3/October/18:

Winners and Losers in the Urban System*

Anthony J. Venables,

University of Oxford, CEPR and International Growth Centre.

Abstract

The urban system divides into cities that are internationally competitive in goods or service production and those that produce just for the domestic economy. Agglomeration economies make it difficult to start new tradable activities, with the consequence that negative trade shocks – and also many positive ones – have the effect of increasing the number of a country's cities that produce just non-tradables. This paper sets out a simple model of this process and shows how such shocks widen the gap between city types, causing cities specialising in non-tradables to lose population and income (especially land rent). Responses to this growing urban dichotomy are discussed.

Keywords: Globalisation, urban, de-industrialisation, polarisation.
JEL classification: F12, F60, R11, R12.

* Paper written for 'Urban Empires', eds. E. Glaeser, K. Kourtit and P. Nijkamp. Thanks to referees for useful comments.

Author's address:
A. J. Venables,
Department of Economics,
Manor Road,
Oxford OX1 3UQ, UK

tony.venables@economics.ox.ac.uk

1. Introduction

The success of many cities is rightly celebrated by authors writing in this and other volumes.[1] Triumphant cities are centres of innovation, driving up incomes for their inhabitants and delivering the fruits of technical progress to the wider world. Their success is based on cumulative processes of learning and productivity growth that go under the collective title of agglomeration economies. Such spatially concentrated economies of scale are particularly important in sectors where high frequency exchange of information is important, such as high-technology innovation, creative activities, media, and in the financial sector.

While triumphant cities are part of the new urban world, study and analysis of this world is incomplete without incorporating several further facts. One is that most economic activity does not involve high frequency exchange of information or any of the other drivers of agglomeration. Most of a modern economy is engaged in relatively low-skill service activity in sectors such as retail, personal services, leisure, government, logistics, property maintenance, construction and so on. These are sectors that show little propensity to agglomerate spatially or that would create much by way of increasing returns to scale if they were to do so. A further 10% or so of the labour force is engaged in traditional manufacturing sectors. There are reasons for manufacturing sub-sectors to cluster together (e.g. auto-industry clusters forming to concentrate components and assembly) but these are no longer sectors that are driving dynamic cities. Furthermore, many manufacturing supply chains are global, rather than urban. The rest of the labor force is employed in the primary sectors of agriculture, resource extraction, and the processing of these products. The location of these sectors is driven largely by availability of land and resources, and these sectors are not principal drivers of urban growth.

Following from this, most cities do not have much or any employment in the sectors that exhibit strong agglomeration economies.[2] These sectors are inherently concentrated in relatively few places. There are simply not enough such sectors (defined for present purposes as distinct sectors not closely linked through agglomeration forces) for it to be possible for them to play a major role in more than a small subset of the cities of the developed and emerging world – and even less so in cities of the less developed world.

A further fact is the divergence of city performance. While some cities have become global superstars and others at least national or regional champions, many more have performed relatively poorly. There is polarisation of the urban system, with booming cities mirrored by a larger number of stagnant areas of relative, if not absolute, decline. Florida et al. (2017)

---

[1] Notably Glaeser (2012), Moretti (2013).

[2] Glaeser and Gottlieb (2009) demonstrate that in the US agglomeration appears strong only in relatively skilled labour abundant cities.

report how 49 'alpha cities' have a dominant share in output, venture capital, and wealth. Focusing on house prices, Gyourko et al. (2013) point to the widening gap between prices the most expensive and the average US cities. Storper (2018) reports that metropolitan areas with population greater than 1 million created three-quarters of the net employment increase in the USA from 2010 to 2016. The lagging performance of many towns and cities shows up not only in economics, but also in social, health, and political outcomes, and have been widely documented.[3]

Divergent performance of cities is one element of a broader picture of increasing regional inequality within many countries in the developed world (see Storper 2018 for a review of evidence and of alternative explanations of this increase). The present paper focuses on the urban element, and in particular on the response of cities to economic shocks. The traditional specialisations of many cities have been undermined by globalisation and by changes in technology. When a city is hit by a negative shock of this type, what economic adjustment processes operate? Do they disperse the impact of the shock across places in the country, consistent with convergent performance of different cities? Or do they tend to amplify the impact on the directly affected place, thereby creating divergence? Standard economic reasoning suggests that adjustment will be a force for convergence, as there are complementarities between economic activity in different places. Thus, a positive shock to one place would typically be expected to benefit others, as rising income in one city increases demand for goods supplied from other cities and creates upwards pressure on wages throughout the economy. Similarly, the impact of a negative shock will be spread around the system.

The core argument of this paper is that – while these complementarities are present – they are dominated by an alignment of factors that tend to amplify divergence. These create particularly powerful increasing returns in some sectors and cities, but not in others. These agglomeration forces are the product of two factors. One is a range of technological conditions and market imperfections that create increasing returns to scale. The other is tradability, allowing production in a place to expand without encountering diminishing returns imposed by the limited size of the local market. Thus, the net strength of agglomeration forces depends on both the strength of increasing returns to scale and on whether a city's output is internationally tradable, supplying a global market, or is 'non-tradable', supplying just a local or national market.

One can imagine two worlds. In one, increasing returns and tradability are not aligned. Trade is largely in primary products (not subject to increasing returns) while sectors that

[3] There is a large American literature on the topic, see for example Cramer (2016) and references in Storper (2018). Rodriguez-Pose (2017) adds a European angle.

exhibit increasing returns (in this world manufactures and some services) are largely traded within rather than between countries. In this world national centers of sectors such as engineering or financial services may form, but they are not of global scale. Sectors with the most agglomeration potential are constrained by a relatively low degree of tradability, while the most internationally tradable goods (primary products) are not subject to increasing returns.

In the other world, some sectors that are technologically prone to increasing returns – financial services, creative and innovative sectors and some manufactures – are also internationally tradable. At the same time sectors that are neither globally tradable nor subject to increasing returns – personal services, customer services, government – become an increasingly large share of the economy. In this world the forces of increasing returns and tradability have become aligned, with two consequences. One is that centres of increasing returns activities may become much larger, with all the implications that follow in terms of productivity, city size and land prices. The other is that trade shocks now operate on these increasing returns sectors and cities, rather than on sectors that are more oriented to primary production. As we will argue in this paper, economic adjustment mechanisms can then amplify rather than dampen the impact of trade shocks.

This paper is not an empirical or historical study of these two possible worlds. Instead, it takes the second world in which increasing returns and tradability are aligned and thinks through the implications of this for the way in which the economy responds to shocks and the possibility of persistent urban divergence. It does so by developing a simple economic model.[4] This is reduced to bare essentials, focussing attention on the distinction between two types of cities. Those that are competitive in increasing returns and internationally tradable sectors, and those where employment is concentrated in non-tradables, producing only for the domestic market. This distinction arises endogenously, as all cities have identical fundamentals but nevertheless develop different specialisations (in the manner of Henderson 1974), The dichotomy emerges even though the model abstracts from differences between cities assuming, amongst other things, that there is a single type of labor, undifferentiated by skill level.[5] The model allows for central arguments to be made in a compact way, and is the subject of section 2 of the paper.

Sections 3 and 4 look at shocks, and at the short- and long-run response of the urban system to these shocks. We argue that the main effect of import competition (section 3) is to cause

[4] A fuller and more technical version of the model is given in Venables (2018).

[5] The importance of skill differentials is highlighted by Diamond (2016) who argues that changes in local labor demand in the US is associated with both an increasing graduate premium and further concentration of graduates in high wage, high rent cities. Venables (2018) extends the present model to contain two skill levels.

cities to switch type, and this increases polarization of the urban system. Booming export sectors (section 4) can have exactly the same effect, benefiting some cities but damaging others. Section 5 of the paper turns to possible policy responses.

2. City systems and the urban dichotomy:

Cities are expensive, yet productive, centers of economic activity. Their size and economic performance are the outcome of the balance between 'urban costs' – the additional costs of living that are incurred in an urban environment – and urban productivity, the productivity premium associated with intense economic interaction and agglomeration economies.

Urban costs are manifest simply in the fact that larger cities are more expensive than small ones. The higher costs are partly real costs, such as commuting, congestion, and provision of infrastructure needed to mitigate these costs. And they are partly land rent, not a real resource cost but a transfer payment, i.e. a cost to the renter and a benefit to the landowner. For the purposes of the analytical arguments and thought experiments of this paper we suppose that urban costs increase with city size, and the rate at which they do so is the same for all cities; in other words, cities are identical in their fundamentals although not necessarily in the economic outcomes that emerge.

Turning to urban productivity, the productivity advantages of the economic scale and density that cities can deliver are well known. The agglomeration forces that support increasing returns come from well-documented sources: scale and specialisation, matching and knowledge spillovers.[6] Each of these is facilitated by close and intense economic interaction often involving face-to-face contact. Thus, thick labour markets enable better matching of workers to firms' skill requirements. Better communication between firms and their customers and suppliers enables knowledge spillovers, better product design and timely production. A larger local market enables development of a larger network or more specialised suppliers. A good example is given by specialist workers or suppliers. The larger the market the more likely it is to be worthwhile for an individual to specialise and hone skills in producing a particular good or service. The specialist will be paid for the product or service supplied but, depending on market conditions, is unlikely to capture the full benefit created. Since the benefit is split between the supplier and her customers there is a positive (pecuniary) externality. This creates a positive feedback – more firms will be attracted to the place to receive the benefit, growing the market, further increasing the returns to specialisation, and so on. This is the classic process of cluster formation, and occurs even

[6] For a survey of analytical work in the area see Duranton and Puga (2004), and of empirical work Rosenthal and Strange (2004) and Combes and Gobillon (2015).

though these interactions might apply to only a small part of the final product – for example R&D but not assembly. The strength of these mechanisms varies across sectors, and there is some consensus that the effects are strongest in high-skill activities and/ or creative activities, such as R&D intensive sectors, media, business services and finance.

As noted in the introduction, these productivity effects can be offset by price effects, so the 'tradability' of the sector's output matters. 'Tradable' goods have price set on the world market, essentially independent of the output of any one producer or city. 'Non-tradables' have sales limited by the size of the local market, so increasing supply reduces the price. For the purposes of this paper we simply assume that increasing returns and tradability are perfectly aligned at the sector level, so there are two-types of productive sectors. Sectors with significant agglomeration economies are also readily tradable and we refer to sectors of this type as T-sectors, T standing for tradable, or perhaps technology. Other sectors we refer to as N-sectors, standing for non-tradables. These include services that have to be consumed at the place of production (haircuts and restaurant meals), and those that are traded nationally but not internationally – the maintenance, logistics, customer service centers, etc, that form much of urban activity. These sectors may be produced under diminishing, constant, or weakly increasing returns to scale. However, the fact that they are traded on a national market, rather than on the much larger global market, means that an increase in supply of these products decreases their prices by a relatively large amount. Thus, N-sectors exhibit overall diminishing returns to scale.

A consequence of these assumptions is that cities specialise. Some cities specialise in sector-T activities, and we refer to these as type-T cities.[7] The remainder go to the default option, specialising in sector-N goods, and are referred to as type-N cities. This specialisation follows from the presence of agglomeration economies operating at the city level and within a particular productive sector (localisation economies). This is the fundamental dichotomy in city types that emerges as a property of equilibrium and, as we shall see, can lead to divergence of economic performance.

What about the economic adjustment mechanisms that equilibrate the city system in response to shocks? In the international context we expect shocks to be largely absorbed in wages, and it is this that enables adjustment. Thus, if a *country's* export sector has a negative shock the adjustment mechanism is a real depreciation, i.e. a reduction in its wage and unit costs relative to its trading partners, and this reduction continues until other sectors become competitive. However, within a country the performance of a city depends largely on its

[7] For simplicity, we assume that all sector-N goods are tradable nationally (i.e. we ignored haircuts and restaurant meals), so that specialisation is complete. Relaxing this assumption complicates analysis but does not qualitatively change it.

absolute advantage, not its comparative advantage. Thus, if a *city* within a country suffers a negative shock there may be little flexibility of relative wages between regions because factor mobility means that labour markets are tightly integrated. Of course, there are some immobile factors, such as land and houses. Their prices will fall in the adversely affected region but since these factors only represent a small fraction of costs they have little leverage in bringing other sectors to the point of competitiveness.

There are two distinct reasons for the limited flexibility of spatial wage differentials. One is that there may be wage rigidities and labour market imperfections, arising perhaps from national wage agreements or conventions that ignore local labour market conditions. Such rigidities can create city specific unemployment, and underpins much of the empirical work of labour economists.[8] The other mechanism is simply that there is – in the long-run at least – high intra-country mobility of workers in response to real wage differentials, and this places bounds the extent to which nominal wages vary within a country. In this paper we take the latter route, so there is perfect labour mobility; nominal wages may differ across cities, but real wages are equalised by inter-city migration that leads to adjustment in land rents and house prices.

Adjustment is also shaped by the fact that it is hard to start a type-T activity in a place where the activity is not already operating. Agglomeration economies mean that the returns to investing in a place depend on who else is (or is expected to be) there. This in turn creates a first-mover problem: no one wants to move to a new place while uncertain about its future development. Coordinated action by firms – or a 'large developer' – can solve the problem, but without this no firm wants to be the first in the sector to establish in a new location, uncertain as to whether it will be followed by other firms, and hence cluster creation and consequent high levels of productivity.

It follows from this that the division of cities between the two types is not uniquely determined. The proportion of cities of each type lies in a range, not at a single point. This matters greatly for the adjustment of the system to shocks as will be discussed in following sections. But first, to understand why this range exists, it is useful to think about Figure 1a, based on the full description of the model as set out in the appendix.

The horizontal axis of Figure 1a has the division of a country's cities between types. The number of cities in the country is set at 100, and the number of type-T cities is measured from left to right, the number of type-N from right to left. Thus, at point AA there are 40 type-T cities and 60 type-N. The vertical axis has nominal wages paid, with W_T being those paid by

[8] Notably the work of Autor, Dorn and Hanson (reviewed in Autor et al. 2016) who look at the impact of trade on regional and urban performance, and in particular the impact of Chinese competition on the US labour market.

T-sector firms in type-T cities, and W_N those paid by type-N firms (which, in equilibrium, are all in type-N cities). At point AA these *nominal* wages are higher in type-T cities than type-N, but this is a long-run equilibrium where no worker wants to migrate. The reason is that type-T cities are larger, so workers face higher urban costs of commuting and rent and hence require higher nominal wages to equalise real wages. Total rents earned in a city of each type are given in the lower of the figure. At point AA these are higher in type-T cities, reflecting both larger population and land area in these cities, and higher rents per unit land.

The critical point about this framework is that the equilibrium division of cities between types is not uniquely determined. Any division between {*Tmin, Tmax*} is an equilibrium. Explaining this requires going behind the curves on the figure. The curve W_N gives the wage that is paid in type-N cities if all the sector-N producers in these cities are breaking even, making zero abnormal profits. It slopes upwards because, moving to the right, there are fewer type-N cities, and hence less supply of type-N goods that therefore fetch a higher price; firms enter and jobs are created, this increasing nominal wages.[9] The curve W_T is the corresponding curve for type-T cities. The price of type-T goods is fixed on world markets (removing a potential source of diminishing returns). But moving to the right there are more type-T cities, each of which is therefore smaller. Because each city is smaller it has lower agglomeration economies and lower productivity, this dragging down wages.

The straight line at the bottom of the figure, W_{TN}, gives the wage that a sector-T firm could pay if it were to set up in a type-N city; doing so, it receives no agglomeration benefits from other firms in the same sector, so has low productivity. The gap W_{TN} minus W_T is the productivity disadvantage that would be suffered were a sector-T firm to leave a type-T city and produce without the benefits from being in its sector-T cluster.

Why is any division of cities between {*Tmin, Tmax*} an equilibrium? The reasoning is best seen by looking at divisions that are not equilibria. To the right of *Tmax* (more type-T cities) it would be profitable for a firm producing sector-N goods to set up in a type-T city; it would earn W_N per worker, but the going wage rate in a type-T city is less than this, W_T. Essentially, to the right of *Tmax* there are so few type-N cities that the price of N-goods is high enough for it to be profitable to set up a type-N firm even in a type-T city.

To the left of *Tmin* it would be profitable to produce sector-T goods in a type-N city. Productivity in T-production is lower as agglomeration economies are foregone. However, there are so many type-N cities that W_N is very low. Even without the benefit of agglomeration the per worker earnings of a T-sector firm in a type N-city, W_{TN}, are greater

[9] Fewer type-N cities mean that each such city is larger, but less than proportionately, so that the supply of type-N goods is smaller and their price higher.

than the going wage in a type-N city, W_N. Situations with fewer than *Tmin* tradable cities are therefore not equilibria – some firms will want to change location.

In the interval {*Tmin Tmax*} neither of these arguments apply, so there is no incentive for any firm to relocate. The fundamental reason is the coordination failure – no firm wants to be the first to establish T-sector production in a place with no existing production in the sector.[10] This framework and example shows how cities of different types coexist – even over a range of divisions. With this set up, how does this system react to shocks? As these occur, who gains and who loses?

[10] In the absence of coordination failure (e.g. if a large developer can create the coordinated establishment of a larger number of firms) the equilibrium set would shrink to point S on Figure 1a. At this point cities of each type are the same size.

Figure 1a: Nominal wages paid by each type of firm in each type of city

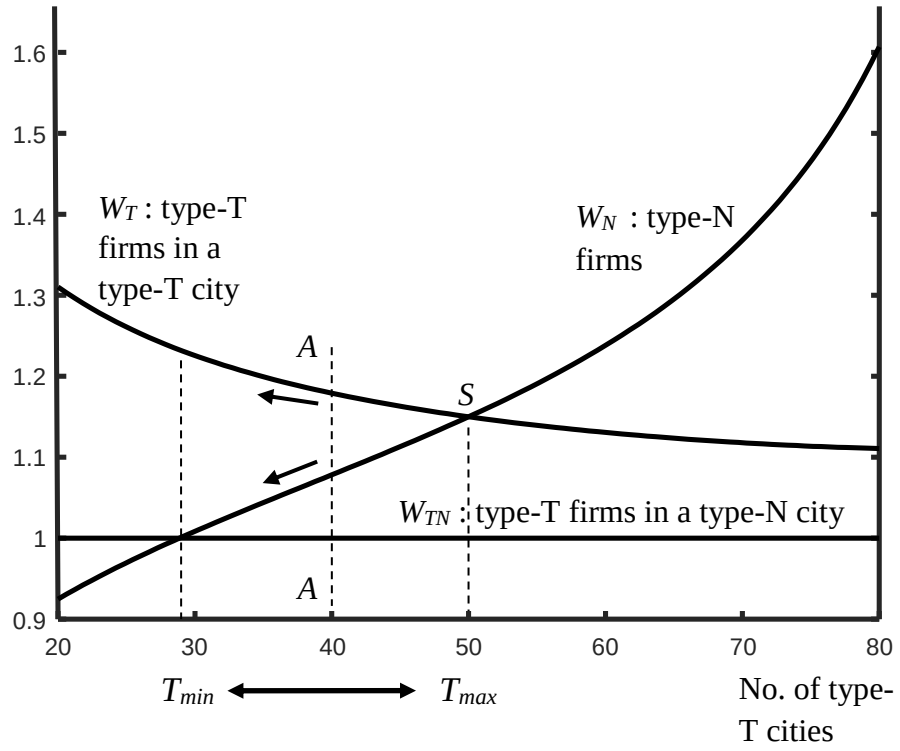
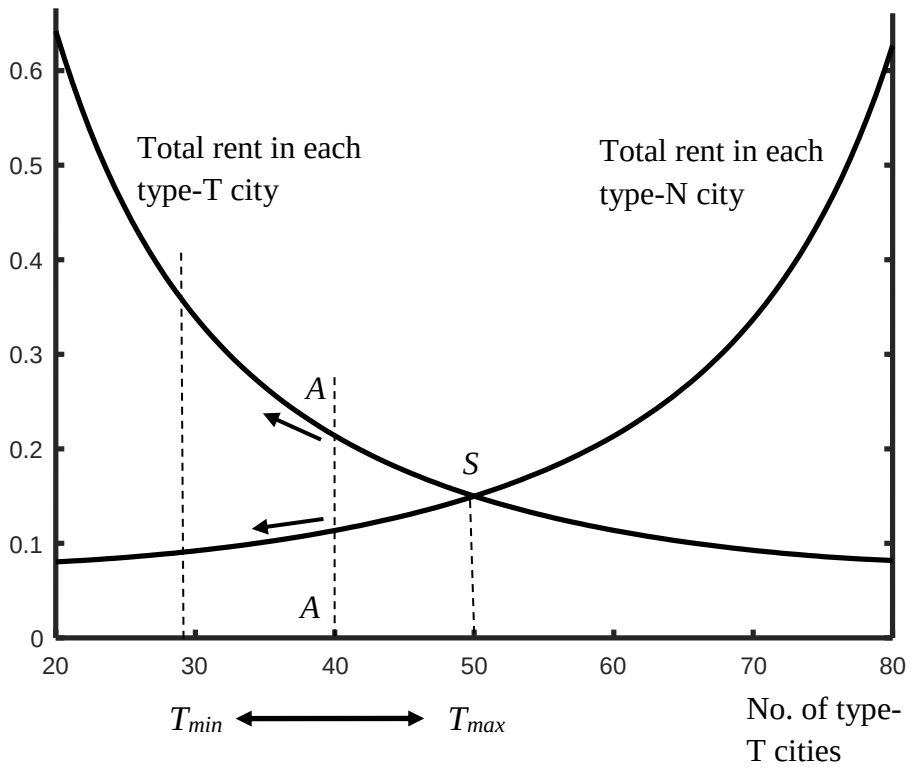


Figure 1b: Real rents earned in each city



3. Import competition

In recent years the competitive advantage of high-income countries in many sectors has been removed by global competition. Some of these sectors were concentrated in type-T cities, specialising in sectors such as textiles and steel. How does the economy adjust to such exogenous shocks? The short-run impact is immediate and direct; the affected city simply loses employment in the sector. In the medium to long-run this elicits two responses; out-migration from the city and a falling nominal wage. When does this process stop?

The answer comes from Figure 1a. Taking AA as the initial position the affected city drops down the vertical dashed line AA. Initially paying wages W_T, loss of its T-sector means that the nominal wage falls until level W_N, at which point non-tradable sector firms find it profitable to set up in the city. The key point is that – in the interior of the range {*Tmin, Tmax*} – W_N is greater than W_{TN}, the wage at which it would become profitable to establish a new sector-T activity from scratch. The city therefore switches from being type-T to type-N. The economy is flexible enough for new jobs to be created, but the first-mover problem means that these are not in T-sectors.[11] Instead, the jobs are created in sectors we have classified as type-N, i.e. customer services, warehousing, logistics and perhaps the (lower value) back-room jobs of the financial and insurance sectors. The nominal wage fall is associated with out-migration from the affected city, and hence lower urban costs and land rents.

There are further implications throughout the economy. As affected cities switch from type-T to type-N so the equilibrium moves to the left on Figure 1a (as illustrated by the arrows). There is now increased supply of sector-N goods, so their price falls and nominal wages, W_N, fall in *all* competing type-N cities. This leads to out-migration from all type-N cities which partially, but not completely, offsets the increase in supply of sector-N goods created by additional type-N cities. The final resting place is as follows:

- Affected cities switch from type-T to type-N.
- All type-N cities are smaller and have lower nominal wage than before.
- The share of the population in type-N cities has increased.
- Remaining type-T cities have higher nominal wages and are larger than before.

Notice that in this process real wages are equalised across city types (following directly from the assumption of perfect labor mobility) although nominal wages diverge. The principal losers are people who own land in type-N cities – all type-N cities, not just those that suffer

[11] It would take a large wage reduction to induce a new T-sector to move in. But the space has already been taken by N-sector production that does not face the first-mover disadvantage.

the direct shock, as the effect is shared via a fall in the relative price of sector-N goods. The gainers are, of course, landowners in the type-T cities. These cities boom as there is an influx of labor, possibly enhancing agglomeration economies and raising productivity; rents are bid up, the amount depending on the supply of land and housing in the city. The dependence of rents on the urban structure of the country is illustrated in Figure 1b. The curves give the total rent generated in a single city of each type; this is decreasing and convex in the number of cities of that type, as more cities means each has smaller population and area, and lower rent per unit area. Starting from a position in which type-T cities are larger than type-N (i.e. anywhere to the left of point S on Figure 1b), the shock causes further divergence of city size and rents. Since rents are increasing and convex in city size, the share of rents in the economy as a whole rises.

The message is that there is an adjustment mechanism, so affected cities converge to a new equilibrium. However, the dichotomy between city types increases, with type-T getting larger and type-N getting smaller. The number of people living in type-N cities increases, the share of rents in the economy increases, and land-owners in remaining type-T cities are the beneficiaries.

4. Export booms

Some tradable sectors receive positive shocks – growing world demand for their output or technical change that raises the productivity of the sector in which they are specialised. The effect of a positive shock of this type depends on whether or not it affects all tradable sectors. In the unlikely event that all T-sectors are affected, it spills over positively throughout the city system. The direct effect is to raise wages in tradable sectors and cities (this shifting the W_T curve on Figure 1 upwards). This sets in train two other forces. One is that this additional income generates additional spending on sector-N goods so their price, and hence the wage in type-N cities, W_N, increases. The increase in W_N is typically less than the increase in W_T, so there is migration from type-N cities to type-T until real wages have once again been equalised. Overall, this is a story of complementarity; booming tradable goods sectors also benefit the rest of the economy.

Tradable goods sectors are, in reality, heterogeneous. Most obviously, there are ‘new’ sectors in technology, creative sectors, and finance, and ‘old’ ones, such as many manufacturing sectors. What if the positive shock is restricted to the ‘new’ sectors? These sectors will expand, raising both the nominal wage and employment in affected cities. This wage increase is however a *negative* shock for old-sector tradable cities. They are competing for labor and the wage has gone up, undermining their competiveness. If the survival of

some of these sectors is marginal then the wage increase puts them out of business, and further effects are as described in section 3. Old-sector type-T cities switch to type-N production, the supply of N-goods increases, reducing their price, reducing W_N, and leading to contraction of all type-N cities. The net effect is that, while tradable cities that receive the positive shock boom, other tradable cities and all non-tradable cities are negatively affected. And as before, much of the effect ends up in changes in rents, with these booming in the growing cities and falling everywhere else. In short, a positive shock to some tradable sectors has a negative impact elsewhere, tending to polarise the urban system.

The intuition behind this result is best understood in terms of the 'Dutch disease'. This is a term from resource economics and refers to the fact that countries with large exports of natural resources (the phrase was coined in the context of natural gas exports from Holland in the 1970s) will tend to have appreciated exchange rates, a consequence of their large foreign exchange earnings. The appreciated exchange rate damages the country's non-resource tradable goods sectors and was argued to lead to de-industrialisation. The argument extends beyond the context of natural resources to other booming tradables sectors. Thus, it is suggested that the strength of London's financial services sector is damaging to the rest of the economy, via this mechanism of raising the exchange rate and making other tradable sectors – and cities – uncompetitive.

While this is the intuitive reasoning, the full formal analysis is given in Venables (2018). Constructing a model with two types of tradable sectors, as well as non-tradables, the paper makes the point that it is just the *relative* price of the outputs of the two types of tradable sectors that matters. Thus, an increase in the world price of one set of tradable sectors is technically identical to a reduction in the price of the other set. A positive trade shock to new tradable sectors has the same impact on urban structure as does a negative trade shock to old tradable sectors. In each case exit of the *relatively* badly affected type-T cities occurs, increasing the number of type-N cities with adverse effects on all of them.

5. Responses

We have argued four main points. The first is that trade and technology shocks can knock-out some tradable goods sectors and hence some cities dependent on these sectors. This most obviously arises through negative shocks – such as direct import competition – but can also arise as a consequence of positive shocks. Positive shocks to some city-sectors will make others less competitive through general equilibrium effects in labor markets and real exchange rate appreciation.

Second, if sectors of activity face localisation economies, then it is hard to start up new activities in new places. The spatial economy is replete with market failures that make adjustment to change difficult while, in contrast, much international economics has put naive faith in adjustment occurring because 'everywhere has a *comparative* advantage'. The latter statement is true, but irrelevant in an integrated economy in which there is relatively little divergence in wages between places. It becomes positively misleading if there are barriers to starting activities in new places, such as those created by localisation economies and coordination failure.

Third, the adjustment mechanism in response to shocks takes the form of cities that loose tradable activity switching to non-tradables, sectors where localisation economies and coordination failure are less important. The identification of sectors with localisation economies with tradable goods, and vice versa, is of course not exact; there are surely some tradable goods that do not cluster, and some non-tradables that do. But this assumption seems to hold good for many of the sectors that we see in booming cities.

Fourth, if the adjustment mechanism is such that a wide range of shocks leave the economy with fewer cities producing tradable goods and more producing non-tradables, then there will be increasing polarisation of the urban system. The relative price of non-tradables falls, and with it the wages of people in towns and cities producing such goods. Mobile factors will move in response to this loss, so the negative impact is transferred to factors of production that cannot move. In this simple model this is just land, but in reality includes individuals who are unable or unwilling to move. The social, economic, and political consequences of this are apparent.

The core of the economic problem is that the market mechanism does not create sufficient incentives to start new tradable activities (or more generally, new activities that can achieve high value productivity through agglomeration economies and increasing returns to scale) in places that have lost historic specialisms. What can start such activities?

One possibility is based on innovation. If the economy is creating new activities, not linked to existing agglomerations, then it is possible that they start up in relatively low cost 'type-N' places. This possibility is emphasised in Moretti (2013), and exemplified by Seattle. In the 1970s Seattle was a city with a declining port and manufacturing sector, unemployment twice the US national average, losing population and famous for the 1971 billboard saying 'will the last person leaving Seattle turn out the lights'. Microsoft arrived, losing little if any productivity in moving from its original base of Albuquerque. The cluster of software activity then grew up around Microsoft. Of course, the move of Microsoft from Albuquerque was due to the fortuitous circumstance that both Bill Gates and Paul Allen had grown up in Seattle. Thus, while the innovation route has been successful in transforming some cities, it

seems unlikely that there are enough distinct new innovative clusters for this to be a solution for more than a few fortunate cities.

A second route is to try to address the coordination failure by policy that targets particular places for economic development, possibly in specific sectors. Economic reasoning suggests that 'large developers' may be able to internalise the externalities created by agglomeration, overcoming coordination failure by launching development at scale. Public policy to support this may take the form of city plans and the location of infrastructure (e.g. placement of transport hubs). Special economic zones offer regulatory, fiscal, and infrastructure benefits, concentrated in one place with the hope of creating cluster benefits. Developing countries offer some successful example, such as Shenzhen, Dhaka, and Penang. However, there are many more failures.[12] Developed countries have used fiscal incentives in the form of regional investment or employment subsidies and subsidies to influence plant location decisions. Reviews of such policies suggest that, even if policies have had some impact, they have generally failed to jump-start new economic activities and trigger the development of self-sustaining private sector clusters (see e.g. Neumark and Simpson 2015, Kline and Moretti 2014). Perhaps one reason for this is that such policies succeed in attracting non-tradable activities, moving public sector jobs or securing investments in warehousing or customer service centres; policies may evn be targeted at these sectors. They fail to attract internationally competitive tradable sectors, so their effect is simply to accelerate the adjustment process described in this paper, dragging down incomes in all other non-tradable cities.

A third possibility is that parts of T-sector activities are able to split off from their core city and relocate to a lower wage type-N city. This brings the dual benefit of job creation in type-N cities and creating space for core activities to expand in type-T cities. The extent to which this is possible depends on the scope – functional and spatial – of agglomeration economies. The growth of offshoring indicates that it is possible to geographically separate back-office activities from parts of the business that benefit from presence in a cluster (e.g. in some finance and insurance sectors). However, two caveats are in order. First, the international context is one of much larger nominal wage differences than those that arise between cities within a country. And second, the move to 'reshoring' is illustrative of the fact that many firms found the costs of geographical fragmentation of activities to be greater than anticipated.

The final option is to accept increasing polarization of the urban structure and the attendant decline of many towns and cities. In a simple framework – such as that presented in this paper – this could be welfare improving, particularly if booming cities are enabled to expand

[12] See Duranton and Venables (2017) for analysis of place-based policies developing economies.

by constructing infrastructure and housing to mitigate effects on commuting costs, land prices and rents.[13] However, the further costs are evident. Many people are unable or unwilling to move, with economic, social and health costs that are apparent. Mitigating these costs is important, and we note that there is a significant resource that could be used to finance this. As we saw above, the beneficiaries of urban polarisation are those fortunate to receive windfall gains from owning land in the booming cities, gains which are administratively, economically, and ethically – if not politically – ripe for taxation.

References:

Autor, D.H., D. Dorn and G. H. Hanson (2016) 'The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade', *Annual Review of Economics*, 8:205-40.

Combes, P-P. and L. Gobillon (2015). The Empirics of Agglomeration Economies. In Gilles Duranton, Vernon Henderson, and William Strange (eds.) *Handbook of Regional and Urban Economics, volume 5A*, 247–348. Elsevier, Amsterdam.

Cramer, K. J. 2016. The politics of resentment. Rural consciousness in Wisconsin and the rise of Scott Walker. Chicago: University of Chicago Press

Diamond, R. (2016) 'The determinants and welfare implications of US workers' diverging location choices by skill, 1980-2000' *American Economic Review*, 106, 479-524.

Duranton, G. 2007. 'Urban Evolutions: The Fast, the Slow, and the Still' *American Economic Review*, 97(1): 197-221

Duranton, G. and D. Puga, (2004). 'Micro-foundations of urban agglomeration economies', in: J. V. Henderson & J. F. Thisse (eds.), *Handbook of Regional and Urban Economics, volume 4*, 2063-2117. Elsevier, Amsterdam.

Duranton, G. and D. Puga, (2015). 'Urban Land Use', in: G. Duranton, J. V. Henderson and W. Strange (eds.), *Handbook of Regional and Urban Economics*, volume 5, chapter 8, pages 467-560, Elsevier, Amsterdam.

Duranton, G. and A.J. Venables (2017). 'Place-based policies for development', World Bank, Washington DC.

Florida, R., C. Mellander and K.M. King (2017) 'Winner-Take-All-Cities', Martin Prosperity Institute, dp 2017-MPIWP-002.

[13] Hsieh and Moretti (2018) argue that restrictions on housing construction in booming urban areas has reduced aggregate US growth by 1/3rd over the period 1964-2009.

Glaeser, Edward L, & Gottlieb, Joshua D. (2009) 'The wealth of cities: Agglomeration economies and spatial equilibrium in the United States'. *Journal of Economic Literature*, 47.

Glaeser E.L. (2012) *The Triumph of the City*, Macmillan, London

Gyourko, J., Mayer, C., Sinai, T. (2013) "Superstar Cities," *American Economic Journal: Economic Policy*, *5*(4), 167–99.

Henderson, J. V. (1974). "The Sizes and Types of Cities." *American Economic Review* 64(4), 640–656.

Henderson, J.V. and A.J. Venables (2009), 'The dynamics of city formation', *Review of Economic Dynamics*, 12, 233-254.

Hsieh, C-T. and E. Moretti (2018) 'Housing Constraints and Spatial Misallocation', *American Economic Journal: Macroeconomics*, forthcoming.

Kline, O. and E. Moretti, (2014) People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Policies *Annual Review of Economics*, 6 629-662.

Moretti, E. (2013) *The New Geography of Jobs*, Mariner Books, New York.

Neumark, D. and H. Simpson. (2015). 'Place-based policies' In: G. Duranton, J. V. Henderson, and W. C. Strange, *Handbook of Regional and Urban Economics, volume 5B.* Amsterdam: Elsevier, 1197-1287.

Rodriguez-Pose, A. (2017) 'The revenge of the places that don't matter (and what to do about it)', CEPR dp no 12473.

Rosenthal S. & W. Strange. (2004). 'Evidence on the nature and sources of agglomeration economies' in V. Henderson & J-F. Thisse (eds) *Handbook of Regional and Urban Economics*,vol 4, pp 2119-2171. Elsevier, Amsterdam.

Storper, M. (2018) 'Separate Worlds? Explaining the current wave of regional economic polarization', *Journal of Economic Geography*, 18, 247-270

Venables, A. J. (2018) 'Globalisation and Urban Polarisation', *Review of International Economics*, forthcoming.

Appendix

A small open economy is endowed with a single factor of production, labour in quantity L, which is mobile between cities. The number of cities is fixed at M, each has a geographical structure of the usual Alonso-Mills-Muth form, in which jobs are located in the central business district (CBD), workers occupy residential land, commuting is costly, and land rent adjusts to make workers indifferent between residential locations. There are two types of production sector, tradables and non-tradables. Non-tradables are produced under constant returns to scale and are freely traded within the country but not internationally; their price is set on the national market. Tradable sectors produce goods or services that are freely tradable at fixed world prices and are subject to city specific agglomeration economies (localisation economies). Each city specialises, being either type-T (specialising in a single tradable sector), or type-N (producing only non-tradables). The number of type-T cities is endogenous and denoted M_T, and the remainder are type-N, producing non-tradables ($M_N = M - M_T$).

The price of tradables is p_T, exogenously fixed at the world level, and the number of workers employed in each type-T city is denoted L_T. Productivity, q, in each type-T city is increasing in the level of employment in the city, $q'(L_T) \geq 0$. Labour is the only input to production, so

$$w_T = p_T q(L_T). \tag{1}$$

Non-tradables have price p_N and offer wage w_N. One unit of labour produces one unit of non–tradable output so $p_N = w_N$. Employment in each type-N city is L_N. The price (and hence wage w_N) of non-tradables is determined by market clearing condition

$$(M - M_T)L_N w_N = \theta\left[M_T w_T L_T + (M - M_T)w_N L_N\right]. \tag{2}$$

The left hand side is the value of supply of non-tradables, and the right hand side is the value of demand, where the term in square brackets is total income and $\theta < 1$ is the share of income spent on non-tradables. All income in each city is spent on a composite good which is a Cobb-Douglas aggregate of tradables and non-tradables with price index $P = p_T^{1-\theta} w_N^{\theta}$.

Taking tradable goods as the numeraire, $p_T = 1$ and the price index of the composite good is

$$P = w_N^{\theta}. \tag{3}$$

Workers' per capita utility in a city of type-i is $u_i = \left(w_i - bPL_i\right)/P = w_i/P - bL_i$. The term in brackets is the wage net of urban costs, these consisting of commuting and rent payments. Urban costs increase with city population L_i at rate b and are incurred in units of the composite good with price index P. Expenditure net of urban costs goes on the composite good, so utility is spending net of urban costs deflated by the price index. Labour mobility equalises utility across all cities so

$$w_T / P - bL_T = w_N / P - bL_N \,. \tag{4}$$

Finally, national labour market clearing is

$$L = M_T L_T + (M - M_T) L_N \,. \tag{5}$$

Equilibrium conditional on the number of type-T cities, M_T, is the solution of the five equations above for endogenous variables $\{w_T, w_N, L_T, L_N, P\}$ and gives the wage curves of Figure 1a.

The 'urban costs' borne by each worker are divided between commuting costs and land rent according to residential location.[14] For a worker at the city edge they are entirely commuting costs while adjacent to the CBD they are entirely rent. If the city is linear and commuting costs are linear in distance then, since the commuting cost paid by the marginal worker (living at the city edge) is bPL_i, (i = T, N), the total of commuting costs and rent city wide is bPL_i^2. Total commuting costs are half this, $bPL_i^2/2$, the remainder being rent, $bPL_i^2/2$. Rents are spent on the composite good so real rents (deflated by P) are $R_i = bL_i^2/2$. Notice that total real rents are increasing and convex in city size. The quadratic form comes from our assumptions of a linear city with linear commuting costs, but convexity is a much more general property as large cities have both more land and higher average rent. The total utility generated by a city of type i is the real income of workers plus land rents, $u_i L_i + R_i$.

Figure 1a traces out values of variables solving eqns. (1) – (5) as a function of M_T, with the horizontal axis giving the proportion of cities that are type-T and the vertical giving wages. Figure 1b gives the corresponding total real rents in a single city of each type. The figure is constructed with L = 100, M = 100, θ = 1/2, b = 0.3, $q(L_T) = 1 + 0.15 L_T$.

[14] For detailed exposition of the Alonso-Muth-Mills model of urban land-use see Duranton and Puga (2015).