economic Review 78*(4), 678–690.

Aghion, P., T. Besley, J. Browne, F. Caselli, R. Lambert, R. Lomax, C. Pissarides, N. Stern, and J. V. Reenen (2013, January). Investing for Prosperity: Skills, Infrastructure and Innovation. CEP Special Papers 28, Centre for Economic Performance, LSE.

Aghion, P. and P. Howitt (1992, March). A model of growth through creative destruction. *Econometrica 60*(2), 323–51.

Aloi, M., J. Poyago-Theotoky, and F. Tournemaine (Forthcoming). The geography of knowledge and R&D led growth. *Journal of Economic Geography*.

Andersson, F. and R. Forslid (2003). Tax competition and economic geography. *Journal of Public Economic Theory 5*(2), 279–303.

Ang, J. B. and J. B. Madsen (2011). Can second-generation endogenous growth models explain productivity trends and knowledge production in the asian miracle economies. *Review of Economics and Statistics 93*(4), 1360–1373.

Arkolakis, C., M. Peters, and S. K. Lee (2020). European Immigrants and the United States' Rise to the Technological Frontier.

Audretsch, D. B. (2003). Innovation and spatial externalities. *International Regional Science Review 26*(2), 167–174.

Audretsch, D. B. and M. P. Feldman (1996). R&D Spillovers and the geography of innovation and production. *American Economic Review 86*(3), 630–40.

Baldwin, R. and J. Harrigan (2011). Zeros, Quality, and Space: Trade Theory and Trade Evidence. *American Economic Journal: Microeconomics 3*(2), 60–88.

Baldwin, R. E. and R. Forslid (2000a). The core-periphery model and endogenous growth: stabilizing and destabilizing integration. *Economica 67*(267), 307–24.

Baldwin, R. E. and R. Forslid (2000b). Trade liberalisation and endogenous growth: A q-theory approach. *Journal of International Economics 50*(2), 497–517.

Baldwin, R. E. and R. Forslid (2010). Trade liberalization with heterogeneous firms. *Review of Development Economics 14*(2), 161–176.

Baldwin, R. E., R. Forslid, P. Martin, G. Ottaviano, and F. Robert-Nicoud (2003). *Economic Geography and Public Policy*. Princeton University Press.

Baldwin, R. E. and P. Martin (2004). Agglomeration and regional growth. In J. V. Henderson and J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics*, Chapter 60, pp. 2671–2711. Elsevier.

Baldwin, R. E., P. Martin, and G. I. Ottaviano (2001). Global income divergence, trade and industrialisation: The geography of growth take-offs. *Journal of Economic Growth 6*, 5–37.

Balland, P.-A., R. Boschma, J. Crespo, and D. L. Rigby (2018). Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional Studies 0*(0), 1–17.

Balland, P.-A., C. Jara-Figueroa, S. G. Petralia, M. P. A. Steijn, D. L. Rigby, and C. A. Hidalgo (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*.

Behrens, K., G. Duranton, and F. Robert-Nicoud (2014). Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy 122*(3), 507–553.

Bettencourt, L., J. Lobo, D. Helbing, C. Kühnert, and G. West (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America 104*(17), 7301–7306. cited By 1072.

Bettencourt, L. M., J. Lobo, and D. Strumsky (2007). Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy 36*(1), 107 – 120.

Black, D. and V. Henderson (1999). A theory of urban growth. *Journal of Political Economy 107*(2), 252–284.

Blaug, M. (1980). *The methodology of economics: Or how economists explain*. Cambridge surveys of economic literature. Cambridge University Press.

Blind, K. and H. Grupp (1999). Interdependencies between the science and technology infrastructure and innovation activities in german regions: empirical findings and policy consequences. *Research Policy 28*(5), 451 – 468.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020, April). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–44.

Boland, L. (1992). *The Principles of Economics: Some Lies My Teacher Told Me*. Routledge.

Bond-Smith, S. (2019). The decades-long dispute over scale effects in the theory of economic growth. *Journal of Economic Surveys 33*(5), 1359–1388.

Bond-Smith, S. and P. McCann (2014). Incorporating space in the theory of endogenous growth: Contributions from the new economic geography. In M. M. Fischer and P. Nijkamp (Eds.), *Handbook of Regional Science*, pp. 213–236. Springer Berlin Heidelberg.

Bond-Smith, S., P. McCann, and L. Oxley (2018). A regional model of endogenous growth without scale assumptions. *Spatial Economic Analysis 13*(1), 5–35.

Bond-Smith, S. C. and P. McCann (2019). A multi-sector model of relatedness, growth and industry clustering. *Journal of Economic Geography*. lbz031.

Boschma, R. and K. Frenken (2009). Technological relatedness and regional branching. *Papers in Evolutionary Economic Geography (PEEG)* (0907).

Breinlich, H., G. I. Ottaviano, and J. R. Temple (2014). Regional Growth and Regional Decline. In *Handbook of Economic Growth*, Volume 2 of *Handbook of Economic Growth*, Chapter 4, pp. 683–779. Elsevier.

Ciftci, M. and W. M. Cready (2011). Scale effects of r&d as reflected in earnings and returns. *Journal of Accounting and Economics 52*(1), 62 – 80.

Combes, P.-P. and G. Duranton (2006, January). Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics 36*(1), 1–28.

Curtis, A., S. McVay, and S. Toynbee (2020, June). The changing implications of research and development expenditures for future profitability. *Review of Accounting Studies 25*(2), 405–437.

Daboín, C., M. Escobari, G. Hernández, and J. Morales-Arilla (2019). Economic complexity and technological relatedness: Findings for american cities. *Brookings Institution Technical Paper*.

Davis, C. and K.-i. Hashimoto (2014). Patterns of technology, industry concentration, and productivity growth without scale effects. *Journal of Economic Dynamics and Control 40*(C), 266–278.

Davis, C. and K.-I. Hashimoto (2015). Industry concentration, knowledge diffusion and economic growth without scale effects. *Economica 82*(328), 769–789.

Davis, C. R. (2009). Interregional knowledge spillovers and occupational choice in a model of free trade and endogenous growth. *Journal of Regional Science 49*, 855–876.

Davis, D. R. (1998). The Home Market, Trade, and Industrial Structure. *American Economic Review 88*(5), 1264–76.

Davis, D. R. and J. I. Dingel (2019, January). A spatial knowledge economy. *American Economic Review 109*(1), 153–70.

de Groot, H., J. Poot, and M. Smit (2016). Which agglomeration externalities matter most and why? *Journal of Economic Surveys 30*(4), 756–782.

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

Desmet, K. and E. Rossi-Hansberg (2009). Spatial growth and industry age. *Journal of Economic Theory 144*(6), 2477–2502.

Desmet, K. and E. Rossi-Hansberg (2010). On Spatial Dynamics. *Journal of Regional Science 50*(1), 43–63.

Desmet, K. and E. Rossi-Hansberg (2012). Innovation in space. *American Economic Review 102*(3), 447–52.

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

Dinopoulos, E. and P. Thompson (1998). Schumpeterian Growth without Scale Effects. *Journal of Economic Growth 3*(4), 313–35.

Diodato, D., F. Neffke, and N. O'Clery (2018). Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics 106*, 1 – 26.

Duranton, G. (2006). Some foundations for Zipf's law: Product proliferation and local spillovers. *Regional Science and Urban Economics 36*(4), 542–563.

Duranton, G. (2007). Urban evolutions: The fast, the slow, and the still. *American Economic Review 97*(1), 197–221.

Duranton, G. and D. Puga (2004). Micro-foundations of urban agglomeration economies. In J. V. Henderson and J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics*, Volume 4 of *Handbook of Regional and Urban Economics*, Chapter 48, pp. 2063–2117. Elsevier.

Duranton, G. and D. Puga (2005). From sectoral to functional urban specialisation. *Journal of Urban Economics 57*(2), 343 – 370.

Duranton, G. and D. Puga (2019, December). Urban Growth and its Aggregate Implications. NBER Working Papers 26591, National Bureau of Economic Research, Inc.

Eeckhout, J., R. Pinheiro, and K. Schmidheiny (2014). Spatial Sorting. *Journal of Political Economy 122*(3), 554–620.

Escobari, M., I. Seyal, J. Morales-Arilla, and C. Shearer (2019). Growing cities that work for all: A capability-based approach to regional economic competitiveness. Technical report, Brookings Institution.

Frenken, K., F. V. Oort, and T. Verburg (2007). Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies 41*(5), 685–697.

Friedman, M. (1953). The methodology of positive economics. In M. Friedman (Ed.), *Essays in Positive Economics*, pp. 3–43. University of Chicago Press.

Fujishima, S. (2013). Growth, agglomeration, and urban congestion. *Journal of Economic Dynamics and Control 37*(6), 1168 – 1181.

Fukuda, K. (2017). The effects of globalization on regional inequality in a model of semi-endogenous growth and footloose capital. *Asia-Pacific Journal of Accounting & Economics 24*(1-2), 95–105.

Glaeser, E. L. (2003). The new economics of urban and regional growth. In G. L. Clark, M. S. Gertler, and M. P. Feldman (Eds.), *The Oxford Handbook of Economic Geography*. Oxford University Press.

Glaeser, E. L. and J. D. Gottlieb (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., H. D. Kallal, J. A. Scheinkman, and A. Shleifer (1992). Growth in Cities. *Journal of Political Economy 100*(6), 1126–52.

Greasley, D., J. B. Madsen, and M. E. Wohar (2013). Long-run growth empirics and new challenges for unified theory. *Applied Economics 45*(28), 3973–3987.

Grossman, G. M. and E. Helpman (1991). Quality ladders in the theory of growth. *Review of Economic Studies 58*(1), 43–61.

Ha, J. and P. Howitt (2007). Accounting for trends in productivity and R&D: A Schumpeterian critique of cemi-endogenous growth theory. *Journal of Money, Credit and Banking 39*(4), 733–774.

Helpman, E. and P. Krugman (1985). *Market structure and foreign trade: Increasing returns, imperfect competition, and the international economy*. MIT press.

Henderson, V. (1997). Externalities and industrial development. *Journal of Urban Economics 42*(3), 449 – 470.

Henderson, V., A. Kuncoro, and M. Turner (1995, October). Industrial Development in Cities. *Journal of Political Economy 103*(5), 1067–1090.

Hidalgo, C., P.-A. Balland, R. Boschma, M. Delgado, M. Feldman, K. Frenken, E. Glaeser, C. He, D. Kogler, A. Morrison, F. Neffke, D. Rigby, S. Stern, S. Zheng, and S. Zhu (2018). *Unifying Themes in Complex Systems (IX)*, Chapter The Principle of Relatedness, pp. 451–457. Springer.

Howitt, P. (1999). Steady endogenous growth with population and r & d inputs growing. *Journal of Political Economy 107*(4), 715–730.

Isaksson, O., M. Simeth, and R. Seifert (2016, 01). Knowledge spillovers in the supply chain: Evidence from the high tech sectors. *Research Policy 45*, 699–706.

Izushi, H. (2008). What does endogenous growth theory tell about regional economies? empirics of r&d worker-based productivity growth. *Regional Studies 42*(7), 947–960.

Jacobs, J. (1969). *The economy of cities*. New York: Random House.

Jones, C. I. (1995a). R&d-based models of economic growth. *Journal of Political Economy 103*(4), 759–84.

Jones, C. I. (1995b). Time Series Tests of Endogenous Growth Models. *The Quarterly Journal of Economics 110*(2), 495–525.

Jones, C. I. (1999). Growth: With or Without Scale Effects? *American Economic Review 89*(2), 139–144.

Kortum, S. S. (1997). Research, Patenting, and Technological Change. *Econometrica 65*(6), 1389–1420.

Krugman, P. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics 9*(4), 469–479.

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

Laincz, C. and P. F. Peretto (2006). Scale effects in endogenous growth theory: an error of aggregation not specification. *Journal of Economic Growth 11*(3), 263–288.

Lesage, J. P. and M. M. Fischer (2008). Spatial Growth Regressions: Model Specification, Estimation and Interpretation. *Spatial Economic Analysis 3*(3), 275–304.

Lucas, R. J. (1988). On the mechanics of economic development. *Journal of Monetary Economics 22*(1), 3–42.

Luintel, K. B. and M. Khan (2017). Ideas production and international knowledge spillovers: Digging deeper into emerging countries. *Research Policy 46*(10), 1738 – 1754.

Madsen, J. B. (2008). Semi-endogenous versus Schumpeterian growth models: testing the knowledge production function using international data. *Journal of Economic Growth 13*(1), 1–26.

Madsen, J. B., J. B. Ang, and R. Banerjee (2010). Four centuries of British economic growth: the roles of technology and population. *Journal of Economic Growth 15*(4), 263–290.

Martin, P. and G. I. Ottaviano (1999). Growing locations: Industry location in a model of endogenous growth. *European Economic Review 43*(2), 281 – 302.

Martin, P. and G. I. P. Ottaviano (2001). Growth and agglomeration. *International Economic Review 42*(4), 947–68.

Martin, R. and P. Sunley (1998). Slow convergence? the new endogenous growth theory and regional development. *Economic Geography 74*(3), 201–227.

McCann, P. and Z. Acs (2011). Globalization: Countries, Cities and Multinationals. *Regional Studies 45*(1), 17–32.

McCann, P. and R. Ortega-Argilés (2015). Smart Specialization, Regional Growth and Applications to European Union Cohesion Policy. *Regional Studies 49*(8), 1291–1302.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review 1*(6), 894–904.

Minerva, G. A. and G. I. Ottaviano (2010). *Handbook of Regional Growth and Development Theories*, Chapter Endogenous growth theories: Agglomeration benefits and transportation costs. Cheltenham, England: Edward Elgar.

Minniti, A. and C. P. Parello (2011). Trade integration and regional disparity in a model of scale-invariant growth. *Regional Science and Urban Economics 41*(1), 20–31.

Nagy, D. K. (2020). Hinterlands, city formation and growth: Evidence fromthe U.S. westward expansion. Unpublished manuscript.

Neffke, F., M. Henning, R. Boschma, K.-J. Lundquist, and L.-O. Olander (2011). The dynamics of agglomeration externalities along the life cycle of industries. *Regional Studies 45*(1), 49–65.

Ottaviano, G., T. Tabuchi, and J.-F. Thisse (2002). Agglomeration and trade revisited. *International Economic Review 43*(2), 409–436.

Palivos, T. and G. Karagiannis (2010, November). The Elasticity Of Substitution As An Engine Of Growth. *Macroeconomic Dynamics 14*(5), 617–628.

Peretto, P. F. (1998). Technological change and population growth. *Journal of Economic Growth 3*(4), 283–311.

Peretto, P. F. (2018). Robust endogenous growth. *European Economic Review 108*, 49 – 77.

Proost, S. and J.-F. Thisse (2019). What can be learned from spatial economics? *Journal of Economic Literature 57*(3), 575–643.

Redding, S. J. (2020, September). Trade and Geography. Working Paper 27821, National Bureau of Economic Research.

Redding, S. J. and E. Rossi-Hansberg (2017). Quantitative spatial economics. *Annual Review of Economics 9*(1), 21–58.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy 98*(5), S71–102.

Romer, P. M. (2015). Mathiness in the theory of economic growth. *American Economic Review 105*(5), 89–93.

Segerstrom, P. S. (1998). Endogenous Growth without Scale Effects. *American Economic Review 88*(5), 1290–1310.

Simmie, J. (2003). Innovation and urban regions as national and international nodes for the transfer and sharing of knowledge. *Regional Studies 37*(6-7), 607–620.

Storper, M. (2010, 12). Why do regions develop and change? The challenge for geography and economics. *Journal of Economic Geography 11*(2), 333–346.

Tether, B. S., I. J. Smith, and A. T. Thwaites (1997, March). Smaller enterprises and innovation in the UK: the SPRU innovations database revisited. *Research Policy 26*(1), 19–32.

Ulku, H. (2007). R&D, innovation, and growth: evidence from four manufacturing sectors in OECD countries. *Oxford Economic Papers 59*(3), 513–535.

Venables, A. J. (1996). Equilibrium locations of vertically linked industries. *International Economic Review 37*(2), 341–59.

Venables, A. J. (1999). The international division of industries: Clustering and comparative advantage in a multi-industry model. *Scandinavian Journal of Economics 101*(4), 495–513.

Venturini, F. (2012). Product variety, product quality, and evidence of endogenous growth. *Economics Letters 117*(1), 74–77.

Verginer, L. and M. Riccaboni (2021). Talent goes to global cities: The world network of scientists' mobility. *Research Policy 50*(1), 104127.

Walz, U. (1997). Growth and deeper regional integration. *Review of International Economics 5*(4), 492–507.

WIPO (2019). World Intellectual Property Report 2019: the geography of innovation: Local hotspots, global networks. Technical report, World Intellectual Property Organization.

Yamamoto, K. (2003). Agglomeration and growth with innovation in the intermediate goods sector. *Regional Science and Urban Economics 33*(3), 335–360.

Young, A. (1998). Growth without scale effects. *Journal of Political Economy 106*(1), 40–63.

Zachariadis, M. (2003). R&D, innovation, and technological progress: a test of the Schumpeterian framework without scale effects. *Canadian Journal of Economics 36*(3), 566–586.

Zhang, W.-B. (2007). A multiregion model with capital accumulation and endogenous amenities. *Environment and Planning A: Economy and Space 39*(9), 2248–2270.