Why You'd Want to Live Here

Spatial Equilibria from the Industrial Revolution to the Information Age

Castleforge Partners
The Heal's Building, 22-24 Torrington Place, London WC1E 7HJ | +44 (0)207 637 4335 | info@castleforgepartners.com

PREFACE

When the facts change, I change my mind. What do you do? – JM Keynes

Real estate is an industry of heuristics. Practitioners often speak about the dynamics behind asset classes or the supply and demand drivers of valuation in such broad terms that the layperson could confuse these statements as fact, not of circumstance (indeed, so often, too, does the non-layperson). "Multifamily cap rates are lower than office ones." "Location, location and location are the drivers behind value and returns." "Hotels are relatively risky because of the operational component." "Longer leases are more valuable than shorter leases." "Shopping malls are dead." We industry folk are familiar with many of them.

It would appear that first-principles based thinking in real estate is simply not done. And why should it be? Thinking from first principles is difficult, and it is time-consuming. In any industry where change occurs infrequently, like it often does in real estate, yesterday's conclusion is about as good as today's.

And so, thinking by heuristics has been an efficient and successful route for many investors. Indeed, entire investing careers have been built in the horizon of a single real estate paradigm. A GI returning home after the war having purchased a home in Levittown could have become very wealthy by becoming a real estate professional tied to the growth in the American suburban lifestyle, from homes to shopping malls and big box retail to lifestyle centres. My own father moved from Hungary to New York in 1974. Had he hung up the stethoscope to start a career investing in US "gateway city" real estate, I might now be living a life of leisure. And at the end of those two hypothetical careers, almost nothing would have suggested that the dynamics that drove the success all along were waning. Such is the pace at which real estate fundamentals typically change.

For Castleforge's research department, on the other hand, that's not good enough. The real estate practitioner who uses only heuristics often does have a remarkable way of pulling rabbits out of a hat. But for those of us who don't believe in magic, it's more fascinating (and necessary) to discover how the rabbit got there in the first place.

We start with the supposition that the most important driver of what creates value in the built environment, fundamentally, has to be population. If there were no people to use the real estate that firms create and own, then there would be no cash flow stream from which to derive any value. Resultingly, how people choose to orientate themselves and where they choose to live and work and play—the primary purpose of this paper—are the first set of questions that any real estate professional needs to answer in order to understand what drives value in her industry.

Questions as varied as "how many warehouses will we need in Germany and where should they be built?", "what will be the demand for middle-range apartment rentals in Portland?", "how many square feet and on what sort of lease will businesses in Osaka demand?" can only begin to be answered after one has a vague idea about the population dynamics in these cities and countries. Yet change is occurring at a rapid pace: a combination of technological advances, climate change, Covid-19, and other factors are making it difficult to know what the next five years looks like, let alone the 50-year horizon of a real estate investment career for today's graduates.

What is needed is a general framework to understand how people have arranged themselves onto this earth at the present. That framework must be able to explain why the past looked like it did and how we got to where we are today. This framework is key, because it also helps the investor to understand how the future might look as various changes taking place impact how populations will re-orientate themselves over time.

Fortunately, others have led the way. Over the past few decades, economists, historians, and other scholars have worked to create this framework. Thanks to the likes of Paul Krugman and those since,

the field of New Economic Geography ("NEG") has literally transformed the way that economists think about how populations orientate themselves. Still, few in our industry are familiar with the field of NEG, much less how it would apply to investing in real estate. What we have attempted in the following report is a summary of the literature and an explanation of the NEG framework in an effort to help to make sense of how we see the world changing in the medium to long term.

Our own view at Castleforge is that the facts have changed. After a long period of relative stability in multiple asset classes and geographies, the old paradigms are now clearly in a state of flux. These changes will create winners and losers across the real estate industry. Those able to spot the winners of the future will be in an excellent position to benefit from structural tailwinds in the decades to come. In an uncertain world so susceptible to change and disruption, practitioners may now want to familiarise themselves with this framework in an effort to re-think from first principles all the "truisms" that our industry thought we knew as fact.

Michael Kovacs

December 2021

INTRODUCTION

We take up a very basic question in this paper: *why do people live where they do?* In the pages below, we draw on insights from economists, historians, and urbanists to try to understand why some places thrive, why other places flounder, and why the fates of particular places evolve over time.

These questions are crucial ones for real estate investors to reckon with. After all, location famously drives value in real estate. When a place succeeds, property in that place tends to outperform. Alternatively, when a place fails, property there tends to struggle. This implies that, if investors want to predict how real estate markets in particular locations will change, they must form broader views about how those places themselves will change. This paper aims to provide a foundation for developing such views.

Our discussion contains three parts.

In *Section 1*, we sketch a theoretical framework, relying on ideas formulated by spatial and urban economists. Our framework involves three core claims. First, individuals decide to live in particular locations to the extent that those places can deliver the right mix of benefits, including good *wages*, affordable *rents*, and high-quality *amenities*. Second, wages, rents, and amenities are determined by what we call *local advantages in production and consumption*, and advantages in production and consumption stem from *natural endowments*, *human interventions*, or *economies of agglomeration*. Third, local advantages evolve when new technologies emerge, government policies change, or cultural preferences adjust. As we try to show, these are the basic forces that propel certain places, like twenty-first century London, to success—and doom other places, like late-twentieth century Detroit, to failure.

Next, in *Section 2*, we turn to the past, applying our theoretical framework to two historical episodes. First, we consider the original wave of urbanisation that swept Britain in the nineteenth century. We show how two sets of technological developments—improvements in agricultural practices and the transition to a coal-dominated energy mix—paved the way for rapid growth in cities like those in Britain, including London. Second, we explore urban decentralisation in twentieth-century America. We explain how the introduction of the automobile, combined with newly evolved cultural preferences, nudged Americans out of the old, cold, and dense cities in the North and pulled them into newer, sunnier, and lower-density settlements, particularly in the Sun Belt.

Finally, in *Section 3*, we consider the future, reflecting on the ways that new communications technologies may change where people decide to live and work in the coming decades. We argue that the advantages in production that superstar cities have monopolised over the last quarter-century have started to disperse—and that a more egalitarian urban landscape, where secondary cities flourish, is now emerging as a result.

SECTION 1 | SETTING UP A FRAMEWORK

Economic theories of urban growth start with a basic insight: in order to explain why human geography evolves over time, we must understand what motivates people and firms to locate where they do. After all, in places where people enjoy freedom of movement, settlement patterns should reflect utility-maximising decisions made by individuals. This idea is intuitive: desirable locations attract residents, undesirable locations repel them, and just like in ordinary markets, prices—here, in the form of rents and wages—adjust to ensure equilibrium. There are caveats and exceptions, which we will eventually consider, but let's begin by developing that basic idea in more depth.

SECTION 1.1 | THE CONCEPT OF SPATIAL EQUILIBRIUM

As we have just suggested, it's useful to conceive human settlement patterns as market outcomes, which are driven by individual decision-making aimed at maximising utility. This suggestion raises

obvious questions. When considering where to locate, how exactly do individuals maximise utility? What kinds of costs and benefits enter into their consideration?

With those questions in mind, we start with a simple model, drawn from the spatial economics literature.[1] The main pecuniary cost to an individual of living in a particular place is the price of local housing, and the main pecuniary benefit to an individual of living in particular place is the local wage. Still, while money surely matters, so too do various non-pecuniary considerations. These include: local climate, natural beauty, access to entertainment, access to food produce, commute times, school quality, public safety, and transportation coverage. We call these *amenities*. Finally, people may also weigh their idiosyncratic *preferences*. (For example, somebody may feel emotionally attached to the place where he was raised.) We will therefore say that the total utility a person obtains by living somewhere depends on the level of *rents*, *wages*, *amenities,* and *preference-satisfaction*, which he receives in that place. A simple utility function, for individual *i* in location *A*, follows:

$$U_A^i = wage_A - rent_A + amenities_A + preference_A^i$$

This function implies that, all else equal, lower rents, higher wages, or better amenities will improve *i*'s utility in *A*; and that higher rents, lower wages, or worse amenities will degrade *i*'s utility in *A*. Leaving aside idiosyncratic preferences (which anyway do little to illuminate population-level trends), we can assume that comparatively lower rents, higher wages, or better amenities make a location more advantageous, which encourages in-migration, while comparatively higher rents, lower wages, or worse amenities make a place less advantageous, which encourages out-migration.

How do rents, wages, and amenities interact? The *spatial equilibrium assumption* enters at this point. According to the spatial equilibrium assumption, "there are not [economic] rents to be gained by changing locations" (Glaeser 2008, p. 47). While different places provide different benefits and impose different costs, these both adjust so that, in equilibrium, the average individual cannot gain by relocating. As Edward Glaeser puts it, "if identical people are choosing to live in two different places then those two different places must be offering an equivalent bundle of advantages, like wages, prices, and amenities. Essentially, there must be no potential for arbitrage across space" (Glaeser 2008, p. 4). This just means that, for two locations, *A* and *B*:

$$U_A^i = U_B^i$$

Spatial equilibrium is reached via the usual mechanisms of supply and demand. If residents of Tokyo could increase their total utility by moving to Osaka—for example, because Osaka provided relative benefits, like higher-paying jobs, cheaper housing, or lower crime, and no commensurate costs—then that's exactly what they would do. As people moved, the demand for housing and the supply of labour in Osaka would both expand, raising rents and reducing wages. Residents of Tokyo would continue to move until the point at which the net benefits of relocating to Osaka fell to zero, when equilibrium was restored, and spatial arbitrage was no longer possible.

The crucial point here is that the location-specific costs and benefits of different places must tend to offset each other. Or, put differently: so long as $U_A^i = U_B^i$, if wages in *A* exceed wages in *B*, then rents in *A* must exceed housing costs in *B*, or amenities in *B* must be superior to amenities in *A*, or both.

For example, in the early twentieth century, when government investment in public health brought infectious disease in American cities under control for the first time, the urban mortality penalty

[1] This model was originally developed by Rosen (1979) Roback (1972). See Glaeser (1998) for a summary.

shrank. In tandem, real wages in American cities fell (relative to wages elsewhere in the United States), and the country's urban population grew.[2] Why did that happen?

Spatial equilibrium tells us that utility must equalise across space. When public health (a kind of amenity) in cities improved relative to rural or suburban areas, cities became more attractive places to live on a relative basis. As people relocated to cities to capture that improvement, the urban labour supply expanded, pushing down wages. More intuitively, the spatial equilibrium idea explains how costly locations, like London and New York, nonetheless retain millions of residents: these places tend to provide commensurately high wages or appealing amenities.

Spatial equilibrium is a powerful idea, but it only gets us so far. Although it's correct to point out that high wages help offset London's high housing costs (and therefore help keep people there), a full explanation of urban development in London should tell us something about what causes London's wages to be so high at all. At a more fundamental level, we want to explain why prices, like wages and rents, vary by location and time in the first place, leading to changes in migration flows. We will turn to that topic next.

SECTION 1.2 | ADVANTAGES IN PRODUCTION AND CONSUMPTION

Fundamentally, wages and rents vary by location because locations vary by their capacity to sustain (1) productive labour and (2) high-quality life. In other words, different locations have different *local advantages*, which benefit individuals in their dual roles as producers (i.e., workers) and consumers (i.e., residents). We will call these *advantages in production* and *advantages in consumption*, respectively. For now, brief definitions will suffice. Advantages in production are location-specific features that contribute to the productivity of local labour, while advantages in consumption are location-specific features that enhance the residential quality of a place. Flat topography makes northern Iowa a productive place to cultivate corn, and sunny weather makes Sydney a high-quality place to live.

Importantly, these local advantages change over time. As we explore throughout this paper, local advantages evolve when (1) new technologies emerge, (2) government policies change, or (3) shared cultural preferences shift. We call these events *shocks*. Shocks to local advantage are transmitted through local markets, leading to adjustments in local wages and rents. In particular, labour markets (and therefore wages) respond most directly to new advantages in production, just as housing markets (and therefore rents) respond most directly to new advantages in consumption.

Let's start by cataloguing the first-order effects of these shocks. Consider four possibilities. First, when the production advantages of a place improve, local wages tend to rise, as the productivity of local labour increases, and demand for local labour expands. Second, when production advantages deteriorate, local wages tend to decline, as the productivity of local labour decreases, and demand for local labour contracts. Third, when the consumption advantages of a place improve, local rents tend to increase, as demand for local housing surges. Fourth, when consumption advantages deteriorate, local rents tend to fall, as demand for local housing declines.

[2] Boustan (2019).

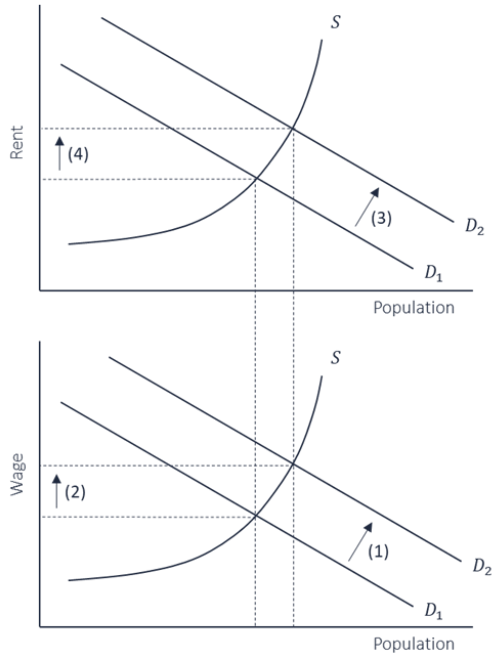
FIGURE 1: FOUR TYPES OF SHOCK AND THEIR FIRST-ORDER EFFECTS

Shock type	Housing demand	Labour demand	Rents	Wages
Production advantages improve	Held constant	Increases	Held constant	Increase
Production advantages deteriorate	Held constant	Decreases	Held constant	Decrease
Consumption advantages improve	Increases	Held constant	Increase	Held constant
Consumption advantages deteriorate	Decreases	Held constant	Decrease	Held constant

But the story doesn't end there. As we just saw, when the production advantages of a place improve, labour there enjoys productivity gains, which push up local wages. However, the spatial equilibrium assumption predicts that that arbitrage opportunity will be exploited until it disappears. Here, spatial equilibrium is realised as the higher wage level entices newcomers, who bid up housing costs. The result is higher rents and a larger equilibrium population. Similar progressions follow the three other types of shock, and in this way, population growth or decline is driven, proximately, by changes in wages, rents, and amenities. We illustrate these progressions in Figure 2.

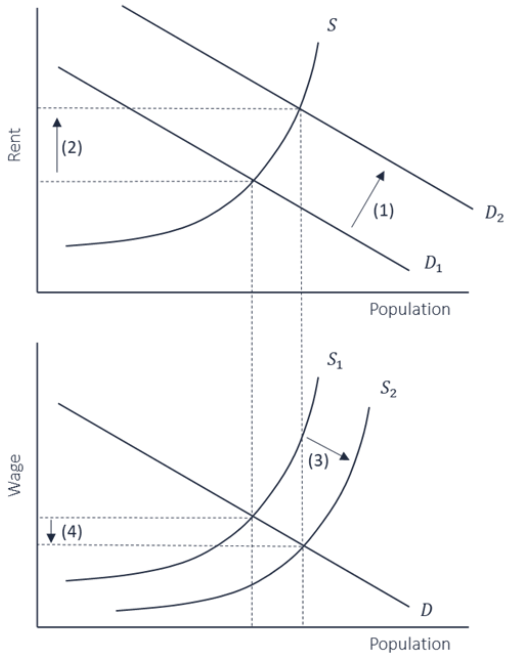
FIGURE 2: SPATIAL EQUILIBRIUM WITH ADVANTAGES IN PRODUCTION AND CONSUMPTION

2(a): Production advantages improve



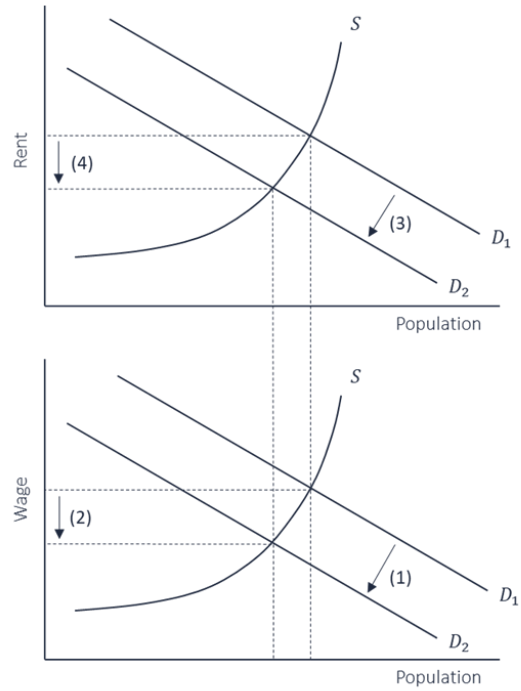
(1) Demand for labour increases. (2) Wages increase. (3) Demand for housing increases. (4) Rents increase. Population increases.

2(c): Consumption advantages improve



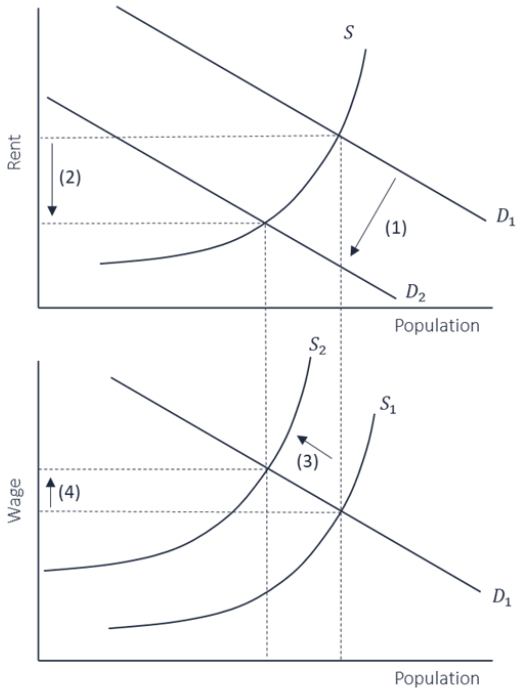
(1) Demand for housing increases. (2) Rents increase. (3) Supply of labour increases. (4) Wages decrease. Population increases.

2(b): Production advantages deteriorate



(1) Demand for labour decreases. (2) Wages decrease. (3) Demand for housing decreases. (4) Rents decrease. Population decreases.

2(d): Consumption advantages deteriorate



(1) Demand for housing decreases. (2) Rents decrease. (3) Supply of labour decreases. (4) Wages increase. Population decreases.

SECTION 1.3 | THE SOURCES OF LOCAL ADVANTAGE

Advantages in production and consumption can be put into three categories: natural advantages, artificial advantages, and agglomerative advantages. In this section, we consider each in turn.

Natural advantages

First, a location can be advantaged by virtue of its natural endowments. Call these *natural advantages*.

A *natural advantage in production* is any local natural endowment that contributes to a location's productivity—for example: proximity to major waterways; possession of natural resources, like energy, stones, or precious metals; physical elevation (which enhances security and protects capital goods); and the presence of soil well-suited to cultivating crops. Looking at the long arc of human history, we see that settlements often arose in specific places for reasons associated with natural advantages in production. Mesopotamia was famously fertile; Luxor, or ancient Thebes, sits on the shores of the Nile; and Jerusalem is, literally, a city on a hill.[3]

A *natural advantage in consumption* is any local natural endowment that improves a location's residential quality. This includes the amenities we noted above, like pleasant weather, proximity to mountains or beaches, and natural beauty. In a world of increasingly abundant wealth, decentralised production, and digital entertainment, natural advantages in consumption might be more important than ever. Consider one suggestive illustration: while the population of American counties with mean January temperatures above 45 degrees Fahrenheit (7 degrees Celsius) grew by over 9 percent between 2000 and 2010, the population of counties with mean January temperatures below 22 degrees (minus 6 degrees Celsius) grew by less than 2 percent during that same period.[4] (We will return to this theme in *Section 3*, when we consider the future of cities.)

Natural advantages in production and consumption influence labour and housing markets in precisely the ways we outlined in the graphs when considering different kinds of shocks. When natural consumption advantages improve, the population level rises, in the way that Figure 2(c) describes. For instance, starting in the 1960s, the proliferation of residential air conditioning in the United States improved living conditions in regions with hot climates, helping to fuel subsequent population growth in places like Arizona and Florida.[5]

While natural advantages may fade over time—for example, as new technologies render old ways of transportation obsolete or allow residents to more fully utilise consumption advantages—path dependencies can help cities persist.[6] Manhattan started as a Dutch trading post, linking producers of beaver fur near present-day Albany to the Atlantic Ocean, via the Hudson River. Those days are long gone, but New York obviously survived, partly thanks to local industries that originally developed to support Manhattan's initial role as a hub for international trade.[7] We will explore the main mechanism that underlies urban path dependencies below, when we discuss agglomeration.

---

[3] Note that natural advantages are only natural in a certain sense. While natural endowments are largely pre-determined, whether a location's specific endowments count as advantageous depends on contextual factors. For example, while nature is responsible for the fact that Saudi Arabia sits atop a quarter-trillion barrels of oil, that natural endowment would not matter without the invention of the technologies that made oil useful and extractable. As we have already suggested, when new technologies emerge, natural advantages evolve, too.

[4] See Glaeser (2011). Studying population growth in European cities from 1980 to 2000, Cheshire and Magrini (2005) find that a similar trend holds there. Also see Glaeser, Ponzetto, and Tobio (2011) for a much longer-term study of the correlation between mild winter weather and population growth.

[5] Glaeser (2005).

[6] See Bleakley and Lin (2012) on how path dependencies help cities persist.

[7] Glaeser (2005).

Artificial advantages

Next, a place can be advantaged—once again, in production or consumption—by virtue of man-made contributions to it. Call these *artificial advantages in production* and *artificial advantages in consumption*. Just like their natural analogues, artificial advantages in production improve local productivity, while artificial advantages in consumption enhance the quality of local residential life.

Government often plays a major role in the delivery of artificial advantages. For example, local production is buoyed by well-built physical infrastructure, like electrical grids, roadways, telecommunications networks, ports, oil pipelines, and railroads. The productivity of local labour also depends on intangible public goods, like the rule of law, capable political institutions, and favourable corporate regulation and tax policy. Physical infrastructure and intangible public goods are typically supplied by government. Meanwhile, local residents benefit from the usual non-pecuniary amenities, like high-quality schools, good restaurants and entertainment venues, low crime rates, efficient public transportation, and characterful architecture. Residents can sometimes capture gains from pecuniary benefits, too, like local subsidies (e.g., Alaska's Permanent Fund Dividend) and favourable tax policies for households (e.g., Florida's state-income tax rate of zero). These kinds of benefits effectively increase local wages. With respect to their effects on population growth, shocks to artificial advantage operate just like shocks to natural advantage do.

Agglomerative advantages

Finally, a place can be advantaged simply by virtue of the fact that many people co-locate in it. Call this kind of advantage an *agglomerative advantage* (or, equivalently, an *economy of agglomeration*).

Humans are not distributed evenly throughout space—on the contrary: while a small portion of the planet teems with people, most land sits completely empty. Consider the United States. By 1992, just 1.9 percent of the country's total landmass had been built-up or paved, and yet most cities and towns occupy land that's qualitatively similar to unoccupied land located (relatively) nearby. For example, it no longer really matters that Chicago abuts Lake Michigan, but even back when it did, many sites on the shores of Lake Michigan enjoyed that same advantage.[8] The unevenness of human geography therefore presents a puzzle. Why do people live in such close proximity, bidding up land prices and tolerating nuisances like congestion, when much cheaper plots could be inhabited somewhere else?

As we hinted above, the answer has to do with the fact that agglomeration generates its own special kind of advantage. More precisely, we will say that agglomerative advantages are the *location-specific economies of scale that arise when population density generates increasing returns*. Like natural and artificial advantages, agglomerative advantages come in two varieties, most directly benefitting either producers or consumers. Agglomerative advantages in production arise when population growth improves the marginal productivity of local labour. Agglomerative advantages in consumption arise when population growth improves the marginal quality of local residential life.

It's worth noting that agglomeration has a dark side, too; it usually means higher costs in the form of pricier rents, worse congestion, greater pollution, and higher crime. At certain densities, the marginal costs of agglomeration can exceed its benefits, as we will see in *Section 3*. At this point, we can say that a particular location can no longer be expected to increase in density absent a fundamental change in the costs and benefits of agglomeration, for example in the form of a technological change or a global pandemic.

[8] See Boustan (2019) and Cronon (1991) on the role played by Lake Michigan in Chicago's development.

How agglomerative advantages operate

Economies of agglomeration are visible everywhere. They can help justify the existence of industrial clusters, like the information technology clusters in California's Silicon Valley and on Route 128 outside Boston, or the financial services clusters in the City of London and on Wall Street.[9] Despite steep real estate prices in each of these places, firms judge that the benefits of density outweigh the costs, and therefore opt to co-locate there nonetheless. Most importantly for our purposes, the densest mode of human settlement—the city—also owes its existence to economies of agglomeration. In our modern world, urban productivity markedly exceeds non-urban productivity, and that gap is largely accounted for by the advantages in production that agglomeration confers.[10] People flock to cities so that they can live and work among other people, in spite of the costs that other people's presence usually imposes. As Duranton and Puga put it, "we can […] regard cities as the outcome of a trade-off between agglomeration economies or localised aggregate increasing returns and the costs of urban congestion."[11]

Agglomerative advantages operate through three separate mechanisms: (1) *learning*, (2) *sharing*, and (3) *matching*.[12] Density forces individuals and firms into close proximity with each other, enabling the development of new skills among workers, the rapid spread of technical knowledge, and the diffusion of best practices; it facilitates the sharing of high fixed-cost, indivisible goods, like airports, opera houses, marketplaces, and roadways; and it improves the odds of generating successful matches—between financiers and entrepreneurs, between firms and workers, and between individuals searching for spouses, research collaborators, friends, or business partners. When the population size increases, learning, sharing, and matching all tend to improve, yielding superior agglomerative advantages, which, in turn, improve production and consumption advantages. For example, London's large population allows it to host a large collection of museums, a kind of artificial advantage that helps attract further residents to the city.

Agglomerative advantages help account for the feedback loops that characteristically drive urban growth and decay beyond the first order effects of a shock. When a place experiences population growth, agglomerative advantages can intensify, improving local productivity or the quality of local amenities. If higher density improves production advantages somewhere, local wages rise, and more people opt to move to that place. On the other hand, when residents flee a location and the population size shrinks, agglomerative advantages can erode, which reduces local productivity or the quality of local amenities. If lower density weakens production advantages, local wages fall, which prompts out-migration and results in a smaller equilibrium population.

SECTION 1.4 | A SUMMARY OF THE FRAMEWORK

Above, we laid out a framework designed to capture the basic forces that drive changes in human geography. Let's briefly review. As we have outlined, our framework has three key principles:

1. Individuals choose to locate in particular places to the extent that those places can provide the right mix of benefits, including good wages, affordable rents, and high-quality amenities.
2. Wages, rents, and amenities are determined by local advantages in production and consumption, and advantages in production and consumption depend on natural endowments,

[9] See Saxenian (2020) for a more detailed examination of the Silicon Valley and Route 128 innovation clusters.
[10] Glaeser and Resseger (2009).
[11] Duranton and Puga (2004).
[12] Duranton and Puga (2004) propose this taxonomy of the mechanisms through which agglomeration economies operate. Krugman (1991) provides a different one. His includes knowledge spill-overs, specialisation in non-traded local inputs, and deeper and higher-skilled local labour markets.

human interventions, or economies of agglomeration. These correspond to natural, artificial, and agglomerative advantages, respectively.

3. Local advantages evolve when new technologies emerge, government policies change, or cultural preferences adjust. These shocks are the ultimate sources of urban growth and decline.

In the rest of the paper, we apply the framework developed above to explain and predict how specific shocks—in both the past and the future—have shaped, and will shape, where people choose to live.

SECTION 2 | URBAN GROWTH AND DECLINE IN MODERN HISTORY

In this section, we turn to the past, applying the theoretical framework we developed in *Section 1* to particular historical episodes. In *Section 2.1*, we consider the wave of urbanisation that swept Britain in the wake of the Industrial Revolution, showing how enhanced urban production advantages prompted people to move *en masse* to cities in the nineteenth century. As we describe below, London was a major beneficiary of that process. In *Section 2.2*, we explore twentieth-century America, where the introduction of the automobile led to a more decentralised population. As consumption advantages improved in lower-density locations as a result of the lower transportation costs that automobiles afforded, Americans relocated to suburban areas and to younger, polycentric cities in the country's Sun Belt.

SECTION 2.1 | URBANISATION IN INDUSTRIALISING BRITAIN: 1700-1900

Our study of urban development starts in nineteenth-century England, where London was emerging as the headquarters of the powerful British Empire, a major centre of international trade, and a flagship of Britain's nascent industrial sector. Throughout this period, London enjoyed sustained and rapid growth. At the turn of the nineteenth century, London contained one fifth of Britain's total population, or roughly 1 million people. Only half a century later, half of Britain's population lived in the capital. By 1900, London carried a population of 6.5 million, easily outsizing the second and third most populous cities in the world, Paris (with 4 million people) and New York City (with 3.4 million people). Four centuries earlier, it had not even ranked among the top 10 largest cities in Europe.[13]

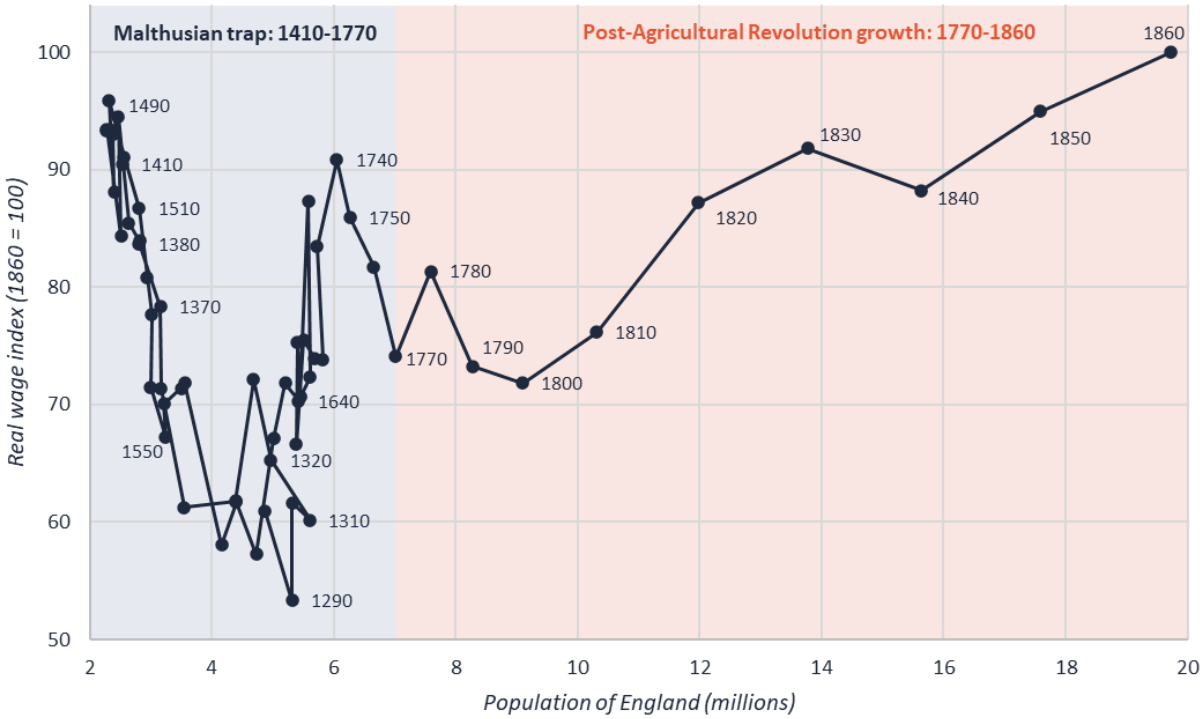
As we show in this section, London's nineteenth-century growth spurt is best understood as a consequence of the intersection of two sets of technology-led changes: (1) extraordinary improvements in British agricultural productivity and (2) Britain's transition to a coal-dominated energy mix during the Industrial Revolution. These twin technological shocks led to improvements in the natural, artificial, and agglomerative advantages in production held by British cities and, in turn, fuelled a massive wave of urbanisation.

The Malthusian trap and rural dominance

Human societies spent most of history narrowly escaping starvation, stuck in so-called *Malthusian traps*. The trap is named for the eighteenth-century economist Thomas Malthus, who argued that it followed mathematically from two observations: while (1) population size has the tendency to grow geometrically, (2) food supply grows arithmetically at best. So long as population growth outpaced the growth of food production, periodic corrections—or, in Malthus's words, "gigantic inevitable famines"—would reduce the population size to serviceable levels from time to time.[14] As a result, Malthus figured, civilisation would remain stuck in a constant struggle to achieve above-subsistence levels of food production. Although the population size could grow over time, per capita incomes would revert to subsistence levels in equilibrium.

[13] Wrigley (2013).
[14] Malthus (1798).

FIGURE 3: THE MALTHUSIAN TRAP[15]



As a backwards-looking explanation set in the eighteenth century, Malthus's theory was plausible. Except for short-lived periods of growth or decline, per capita incomes had stayed roughly flat in Britain for half a millennium prior. England's food supply had grown modestly during that period, but even by 1750, when daily supply stood at 2,237 kilocalories per head, it remained low relative to physical needs.[16] Average incomes were low, and growth was largely stagnant. Malthus's diagnosis seemed correct.

Large-scale urbanisation was impossible in the Malthusian trap. Cities had to remain modest in size as long as food production hovered around the minimum level required for subsistence. This was because producing enough food for survival required relegating most people to rural areas, which possessed the right natural and artificial advantages to sustain agricultural work. Under conditions of food scarcity, marginal agricultural output is extremely valuable relative to other types of output. So, food is relatively expensive in Malthusian societies, agricultural regions are relatively productive (if only because their output is highly valued by consumers), and, consequently, farm wages are relatively high (insofar as non-farm work options do not exist for most people). As a result, most people live and work in the countryside.

Pre-industrial Britain was no exception. Food was costly there. A survey of 187 English household budgets from the late-eighteenth century finds that agricultural families spent about 75 percent of their income on food.[17] The population was overwhelmingly rural, too. Until 1800, less than a fifth of the English and Welsh population inhabited urban areas.[18]

[15] Clark (2005).
[16] Broadberry et al. (2015). Figures represent available supply; due to food waste, actual consumption would have been lower.
[17] Griffin (2018).
[18] De Vries (1984).

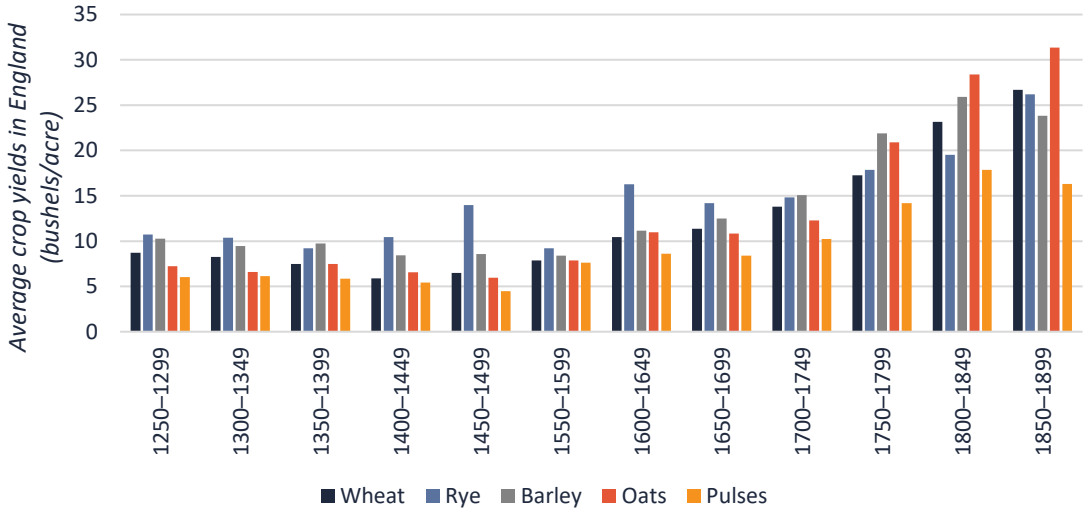
That began to change between the seventeenth and nineteenth centuries, when British agriculture underwent a series of technology-led shifts that significantly upgraded the sector's productivity.

Most importantly, farmers adopted new methods of crop rotation during that period. Farmers had performed crop rotation for millennia, but traditional practices typically involved fallow rotations, which would render up to a third of the arable field area unproductive. With the advent of the Norfolk four-course rotation system in the early eighteenth century, British farmers managed to largely eliminate fallow by introducing turnips and clover into the crop mix.[19] By 1836, clover and turnips occupied nearly half of England's total sown area, up from about a tenth one hundred years prior.[20] All told, a superior crop mix led fallow to decline from about 20 percent of Britain's arable land in 1700 to just four percent by 1871.[21]

Combined with other agricultural innovations, including new fertilisers and farming tools, improvements in Britain's crop mix caused agricultural productivity to climb sharply. In the first half of the eighteenth century, an acre of land in England annually yielded about 14 bushels of wheat, 15 bushels of rye, 15 bushels of barley, 12 bushels of oats, or 10 bushels of pulses. By the mid-nineteenth century, an acre of land yielded 23 bushels of wheat, 20 bushels of rye, 26 bushels of barley, 28 bushels of oats, or 18 bushels of pulses.[22] More generally, real agricultural output per worker increased by about 67 percent in England between 1705 and 1870 and by about 176 percent in Scotland between 1775 and 1860.[23]

[19] Allen (2008). Turnips added humus (decomposed organic matter) to the soil, and clover helped restore the soil's nitrogen content. (The roots of legumes, like clover, host bacteria that convert atmospheric nitrogen into nitrates.) Turnips could also be planted in neat rows, allowing farmers to weed their fields without leaving them fallow. See Brunt (1999).
[20] Mingay (1996).
[21] Overton (1996).
[22] Apostolides et al. (2008).
[23] Brunt and García-Peñalos (2021).

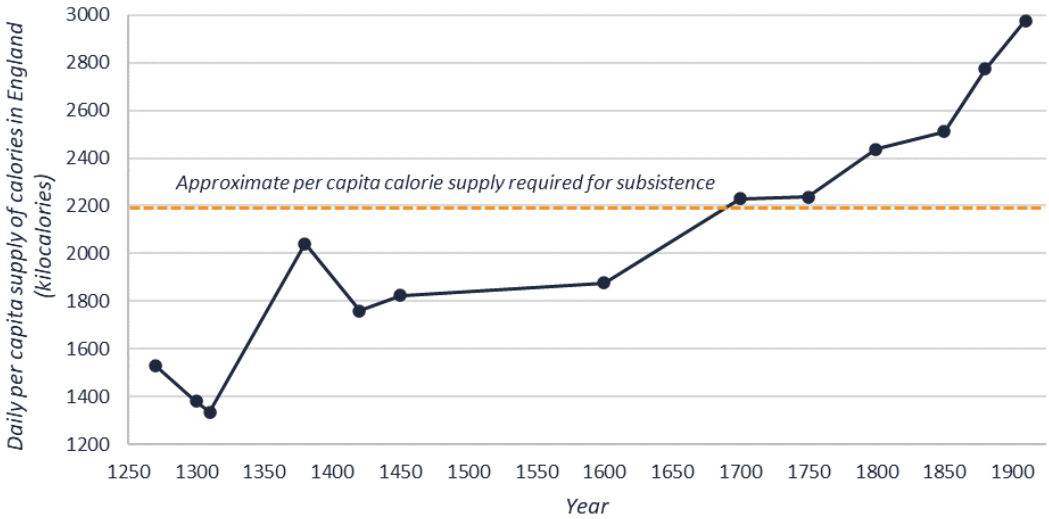
FIGURE 4: CROP YIELDS IN ENGLAND[24]



These improvements pulled Britain out of the Malthusian trap, enabling it to sustain consistent population growth while at the same time achieving per capita income growth.

In 1700, Britain contained about 5 million people, roughly its population at the end of the thirteenth century. By 1851, Britain had 17 million people, and by 1901, it had 30 million people.[25] Despite population growth during this period, England's food supply rose comfortably above subsistence levels for the first time, reaching 2,773 daily kilocalories per head by 1880.[26] Real per capita incomes in England increased by a factor of about 2.5 between 1700 and 1880, and aggregate output rose about tenfold.[27]

FIGURE 5: CALORIE SUPPLY IN ENGLAND[28]



[24] Apostolides et al. (2008).
[25] Broadberry et al. (2015).
[26] Floud et al. (2011).
[27] Broadberry et al. (2015).
[28] Data aggregated by *Our World In Data* (2021). Original sources: Broadberry et al. (2015); Clark, Huberman, and Lindert (1995); Floud et al. (2011).

Transition to a new rural spatial equilibrium

Technological improvements in agriculture transformed the economic geography of Britain. Growth in agricultural productivity translated into absolute improvements in the production advantages of rural Britain, where agriculture was concentrated. But the initial surge in agricultural productivity per hour ultimately reduced the marginal value of agricultural output per worker relative to non-agricultural output, as fewer workers were required to sustain agricultural production once output reached subsistence levels across the population. This, in turn, degraded the production advantages of rural regions relative to urban ones, as workers were no longer needed for agricultural production. In other words, at the margin, once the food supply gap was closed, agricultural work lost value relative to other pursuits. Eventually, by the nineteenth century, Britain was producing enough food that it could afford to release the bulk of its labour force into non-agricultural activities, which were increasingly concentrated in urban areas.

As our framework predicts, the composition of the British labour force adjusted accordingly. In the mid-seventeenth century, 59 percent of the English population worked in agriculture.[29] Two hundred years later, agricultural activities occupied just 21 percent of England's labour force.[30] As we have suggested, this adjustment was triggered by the relative decline in the productivity of agricultural labour. Between 1801 and 1851, labour productivity in agriculture increased by just 5 percent, while labour productivity in industry and services increased by 82 percent and 57 percent, respectively.[31] Wage levels in England and Wales responded in kind. In 1805, farm labour wages were about 110 percent of non-farm common labour wages, 62 percent of collier wages, and 14 percent of engineer and surveyor wages. By 1851, farm labour wages had fallen to about 65 percent of non-farm common labour wages, 52 percent of collier wages, and 6 percent of engineer and surveyor wages.[32] As agricultural production lost value, wages rebalanced, and workers relocated.

While improvements in agricultural productivity pushed Britain's labour force out of the (rural) agrarian sector, a different set of developments pulled workers into the (urban) industrial and services sectors. Shortly after its Agricultural Revolution, Britain started mining coal intensively to support the rising energy needs of its growing population (which we noted above).[33] As coal extraction technologies improved, especially with the advent of the steam engine, and coal was transformed into an increasingly efficient source of energy, Britain eventually transitioned to a coal-dominated energy mix. This generated an unprecedented supply of usable energy. Between the mid-seventeenth and mid-nineteenth centuries, annual per capita energy consumption in England increased by a factor of about five, and coal-generated energy consumption increased by a factor of about 13.[34] Most importantly, Britain's adoption of coal delivered coke-powered blast furnaces, for smelting pig iron, and coal-powered steam engines, used for transportation, manufacturing, and energy production. **Coupled with a labour force that had very recently been given license to pursue non-agricultural activities, these two innovations helped build a new industrial economy.**

[29] Clark et al. (2010).
[30] Broadberry, Campbell, and van Leeuwen (2013).
[31] Broadberry, Campbell, and van Leeuwen (2013).
[32] Own calculations. Nominal wage data from Lindert and Williamson (1983).
[33] Britain's transition to coal was driven by necessity at first. Until the Industrial Revolution, Britain mainly relied on animal muscle for mechanical energy and firewood for heat energy. Around the eighteenth century, that energy mix started to prove inadequate. Animal muscle was incapable of harnessing enough energy to power Britain's budding industrial sector, and, as population growth accelerated in the wake of the Agricultural Revolution, the country's woodlands had started to strain. To overcome that limit, Britain started mining coal intensively.
[34] Wrigley (2013).

FIGURE 6: URBANISATION IN ENGLAND AND WALES

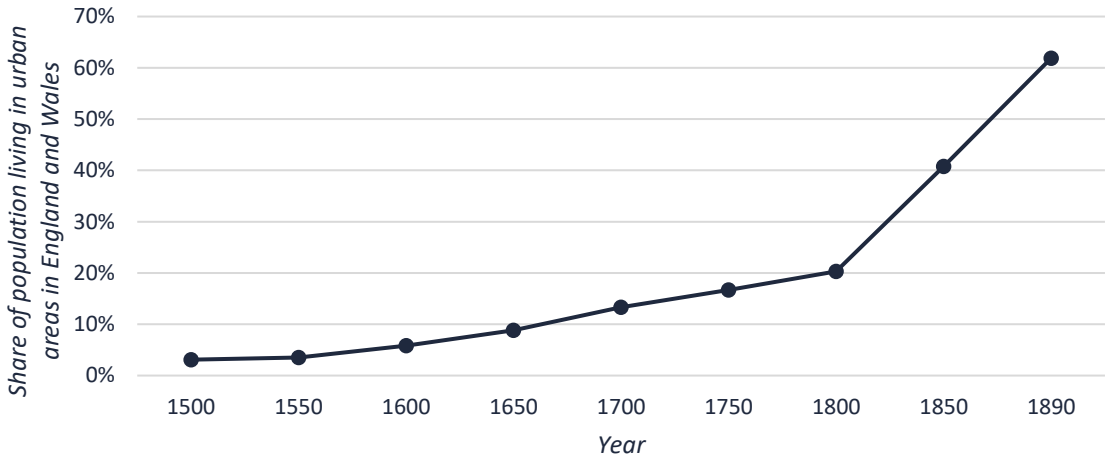
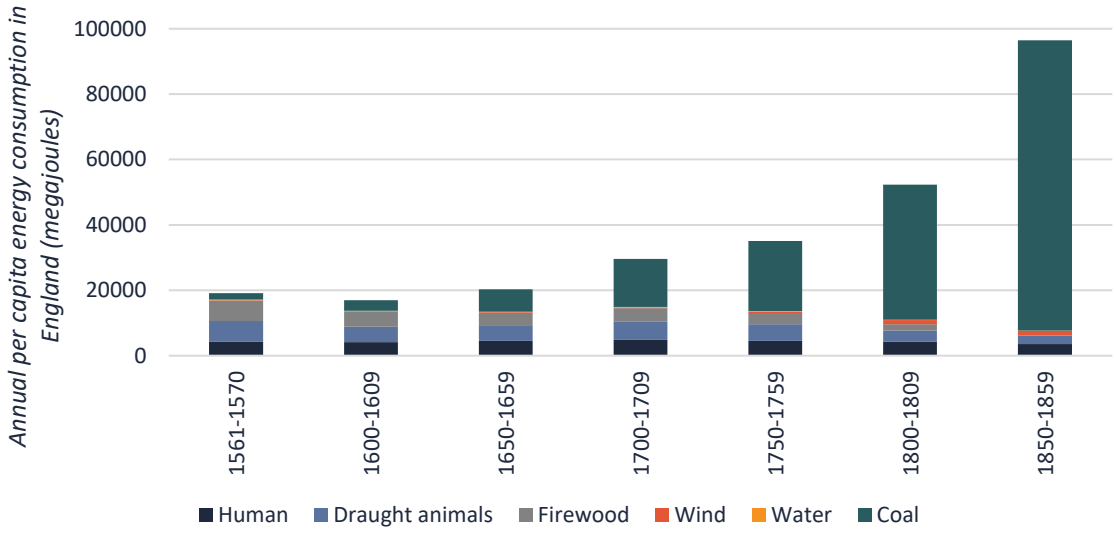
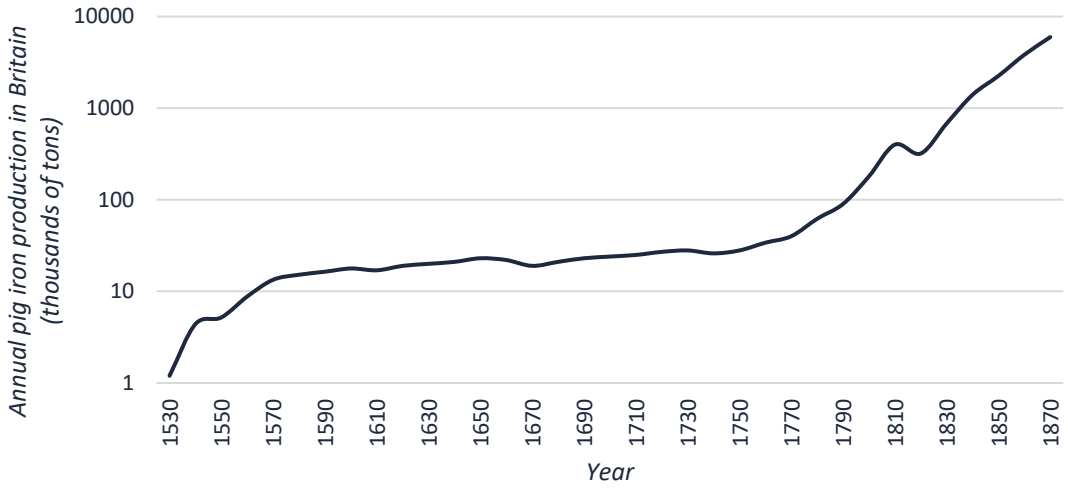


FIGURE 7: BRITAIN'S ENERGY MIX[35]



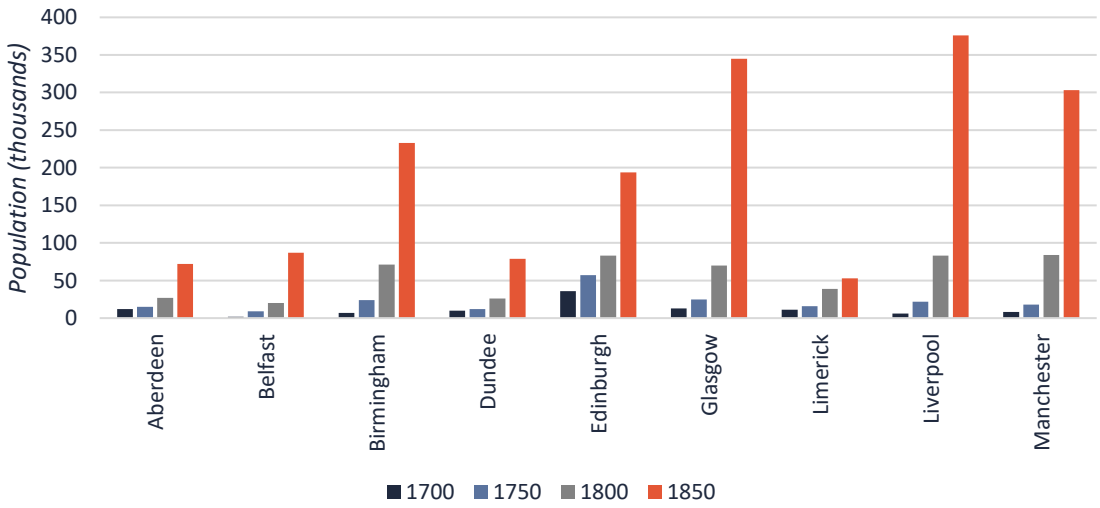
[35] Wrigley (2013).

FIGURE 8: PIG IRON PRODUCTION IN BRITAIN[36]



As the British economy exited its agrarian phase and entered an industrial one, cities blossomed because rising industrial productivity translated into rising urban productivity. Cities gained new advantages in production, and agglomerative advantages were at the core of this development. The new industrial economy specialised in the production of complex goods, involving dozens of material inputs. The manufacture of these goods often mandated coordination and know-how on the parts of the craftsmen and workers who assembled them. This required large and co-located labour forces. The most efficient way to achieve this involved congregating large numbers of specialist workers in big factories, where they could collaborate and learn from each other, and where high fixed-cost physical infrastructure could be shared. As deeper labour pools enabled superior labour specialisation, denser settlements enjoyed growing advantages in production and, as a result, experienced significant population growth.[37]

FIGURE 9: URBAN GROWTH IN BRITAIN[38]



[36] Riden (1977).
[37] On the role of specialisation in industrialisation, and on how this encourages urbanisation, see Goheen (1973).
[38] Brunt and García-Peñalos (2021).

Industrial production was also resource-intensive, requiring a consistent supply of basic materials. Although waterways, railways, and roads all underwent significant improvements throughout the nineteenth century, transportation remained expensive, with particularly high fixed costs. To minimise transportation costs, factories were often built in clusters, which permitted cost sharing, and in cities with relatively cheap access to coal or iron-ore supplies—either by virtue of their direct proximity to coalfields or iron mines (e.g., Glasgow) or by virtue of their proximity to well-connected waterways (e.g., London).[39] Where firms established operations, workers followed close behind. As the production advantages generated by agglomeration loomed larger, urban populations swelled in cities throughout Britain, and the flywheel leading to further production advantages began to spin.

The birth of modern London

London was a major beneficiary of industrialisation. As improvements in production advantages pulled in newcomers from Britain's rural regions, London's population size increased by a factor of 6.8 over the course of the nineteenth century. As London's population expanded, agglomeration economies compounded the city's underlying natural and artificial advantages, driving further population growth in turn. The following paragraphs demonstrate how this growth relates to the framework we've outlined in *Section 1*.

An underlying set of natural advantages in production positioned London for growth during this period. As British manufacturing achieved international prominence—by the nineteenth century, England was dubbed "the workshop of the world"—global trade assumed a central role in the country's domestic economy.[40] London captured some of the resultant gains. The city's proximity to the River Thames, a passage to the Atlantic, made London well-suited to hosting the "finishing industries" that performed the final stages of manufacturing production.[41] This turned London into a hub for trade.

New artificial advantages in production also benefitted London. For instance, London's extensive railway network, which was developed throughout the nineteenth century, supported the city's physical expansion by making it possible to commute from newly formed suburban areas to London's commercial core. One study concludes that without a railway network London's population in the early twentieth century would have remained 30 percent smaller.[42] Additionally, in the earlier phase of the Industrial Revolution, Britain's 4,000-mile network of canals helped link London to coalfields in the North.[43] (However, by the 1840s, canal-based transportation had been mostly replaced by the railroad.)

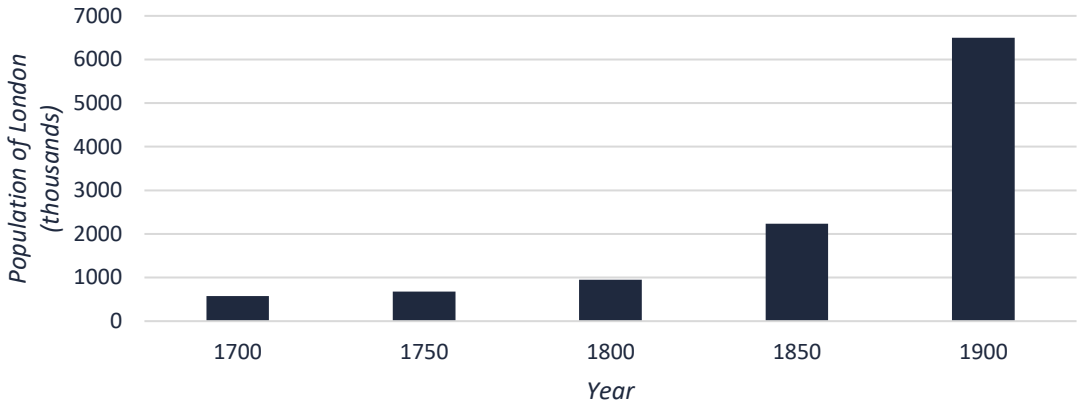
London also developed industrial clusters, which delivered the usual benefits associated with agglomeration. First, they facilitated learning. Aspiring craftsmen could train as apprentices with experts, and businesses could glean trade insights by observing their competitors. This cultivated a higher-skilled workforce and more competent firms. Second, clusters facilitated sharing by reducing transaction costs and permitting cost sharing. For instance, most commercial districts developed marketplaces dedicated to the exchange of their respective specialty goods. Finally, clusters enhanced matching. During a period in which information was mostly shared through face-to-face interactions, specialised markets smoothed trade. This improved matching between producers of final goods and end-use consumers, between the local buyers and sellers of intermediate goods, and

[39] Pollard (1981).
[40] This was supported by the new practice of installing copper sheathing on the hulls of ships, which considerably improved speeds and made long-distance hauls cheaper. See Solar and Ronnback (2015). Exports were also bolstered by Britain's imperial adventures, which opened up new markets abroad. See Ades and Glaeser (1994) for a broader examination of the role of mercantilist trade policies on London's development.
[41] Marshall (1987).
[42] Heblich, Redding, and Sturm (2018).
[43] Matthews (2016).

between local producers and the merchants who purchased their products for export. As London's clusters grew more established, Britain's most skilled workers and most ambitious entrepreneurs relocated to capture the agglomerative advantages that London increasingly provided.

Industrial Revolution-era London represents a paradigm case of the positive urban feedback loops we outlined earlier: as London got larger, its workers got more productive, which, in turn, made London larger and more productive still. This produced impressive results. By the dawn of the twentieth century, London was the most populous city in the world, and the wealthiest one, too.

FIGURE 10: LONDON'S POPULATION[44]



As our framework describes, urban growth is driven by shocks that enhance cities' natural, artificial, and agglomerative advantages in production and consumption. Above, we showed that, in the nineteenth-century, improvements in British agricultural productivity combined with Britain's transition to coal to create an industrial economy. Industrial production required the co-location of workers and firms. This improved production advantages in cities and triggered a wave of urbanisation. In the next section, we apply our framework to tell a very different story. Just as improvements in urban advantages pull people into cities, when urban advantages deteriorate (on a relative basis), people escape cities. As we will now see, this is what happened in the United States a century or so after Britain's Industrial Revolution.

SECTION 2.2 | SPRAWL IN TWENTIETH-CENTURY AMERICA: 1900-1980

Throughout the twentieth century, particularly in the decades that followed the Second World War, the United States witnessed extraordinary economic growth. Shored up by technological innovation, demographic change, and government investment, aggregate output roughly quadrupled in real terms between 1947 and 1987.[45] During that same period, and partly for the same reasons, America's economic geography underwent radical change. The dense cities that dominated the United States at the turn of the twentieth century gave way to an increasingly decentralised urban landscape. In particular, improved consumption advantages outside city centres—triggered at first by new transportation technologies and subsequently accelerated by changing cultural preferences—pushed Americans into suburbs and lower-density cities.

Urban centres as victims of their own success

From the late nineteenth century until the Second World War, the United States urbanised rapidly. Urbanisation in America was largely a function of the same forces that powered urbanisation in Britain

[44] Brunt and García-Peñalos (2021).
[45] Federal Reserve of St. Louis (2021).

a few decades earlier: would-be farmers left the hinterland to work in increasingly productive industrial sectors, which were concentrated in cities. Also like in Britain, successful cities in nineteenth-century America were rich in natural advantages in production (and in the associated artificial advantages built to exploit them). New York provides a clear illustration. New York's harbour housed a port that served as a gateway to global markets, and the Hudson River and East River linked the city to producers of basic materials on Long Island and in Upstate New York, which was itself well-connected by the Erie Canal.[46] Supported by a highly productive manufacturing sector, particularly in garments, and by enormous inflows of immigrants, New York City's population increased more than a hundred-fold between 1800 and 1930, from about 60,000 to 6.9 million.[47]

Although cities in the United States grew much more populous during this period, land-transportation costs remained significant, which meant that cities could only expand their physical boundaries modestly. With barely 2 percent of American households owning cars in 1910, urban workers had to walk or ride streetcars to commute to their jobs.[48] This imposed strict limits on urban sprawl. By 1940, nine of the top 10 largest cities in the United States had population densities greater than 10,000 people per square mile. In comparison, by 1990, seven of the top 10 largest American cities had population densities below 7,500 people per square mile. [49]

As urban density intensified, the costs of agglomeration started to mount, and "inner cities" became synonymous with poverty and crime. Fundamentally, cities became too densely occupied. As a result, residents increasingly wanted to leave for places that could offer superior advantages in consumption. Less populated, warm-weather cities provided one attractive option. Suburbs provided another.

The transition away from urban dominance was primarily led by two shocks: (1) technological and industrial innovations, combined with government-led infrastructure investment, that cut down the cost of land transportation and (2) new cultural preferences that favoured the amenities afforded by lower-density residential environments. These developments ultimately reduced the consumption advantages held by both central business districts and dense, monocentric cities, especially in the North, and they improved the consumption advantages held by both suburban peripheries and more sprawling cities, especially in the South. More generally, because of new transportation technology and infrastructure investment, the transportation-related advantages in production of the densest cities faded in significance. These adjustments reconfigured the spatial distribution of the American population. While sprawling cities like Los Angeles saw spectacular growth, places like Downtown Detroit depopulated.

The advent of the automobile

Cars and trucks were the technologies that contributed most obviously to the diffusion of America's population in this period. In short, the introduction of the mass-produced automobile squashed the cost of land transportation, enabling people to live further apart and firms to locate in places that lacked quick access to waterways or railway hubs. While the new technologies that were built into early automobiles, like internal combustion engines, were prerequisites for this shift, two attendant sets of developments mattered just as much.

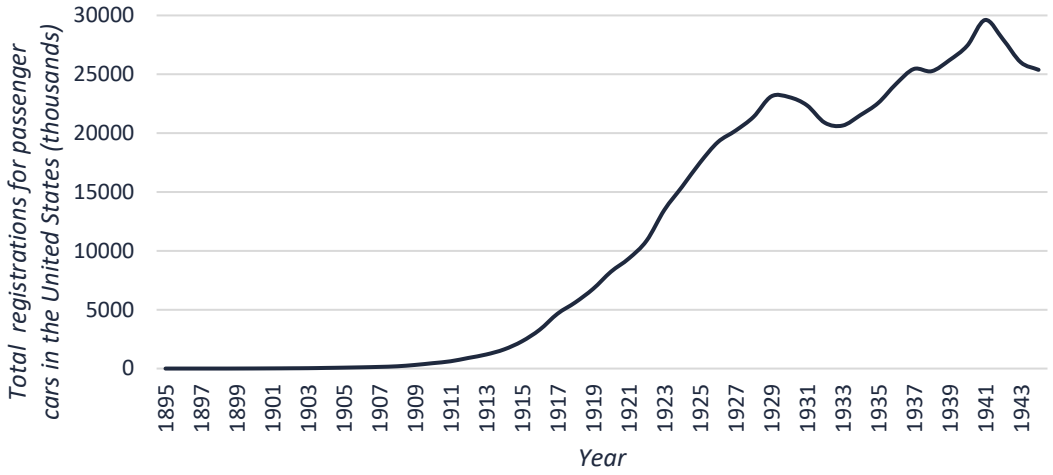
First, government investment in road infrastructure transformed cars into widely useful technologies. In other words, automobiles could make new places habitable only insofar as functional roads would cover a growing portion of America's landmass—and, increasingly, they did. The Interstate Highway System, created by the Federal Aid Highway Act of 1956 and officially completed in 1992, advanced this effort by stitching the country together with a vast network of standardised, multi-lane, high-

[46] Boustan, Bunten, and Hearey (2013)
[47] New York City Department of City Planning.
[48] NBER Macrohistory Database and United States Census Bureau.
[49] Glaeser (2005).

speed roads. This permitted the emergence of long-haul trucking and established an alternative to trains for the movement of goods over long distances.

Second, innovations in industrial organisation made cars commercially viable on a massive scale. Most famously, in 1913, Ford introduced a moving assembly line to manufacture Model-T vehicles at its Highland Park plant in Michigan. Ford's innovation slashed the amount of time it took to assemble its vehicles by more than 80 percent.[50] Moving assembly lines enabled Ford to produce cars with less labour, too. A more efficient production process quicky translated into lower prices for consumers. In 1912, Ford's Model-T cost about 100 percent of an American's average annual earnings; by 1924, that figure had already fallen by four-fifths to 20 percent.[51]

FIGURE 11: CAR REGISTRATIONS IN THE UNITED STATES



As cars became more useful and less expensive, Americans purchased them in much greater numbers. Between 1910 to 1940, automobile registrations for passenger cars increased from about 460,000 to about 27 million.[52] Car adoption was a process that reinforced itself. As private firms continued to develop infrastructure—like gas stations and parking lots—to meet the new demands of car users, car ownership grew even more appealing. As automobiles spread, medium-distance commutes became significantly less costly. In turn, lower commuting costs made older, higher-density cities (e.g., New York, Chicago, and Boston) less attractive relative to newer, lower-density cities (e.g., Los Angeles, Phoenix, and Dallas). Lower commuting costs also made city centres less attractive relative to suburban areas. This was because low commuting costs enabled people to work in central business districts, where many high-wage jobs remained (at least at first), without living within walking or streetcar distance of city centres. In other words, by reducing the costs of medium-distance commutes, automobiles enhanced consumption advantages in places that required residents to undertake medium-distance commutes. These were typically suburban areas and decentralised cities.

Businesses, too, embraced new transportation technologies. This produced lasting effects. Indeed, the real cost of transporting goods in the United States declined by 95 percent over the course of the twentieth century.[53] In turn, longstanding natural and artificial advantages in production faded, particularly ones that were tied to increasingly obsolete modes of transportation. More specifically, as trucks displaced ships and locomotives, cities with major ports or train depots lost some of their advantages in production. This adjustment damaged places like Chicago and New York, which could

[50] Ford corporate material.
[51] Gilder Lehrman Institute of American History.
[52] NBER Macrohistory Database.
[53] Glaeser and Kohlhase (2004).

no longer capture much value by monopolising the distribution of goods that had previously flowed through their transportation networks.[54]

Trucking also eroded the value of centralised distribution hubs in general, since trucks could travel directly to their final destinations, without first stopping to off-load goods in high fixed-cost intermediary hubs.[55] With respect to transportation infrastructure, the cost sharing benefits that agglomeration had previously conferred on mega-cities therefore declined. This nudged manufacturing out of the frozen industrial cities in the Northeast (e.g., New York City and Buffalo) and into warmer mid-sized cities, where costs were cheaper and where people increasingly wanted to live. Ultimately, New York State's share of the country's total manufacturing employment declined from about 15 percent in 1949 to less than five percent by 1997.[56]

The rise of the "consumer city"[57]

No longer obligated to live within a few miles of their workplaces, Americans increasingly based their location decisions on reasons to do with consumption. In particular, a growing number of city residents now sought to avoid the pitfalls of urban life, which had started to accumulate. Most notably, property crime in urban areas doubled from 1960 to 1990.[58] Rising crime encouraged the departures of middle-class city residents, who were less dependent on urban public transportation networks.[59] One study estimates that a typical American city lost one percent of its population for every 10 percent increase in the local crime rate.[60]

More generally, as Americans grew wealthier, they opted to spend more money to occupy more spacious homes and to send their children to better schools. This brought large numbers of well-off Americans to the suburbs, where these amenities were more readily available. All told, in the second part of the twentieth century, the share of metropolitan residents who lived in their metropolitan region's central city (i.e., rather than a suburb or a peripheral city) fell from 58 percent to 36 percent.[61] With some notable exceptions in places like Manhattan, white-collar jobs followed close behind, as firms sought both cheaper rents and closer proximity to suburban workers. By the end of the century, less than 16 percent of employment in a typical metropolitan area was located within three miles of the city centre.[62]

Americans also came to regard climate as an important kind of amenity in the latter half of the twentieth century. During this period, relatively young cities in the Sun Belt saw their populations boom. The sunny climates of cities in the Southeast and Southwest had long been mixed blessings because they often brought intolerably hot summer weather. But the mass adoption of air conditioning in the mid-twentieth century—which was jointly enabled by technological innovation and by increased prosperity—solved that problem. (Until the late 1950s, virtually no American households owned air conditioners. By 1980, almost 60 percent did.[63]) Americans flocked to warm-climate cities to capture this new advantage in consumption. Between 1950 and 2000, the Sun Belt's share of the United States' total population increased by more than 40 percent.[64] Unlike the older cities in the North, however, Sun Belt cities had not been built for a national economy that required

[54] Boustan, Bunten, and Hearey (2013).
[55] Glaeser (2005).
[56] Glaeser (2005).
[57] Glaeser (2001).
[58] Boustan, Bunten, and Hearey (2013)
[59] Glaeser (2005)
[60] Cullen and Levitt (1999).
[61] Boustan, Bunten, and Hearey (2013).
[62] Glaeser and Kahn (2001).
[63] Barreca et al. (2013).
[64] Hoffman (2007).

maximum urban density. These places were instead built for car users, who required the opposite. This further contributed to the decentralisation of America's economic geography in the twentieth century.

A new American spatial equilibrium

When we look at particular cities, we can see that housing prices and population flows reflected newly gained (or lost) advantages in consumption. From 1950 to 1960, the population of Los Angeles grew by 26 percent, while New York's population declined by one percent, Detroit's population declined by 10 percent, Boston's population declined by 13 percent, and Chicago's population declined by two percent.[65] In fact, out of America's 10 largest cities, Los Angeles was the only one that did not suffer de-population during that decade.

FIGURE 12: POPULATION GROWTH FOR 10 LARGEST AMERICAN CITIES IN 1930[66]

	Population in 1930	*Population growth by decade* 1950-60	1960-70	1970-80	1980-90	1990-2000	*Population in* 2000
New York	6,930,446	-1%	1%	-10%	4%	9%	8,008,278
Chicago	3,376,438	-2%	-5%	-11%	-7%	4%	2,896,016
Philadelphia	1,950,961	-3%	-3%	-13%	-6%	-4%	1,517,550
Detroit	1,568,662	-10%	-9%	-20%	-15%	-7%	951,270
Los Angeles	1,238,048	26%	14%	5%	17%	6%	3,694,820
Cleveland	900,429	-4%	-14%	-24%	-12%	-5%	478,403
St. Louis	821,960	-12%	-17%	-27%	-12%	-12%	348,189
Baltimore	804,874	-1%	-4%	-13%	-6%	-12%	651,154
Boston	781,188	-13%	-8%	-12%	2%	3%	589,141
Pittsburgh	669,817	-11%	-14%	-17%	-13%	-10%	334,563
United States	151,325,798	19%	13%	11%	9%	13%	281,421,906

A similar trend persisted in the several decades that followed. As consumption advantages rebalanced, demand-driven population growth led to comparatively higher rents in places now regarded as desirable. Between 1953 and 1986, housing costs in Los Angeles rose by a factor of 5.9, while housing costs in New York rose by a factor of 5.0.[67] By 2000, Detroit, Boston, and Chicago had smaller populations than they had in 1930. New York's population in 2000 was only 13 percent larger than it was in 1930. Meanwhile, Los Angeles's population tripled over that same period.[68]

SECTION 2.3 | THE FRAMEWORK IN THE CONTEXT OF HISTORY

Let's review. In nineteenth-century Britain, new agricultural technologies improved the sector's productivity, ultimately enabling the release of Britain's labour force into non-agricultural pursuits. At the same time, Britain's transition to a coal-dominated energy mix helped create a new mode of industrial production, which relied on congregating large numbers of workers in high-fixed cost factories, where they could share physical capital, raw materials, and energy. This enhanced productivity in places with natural resources (e.g., coal and iron ore), transportation infrastructure (e.g., railroads and waterways), and human density. These places were cities. To put this narrative in the context of the framework we developed in *Section 1*, a series of technological shocks enhanced the natural, artificial, and agglomerative advantages in production held by British cities. This effectively increased urban wages, which, in turn, led to large-scale urbanisation.

[65] Glaeser (2005).
[66] Glaeser (2005).
[67] United States Bureau of Labour Statistics.
[68] Glaeser (2005).

A century or so later, new transportation technologies combined with new cultural preferences to change the spatial distribution of consumption and production advantages in the United States. The advent of the automobile reduced transportation costs, turning suburban areas and decentralised cities, typically in the Sun Belt, into feasible places to live. In other words, the automobile improved the consumption advantages held by these locations relative to cities. During the same period, Americans came to prefer the living conditions that were available in lower-density environments. This further enhanced consumption advantages in suburbs and Sun Belt cities. As our framework anticipates, improved consumption advantages outside city-centres led Americans to depart old, dense city-centres in the North for Sun Belt cities and suburbs. The result was a more decentralised population.

Looking towards the future, we can apply that same framework. As we explore in the next section, a similar set of dynamics can help us understand how urban landscapes may evolve in the coming few decades, as new technological innovations and cultural changes continue to adjust the spatial distribution of local advantages and, in turn, shape where people choose to live.

SECTION 3 | THE FUTURE OF CITIES

A recent history

The latest major wave of urban change started to take shape in the 1980s, in the early days of the Information Age. By the dawn of the twenty-first century, the mass exodus out of major (or *alpha*) cities had reversed, and de-population in many secondary (or *beta* and *gamma*) cities had slowed.[69] Over the past few decades, we have witnessed a period of urban revival. Cities of all types achieved remarkable comebacks, and alpha cities fared best of all. As we describe in this section, their renaissance was largely the product of structural changes in advanced economies that were made possible by new information and communications technologies.

Ever since the Information Age took root, in the world's most dynamic places, production has been fuelled by human knowledge, not physical materials. Starting in the late 1980s, economic activity in advanced economies became increasingly dominated by high-skilled knowledge workers, which includes people who create intellectual property, provide professional services, or participate in financial markets.[70] In tandem with that development, information replaced physical materials as the basic ingredient of the most valuable forms of production. Superstar cities were major beneficiaries of this transformation. In fact, the Information Age enriched superstar cities over the past quarter-century in much the way that industrialisation powered growth in manufacturing-focused megacities a hundred years prior. In that earlier period, materials and manual labour were the most important economic inputs, and so the high cost of transporting those inputs made high-density cities, with the right natural advantages, optimal places for production.

Now, information and intellectual labour are the most important economic inputs. As the value of knowledge soared over the past four decades, advantages in production improved most in places where information flowed best—to wit, major cities with high concentrations of human capital and the densest information and communications infrastructure. The use requirements of early computing technologies illustrate this: until about a decade ago, computing encouraged the concentration of human capital in shared physical spaces. The workplace computer lived alongside a mainframe,

[69] We use the Globalisation and World Cities Research Network's urban classification system, which categorises cities based on their international connectedness and significance to the global economy. *Alpha* cities are the most integrated with the global economy. *Beta* cities are less integrated, and *gamma* cities are less integrated still.

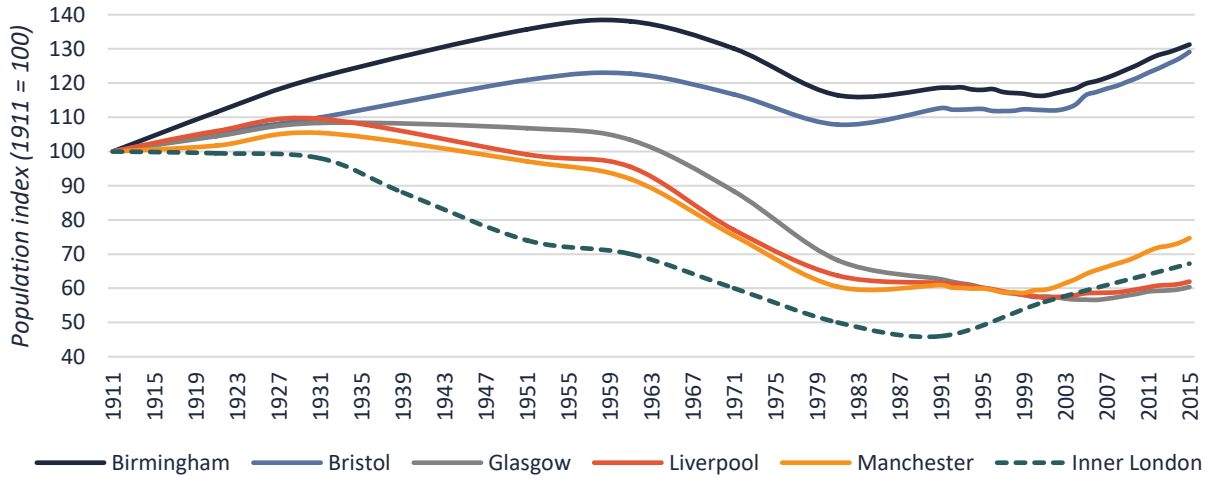
[70] The *information economy* is characterised by its emphasis on forms of production that involve scientific or technological innovation, intellectual property, or high levels of human capital. *Knowledge workers* are people employed in those core information-intensive activities.

copper telephone lines, a wired corporate intranet, and on-site servers. Using a PC meant going to a destination: you had to go to an office and sit at your desk to be on the network.[71] Computers involved high fixed costs, which could be shared as long as their users were co-located. This also favoured density. What's more, tools like Microsoft Teams and Slack, which allow physically distant workers to collaborate effectively in real time, were not available. Although Jack Nilles coined the term "telecommuting" in 1972, it remained a futuristic concept, not a genuine option, until very recently. The result was that superstar cities monopolised information-intensive sectors. As information-intensive sectors dominated advanced economies, so too did superstar cities.

For example, while New York gradually lost its manufacturing sector over the course of the twentieth century, an immensely more lucrative financial sector emerged in its place, dominated by firms for which knowledge was the key ingredient to success. As a result, New York thrived over the past quarter-century, enjoying more sustained population growth during that period than it had experienced since the early twentieth century.[72] New York's recent triumph exemplifies a much broader trend. Indeed, superstar cities like New York—i.e., alpha cities with the highest concentrations of the most high-skilled workers, elite firms, and influential cultural institutions—emerged as the urban victors of the late twentieth and early twenty-first centuries.

However, with the emergence of the latest wave of technological change in information and communication, we believe that that spatial equilibrium, in which alpha cities dominated and other cities fell further and further behind, is no longer stable. In this section, as we contemplate the future of cities, we focus on that theme. Drawing on the framework we developed in *Section 1*, we argue that the advantages in production that superstar cities monopolised in the last quarter-century (at least until five years ago) will increasingly disperse in the next decade, and that this will undercut alpha city dominance. Although superstar cities will remain on top, they will struggle. Meanwhile, secondary cities with affordable housing and high-quality amenities will thrive at their expense.

FIGURE 13: POPULATION GROWTH IN BRITISH CITIES: 1911-2015[73]



[71] See Thompson (2020).
[72] New York City Department of City Planning.
[73] Office for National Statistics.

SECTION 3.1 | CLOUD COMPUTING, MOBILE PHONES, AND THE REPLACEABILITY OF PHYSICAL PROXIMITY

Over the next decade, we think that urban landscapes will be defined by the adoption of new communications technologies that flatten the spatial distribution of advantages in production. Or, to put that idea in simpler words: technology will make different locations less different with respect to their abilities to harbour productive work. This is due to the fact that many forms of production are increasingly doable in a growing number of possible locations as a result of a decreased reliance of proximate physical information and communications infrastructure. While the Covid-19 pandemic brought this phenomenon into focus, its origins preceded the pandemic by about a decade.

In the past decade, cloud computing has made it possible for knowledge-economy workers to securely interact with company materials from any location in the world, including from outside the office. As computer (including mobile phone) hardware has gotten less expensive, more powerful, and more portable, this option has grown even more attractive. Over that same period, improvements in video-conferencing software, online messaging services, and mobile applications have smoothed communications between workers who live in different places. Information is more ubiquitous than ever, and the communication of that information is also cheaper. As the marginal cost of communication declines, the friction that physical distance once imposed is fading. As a result, distance increasingly matters less at the margin, at least in some ways.

Consider a few examples. New communications technologies permit firms to communicate internally with much greater frequency and ease, unifying teams that operate out of offices in different cities. Management consultants in London can review presentations in real time with colleagues who live in Frankfurt, and Big Tech managers in Menlo Park can supervise legions of software developers who are scattered throughout the globe. Communications technologies have also made it easier for outside firms to participate in remote local ecosystems, including markets for labour, specialised corporate services, and financing. Biotech entrepreneurs in Boston can pitch to prospective investors in Silicon Valley via video conference, without stepping on a plane, and start-ups in Salt Lake City can recruit MIT students online, without dispatching interviewers to attend in-person job fairs. Despite their various drawbacks, virtual alternatives to physical co-location now exist in many contexts, and these alternatives continue to improve. The result is that proximity no longer bestows the same privileges that it once did.

Looking to the future, we believe that this basic narrative will be replayed in new ways. The rise of remote and semi-remote (i.e., flexible) work is the latest episode. While it's too soon to tell just how pervasive remote work will ultimately prove—and there are good reasons for doubting that it will come to dominate in the way that some commentators now predict—the direction of change is clear: in the next few decades, people will work remotely (that is, outside of a company's local physical office) with greater frequency and in greater numbers than they did before the pandemic. This will probably further attenuate the benefits of physical co-location, as workers and managers become more adept at organising and executing remotely. Looking even further ahead, we might even contemplate a future in which virtual reality technologies create near-perfect substitutes for many types of in-person interactions.

So, technology is changing how people interact with each other and with the spaces they inhabit. Most significantly for our purposes, technology is making physical proximity less important—or, at least, less irreplaceable. How should we expect that change to shape where people choose to live and work? History offers some hints.

Decentralisation, then and now

Until the late twentieth century, the American economy was oriented around the production of physical things. This economic structure demanded constant engagement with physical objects, and physical objects had to come from somewhere. So long as transportation imposed high fixed costs, as shipping and locomotive infrastructure both did, producers could capture economies of scale by clustering around major transportation hubs. To further minimise producers' costs, transportation hubs were usually located in places with natural advantages that facilitated cheap access to raw materials. Meanwhile, in the absence of automobile transportation, most workers had to live within walking distance of their jobs. Under these conditions, it was most efficient to confine both production and habitation to a small number of dominant locations. This dynamic bred the hyper-dense cities of the early twentieth century.

In the mid-twentieth century, the mass adoption of motor vehicles disrupted that equilibrium. At that point, the structure of the American economy still required most workers to spend their time manipulating physical objects, but personal cars now permitted workers to commute to jobs in city centres or industrial parks from homes in the suburbs, and trucking did not require centrally located, high-fixed cost transportation hubs. As the cost of transporting workers and goods plummeted, the American population decentralised. This translated into a more egalitarian distribution of natural and artificial advantages in production and ultimately resulted in a more diffuse population.

On the other hand, in contemporary superstar cities, high-skilled workers mainly engage with information, not physical materials. Under these conditions, communication, not physical transportation, is what links together the highest-value factors of production. Considering that communication is just a special case of transportation—it is the transportation of information—we could think of virtual communication technologies as twenty-first century analogues to the automotive technologies that surfaced a hundred years ago. Virtual communication technologies permit information to travel longer distances, at higher speeds, and at lower costs than ever before. The spatial distribution of advantages in production is therefore flattening again. With information cheaply available almost anywhere, productive activities in information-intensive sectors are no longer confined to just a few places. This should also encourage the population to diffuse.

However, we do not believe that the world is about to return to suburban living. The move to the suburbs in the mid-twentieth century was preceded by a decline in the advantages in consumption in most mid-sized and large cities. There was little benefit of fleeing crime in New York to end up in downtown Pittsburgh, Cleveland or Chicago; people didn't leave dense, crime-ridden cities to move to another dense, crime-ridden city. Thus, warmer, newer cities built around the car were one escape hatch. The suburbs of major metropolitan areas were another. Suburbs were attractive because they offered proximity to centres of knowledge and culture, without the blights of urban cores; in other words, suburbs had clear consumption advantages over city centres. Today, on the other hand, the suburbs of major metropolitan areas only provide physical proximity to superstar cities in a world where that physical proximity no longer matters as much. What's more, suburbs require car-based commuting, even though fewer and fewer people want to (or can) drive to work today. Fundamentally, suburbs no longer offer (aggregate) consumption advantages relative to cities.

Meanwhile, secondary metropolitan areas, like downtown Pittsburgh, Cleveland and Chicago, are no longer the crime-ridden places they were four decades ago. Some of these cities have emerged as budding centres of knowledge and culture. Moreover, the cost of living in beta and gamma cities is substantially lower than it is in alpha cities. For example, a monthly wage of about $4,300 in Salt

Lake City supports a standard of living that would cost about $8,000 in San Francisco.[74] Indeed, fleeing a place like New York or San Francisco now—likely for cost of living and not crime reasons—could meaningfully improve a resident's overall quality of life, thanks to the improved consumption advantages that secondary cities now offer.

Secondary locations are also seeing significant improvements in their relative production advantages. Cities that previously stood no chance against superstar cities are increasingly able to tap into the key resource that's required for participation in information-intensive economies. For example, although Silicon Valley still contains a high concentration of information about technology businesses, information now flows in and out of Silicon Valley at virtually no cost. This has already started to do damage: the share of American venture capital deals agreed with Bay Area-based start-ups has steadily declined over the last decade, and that figure is expected to fall below 20 percent for the first time in 2021.[75] Start-ups in places like Austin, Denver, and Salt Lake City are securing funding instead.

Therefore, we see a move out of alpha cities (or at least slower in-migration) not necessarily to the suburbs of those same cities, but rather to beta and gamma cities, especially those that can confer other consumption advantages like good climates, lower taxes and natural beauty. That is what we think "decentralisation" will mean in the coming decades. While we believe this shift was already underway before Covid-19, the pandemic is increasing the speed at which this change takes place.

This transition is led by two shocks: (1) new communications technologies, which we described above, and (2) new communication norms, which have only recently started to crystallise. As the product of these shocks, the next phase of the Information Age is coming to be characterised by ultra-low communication costs. Although the cost of communication fell sharply in the past quarter-century, we will suggest that the cost of communication is now declining in a qualitatively new way.

Emerging technologies and new communication norms

Whether or not a new innovation gets adopted *en masse* always depends on background conditions that have little to do with the technical merits of the innovation itself. As we explained above, motor vehicles could not have slashed transportation costs without the highway infrastructure that actually made those technologies useful, or without the mode of industrial organisation that enabled manufacturers to cheaply produce automobiles at scale. By the same token, virtual communication technologies cannot realise their full potential without a certain type of cultural infrastructure in place, one that permits knowledge-economy workers to replace in-person communication with virtual communication in a large number of contexts. Until recently, this cultural infrastructure was relatively undeveloped.

Until only about five years ago, communication-intensive productive activities—business meetings, job interviews, academic conferences, everyday office-based work—were routinely conducted in-person, despite the fact that adequate virtual alternatives sometimes existed. Longstanding social conventions took for granted that certain activities were done in-person, and most people simply obliged. This dynamic was self-reinforcing. For example, if a start-up founder in Miami believed (whether correctly or not) that pitching prospective investors in San Francisco via video-conference would put her at a disadvantage relative to local competitors who would attend in-person meetings, then she had a good reason to incur the additional cost and arrange in-person meetings herself. It did not matter whether those investors *in fact* preferred to meet in-person—just that some entrepreneurs believed that they did. Cultural habits helped uphold the consensus that in-person

[74] Calculations by Numbeo (retrieved in 2021).
[75] *Pitchbook* (2020).

communication was the preferred kind of communication, and that consensus made itself reflexively true.[76]

In short, as long as cultural norms defaulted to a preference for in-person communication, virtual communication could not serve as a viable substitute. This meant that the apparent low cost of communication in the first phase of the Information Age was exaggerated. Even if it (financially) cost almost nothing to conduct a meeting via video conference during this period, if the meeting's participants strongly preferred to conduct their meeting face-to-face, that low cost would have mattered little in practice. The two options were not regarded as adequate substitutes. So long as virtual communication was not regarded as a legitimate substitute for in-person communication, the actual cost of communication was often best represented by the cost of in-person communication, not its cheaper inadequate substitute. This amounted to the cost of transporting relevant parties to a common location, using the same transportation technologies that we had used for decades. So, the *actual* cost of communication did not reflect the current state of technological progress.

However, change has been under way for some time. In the last several years, as virtual communication technologies gradually infiltrated more aspects of our lives, our conventions started to adjust. As members of younger generations, so-called *digital natives*, ascended in the workforce, workplace communication practices evolved, too. (We surveyed some of these practices at the start of this section.)

Most recently, the wholesale adoption of remote-work technologies during the Covid-19 pandemic shattered what remained of old communication conventions. At the start of the pandemic, almost all knowledge-sector jobs immediately switched to remote working arrangements. People adjusted, and new habits formed. Microsoft Teams, Slack, and Zoom are now firmly embedded in workplace communication practices. On top of all that, as efforts to reduce corporate carbon footprints intensify, some organisations may actively prefer their employees to travel less. All told, even though office-based work, in-person meetings, and business travel will surely not disappear, many of the communication conventions adopted during the pandemic will likely endure—particularly as the underlying technologies continue to improve, too. (And perhaps also partly for the same reason that in-person communication was tough to dislodge in the first place: habits tend to last.)

Here is the crucial point: more and more, virtual communication is treated as an adequate substitute for in-person communication. While the cost of virtual communication—and only virtual communication—declined sharply during the first phase of the Information Age, new communication conventions are blurring the distinction between virtual communication and in-person communication. As a result, the actual cost of communication is declining sharply now, too.

The newly emerging spatial equilibrium

As we saw earlier, when the cost of transportation declines, advantages in production and consumption tend to improve in more remote and lower-density locations. The present decline in communications costs should trigger an analogous adjustment. (Recall that communications costs are just a special kind of transportation cost.) According to our framework, when advantages in production or consumption improve in some locations, that place should experience population growth. We therefore believe that, as the cost of communication declines, more remote and lower-density locations will experience growth.

[76] It's worth emphasising that the point is not that virtual communication is equivalent to in-person communication. (It is not, at least not yet.) The point is just that, in many contexts, virtual communication could have replaced in-person communication without any net loss, but vestigial cultural norms got in the way.

As we suggested above, however, we do not expect suburbs to be the main beneficiaries of that adjustment. As people depart the metropolitan areas of alpha cities for the metropolitan areas of beta and gamma cities, we expect that they will tend to resettle in the urban cores of beta and gamma cities (i.e., rather than in those cities' suburbs). In the past, residents of alpha city metropolitan areas were often forced into the suburbs by the high housing costs of those areas' city centres and by urban disamenities, such as congestion and crime. These costs should become less important as populations decentralise, as we expect. First, housing in the urban cores of beta and gamma cities is much less expensive than it is in the urban cores of alpha cities. Congestion and crime tend to be milder in secondary city centres, too. As result, when people relocate to the metropolitan areas of beta and gamma cities, we think that they will generally live closer to city centres than they previously did, primarily because the urban option will be increasingly affordable to them. So, decentralisation does not imply suburbanisation; on the contrary, we think smaller cities will likely grow denser.

As we have already pointed out, secondary locations are becoming more information-rich and therefore more productive. As knowledge-sector workers increasingly perform high-wage activities—which were previously concentrated in a handful of superstar cities, like New York, San Francisco, and London—in dozens of new places—like Denver, Austin, and Edinburgh—wages in places like Denver, Austin, and Edinburgh should tend to rise (relative to wages in New York, San Francisco, and London). As our framework shows, when wages rise in a particular location, people tend to move to that place to capture the available gains. As the take-up of new communications technologies ultimately increases wages outside superstar cities, economic activity should decentralise, and secondary locations should therefore enjoy relatively strong growth. What kinds of secondary locations should we expect to do best?

As we close, we will try to offer some preliminary answers to that question. To do so, we should return to the first part of our discussion. There, we established that the total utility a person obtains by living in a particular location depends mainly on the level of wages, rents, and amenities, which she expects to receive by living in that place, and we assumed that people tend to live in the places where they are best-off. If advantages in production are coming to be distributed more evenly throughout space, as we just argued above, then wages should tend to equalise throughout space, too. This implies that historically high-wage places will be less capable of distinguishing themselves by offering the generous wage premiums that they did in the past. More generally, as wage levels do a worse job of individuating different places, the two remaining considerations—the price of local rents and the quality of local amenities—should wield more influence in individuals' location choices. In other words, if wages are less fixed by location, then non-wage considerations will increasingly determine where people decide to live.

According to our framework, this adjustment should produce several effects. First, it means that locations that impose high rents should seem increasingly unattractive to residents. For example, superstar cities—where wage premiums previously helped to justify exorbitant housing costs—will likely meet tough competition from lower-rent cities, which can attract residents with increasingly competitive wages. At the same time, and for the same reason, advantages in consumption will play increasingly important roles: when people can find well-paying jobs in more places, they should be less inclined to put up with places that cannot provide good qualities of life. As a result, demand for housing should rise in cities with high-quality amenities, *especially* ones that technology cannot easily replace: pleasant climates, attractive tax policies, good restaurants, natural attractions, and trendy cultural scenes. As we suggested above, secondary cities are well positioned to benefit as a result. While housing costs tend to be comparatively low in secondary cities, these cities are often large enough to sustain the artificial advantages that depend on agglomeration (e.g., entertainment venues,

museums, and niche restaurants). Secondary cities can also be rich in natural advantages in consumption. Think of Lisbon's nearby beaches or Salt Lake City's proximity to ski mountains.

Finally, we should expect dislocation to occur most in countries with a great deal of internal variation with respect to the consumption advantages that we believe are coming to matter most. In homogenous countries, residents of alpha cities cannot capture significant non-pecuniary gains by relocating to gamma or beta cities. For example, in countries where the climate is roughly uniform throughout, residents of cold cities cannot relocate to warm cities to capture the benefits of sunnier weather. On the other hand, in climatically diverse countries, this option is available. More generally, in internally heterogenous countries, beta and gamma cities can distinguish themselves by providing residents with more favourable tax policies, climates, and natural attractions.

Compare the United States with the United Kingdom. In the United States, individual states have the power to levy their own income taxes. This leads to a kind of fiscal diversity that the United Kingdom lacks: while Florida has no state income tax, New York's ranges from 4 percent to about 11 percent. (Similarly, the quality and scope of local government services are more variable in America.) The United States also contains more climatic diversity than the United Kingdom. Chicago's winters are frigid, while Phoenix's are mild. Although residents of London could pay lower rents by moving to Manchester, they would not escape London's rainy winters. In short, residents of American superstar cities have a much larger menu of differentiated alternatives than do residents of London. Something similar is true of residents of the European Union. French citizens can freely move to Spain in much the way that residents of London can freely move to Manchester. Yet Europe, like the United States, contains much more climatic and fiscal diversity than the United Kingdom. While Germany has cold winters, and Finland has high taxes, Malta is sunny, and Andorra's maximum tax rate on personal income is 10 percent. Residents of European superstar cities, like Paris, also have a much larger menu of differentiated alternatives than do residents of London. We therefore expect to see the most extreme disruption in regions with the most spatial diversity in consumption advantages, like the United States and Continental Europe.

As new communications technologies flatten the spatial distribution of production advantages, people will relocate. We think that consumption advantage-rich secondary cities will enjoy robust population growth as a result of this adjustment. As we argued above, technology is now providing residents of alpha cities with opportunities to flock to beta and gamma cities that offer superior amenities, more favourable tax policies, and lower-cost housing. As people increasingly capture those opportunities over the course of the next decade, we will yet again move into a new spatial equilibrium.