UNITED NATIONS UNIVERSITY
UNU-WIDER
World Institute for Development Economics Research

Working Paper No. 2010/26

Evolving City Systems

Henry G. Overman[1] and
Anthony J. Venables[2]

March 2010

Abstract

This paper reviews the literature on the forces driving urbanization in developing countries. It presents a model outlining how globalization can lead to the evolution of an urban structure which may approximate Zipf's law. Policy implications are outlined.

Keywords: urbanization, globalization, agglomeration, city systems

JEL classification: O10, O18, R10, R12

1 London School of Economics, email: H.G.Overman@lse.ac.uk, 2 University of Oxford, email: tony.venables@economics.ox.ac.uk

This study has been prepared within the UNU-WIDER project on Beyond the Tipping Point: Development in an Urban World directed by Jo Beall, Basudeb Guha-Khasnobis, and Ravi Kanbur.

UNU-WIDER gratefully acknowledges the financial contributions to the research programme by the governments of Denmark (Royal Ministry of Foreign Affairs), Finland (Ministry for Foreign Affairs), Sweden (Swedish International Development Cooperation Agency—Sida) and the United Kingdom (Department for International Development).

ISSN 1798-7237 ISBN 978-92-9230-261-0

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

www.wider.unu.edu *publications@wider.unu.edu*

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Anne Ruohonen at UNU-WIDER

1 Introduction

The urban population of the developing world is projected to increase by some two billion in the next 30 years. Urbanization rates are strongly correlated with per capita income, productivity tends to be high in cities, and urban job creation is an important driver of economic growth. But urbanization is also one aspect of the widening spatial disparities that often accompany economic development, and many countries have urban structures dominated by their prime city. While cities are highly productive, they create heavy demands for investments in infrastructure and accommodation, in the absence of which slums and informal settlements develop. Urbanization gives rise to numerous policy challenges, both to make cities work better and to ensure that the overall city structure (the number and size distribution of cities) is as efficient as possible. There is no presumption that an unregulated free market pattern of urban development is socially efficient (even when conditional upon appropriate levels of public investment). Urban activity creates many externalities, both positive and negative, so economic theory tells us that an unregulated outcome will not achieve efficiency. We observe the grim conditions of developing mega-cities, and we know that, in some developing countries, the primate city takes a far larger share of population than was the case in much of the developed world at similar stages of development (Bairoch 1988). The performance of the urban sector also bears on overall economic growth. Much job creation – in modern sector activities and in the informal sector – takes place in cities. What determines the attractiveness of a location as a host for investment, and how can city environments be developed to maximize job creation? Do 'bad' city structures impede overall growth?

These concerns point to the need to have a robust understanding of the economics of cities, both theoretically and empirically. Unfortunately, we are far from having such an understanding. Work in development economics over the last several decades has been largely silent on the issue. Some aspects of urbanization have been approached in the contexts of migration and industrial development, but have rarely focused on the particularly urban issues that arise. Urban economics is, however, experiencing a renaissance in the academic literature.[1] There has been improvement in theoretical methods as economic analysis becomes better at analysing economies of scale, and there have been empirical advances as larger datasets and better econometric tools are applied. Some of this work has been applied to developing countries (most notably by Vernon Henderson from Brown University), although the developing country literature remains thin.

The objective of this paper is to draw out messages from this literature in order to understand several issues. The first issue is that of the benefits and the costs of cities. Per capita income and productivity are generally higher in cities than elsewhere, partly because of productivity benefits of cities, and partly because of political access and rent seeking. At the same time, urban living is constrained by the availability of land, and brings congestion and pollution costs. How do these costs and benefits depend on urban scale, and how does urban governance shape them? This question is the starting point

1 See, for example, the new *Handbook of Regional and Urban Economics* (Henderson and Thisse 2004). However, not one of the 20 chapters in this volume deals with developing country issues.

for any discussion of cities in development, and this paper offers a brief survey of the literature.

The second issue we address is that of the determinants of urban structure. What determines the size distribution of cities, how do these evolve through time, and what are the arguments for policy intervention? One aspect of this is to do with the effect of openness to trade, both international and within country. There is a view that globalization is encouraging the move to large cities, and we develop a theoretical model that explores and confirms this conjecture. The other concerns the extent to which the market failures associated with urban development bias the shape of the urban system. Cities are riddled with market and governance failures that cause inefficiencies within cities, distort the city structure, and possibly also reduce the returns to job creation, impeding overall growth. There is likely to be a tendency for cities to be too large and for there to be too few of them, and an argument that policy should remove obstacles to decentralization. We review and discuss these arguments.

We organize the material by first looking at the literature on costs and benefits of cities. We then turn to issues of trade and urban structure, and subsequently cover market failures and urban policy in the final section.

2 Urban economics: theory

Two broad theses are offered to explain urbanization in developing countries and, in particular, the role of the primate city. The first, which we will refer to as 'productivity', is that there are efficiency gains associated with clustering activity; firms and workers are attracted by these benefits. The second, which we will refer to as 'rent seeking', is that city dwellers are better able to extract rents because of preferential access to the political system. These hypotheses are not mutually exclusive; both operate to varying degrees in different countries and cities. Pulling in the opposite direction, there are costs associated with urban centres – including transport and infrastructure costs, and externalities associated with congestion and pollution.

2.1 Productivity

There are some functions, such as government and central administration, that are inherently urban – or at least appear so in virtually all historical contexts. Other activities, such as distribution (markets, exchange, wholesale, and retail activities), other services and manufacturing can operate either in a city or in smaller towns or villages. What determines the benefits of grouping such activities in an urban centre?[2]

The first argument derives from 'thick' goods markets. A system of exchange works better if it operates at a reasonable scale. The variety of goods on offer is greater, search and travel costs are reduced, and competition is more intense. Unsurprisingly then, 'market towns' develop in order to provide locations for trade. Often, these centres are

2 Many of the arguments presented here are developed in greater detail in Duranton and Puga (2004).

home to local agricultural markets and so have important linkages with the surrounding rural economy.[3]

The second urban advantage derives from 'market access'. Suppose that a manufacturing (or service) activity faces transport costs on the goods that it sells. Other things being equal, the most profitable location for such a firm is close to a large mass of consumers – that is, in an urban centre. Models predict an amplification effect, so that manufacturing is drawn towards locations with good market access and, in turn, the jobs and consumer expenditure created mean a large market. Physical geography and transport systems also come into play here; a port or river crossing will have better market access than a mountaintop. These arguments extend to forward and backward linkages, perhaps better labelled as 'cost' and 'demand' linkages.[4] Many firms are primarily engaged in supplying other firms (rather than final consumers) and, for such firms, good market access means proximity to customer firms. This is the backward or demand linkage – firms want to locate close to the sources of demand for their output. But the converse of any backward linkage is a forward or cost linkage. Firms that purchase the output of other firms will want to locate close to their supplier firms. The combined effect of these demand and cost linkages can create a powerful force for agglomeration of activity. We see it in dense networks of firms in related industries – engineering, electronics, and even financial services – where firms that supply specialist financial skills locate near the big financial institutions, and these institutions benefit from access to the skills of the specialists. A developing country example is the surgical instruments cluster in Sialkot, Pakistan, where Nadvi (1999) identifies 'over 20 stages in production, each requiring distinct skills and tools. Surgical instrument making thus lends itself to an extensive division of labour and the process of large numbers of subcontractors in most process activities' (Nadvi 1999: 87).

Several other arguments also point to productivity benefits of cities. Thick market effects arise in the context of labour markets. Large pools of specialist workers and firms using these skills benefit from the better matching of skills with requirements, and also from risk sharing if there are firm- or worker-specific fluctuations in demand or supply. Incentives to acquire skills are greater if several firms are seeking those skills; the worker is then less likely to be subject to the power of a single employer (monopsony). Labour turnover is one – but not the only – mechanism through which firms in a dense cluster of activity can benefit from the skills and knowledge of other firms. There is considerable evidence of productivity spillovers between firms, as they are able to learn about and imitate the practices of other firms in the industry. Silicon Valley provides an example where knowledge exchange – formal and informal – is widespread. The knowledge might be about production methods, marketing skills, or simply knowledge about the location itself. Thus, multinational firms tend to cluster in particular locations, partly because one firm, observing the success (or failure) of another, learns about the quality of the business environment in the location. Hausmann and Rodrik (2002) argue that very narrow patterns of specialization in developing

3 The particular policy issues relating to small and intermediate urban centres, and their potential role in regional and urban development, are beyond the scope of this paper: they are discussed in depth by Tacoli (2004).

4 For a formal analysis, see Fujita *et al*. (1999). An older tradition of development economics considered these linkages, although without recognizing the role of increasing returns to scale and associated market failures.

countries (for example, specialization in soccer ball production) arise as producers learn about the efficiency of a particular location for producing a particular good, this then becoming public knowledge.

The final argument is to do with the provision of public goods. Public goods are inherently associated with increasing returns to scale, since the opportunity cost of consuming them is zero. It therefore makes sense to locate them where demand is high and where any network effects can be maximized. Thus, if a country is to build 100 kilometres of paved road, it is efficient to build most of it joined up rather than scattered around. In this case, an 'urban bias' in public expenditure and provision is an efficient allocation of resources (Arnott and Gersovitz 1986).

These are distinct arguments, but all share several characteristics. First, they are sources of spatially concentrated increasing returns to scale; there are efficiency gains from having things located in the same place. Second, all these arguments are potential drivers of cumulative causation. People choose to set up activities in a location not because of the intrinsic merit of the location, but because other people have already done so, or are expected to. Put differently, there are positive reciprocal externalities; my presence makes the city more attractive to you, your presence makes it more attractive to me, and so on. Furthermore, these mechanisms create 'lock-in' or 'path-dependent' development. A city might be in the wrong place, or an industry in the wrong city – but once there, it will not be profitable for any single producer to move away from the cluster.

Before leaving the issue of the productivity benefits of cities, we should note two further issues. One is that the productivity benefits might have quite different sectoral ranges. Some operate within quite narrow sectors of activity, so do not require large cities to accommodate them, while others (e.g. to do with the operation of the labour market) cut across sectors and are city wide. We return to this distinction between 'localization' and 'urbanization' in the empirical section. The other is that cities can have dynamic effects over and above the static effects on which we have focused here. These dynamic effects will depend on the role that urban environments play in developing new products and processes. Duranton and Puga (2001) argue that large diverse metropolitan areas play a role as a 'nursery'. These information-rich environments allow firms to develop new products and processes. However, once production becomes standardized, firms move out to cheaper specialized locations where they can benefit from localization economies without the high costs of the large urban city. Duranton and Puga (2001) provide evidence on firms across French metropolitan areas that appear to be consistent with their theory. Could large diverse metropolitan areas be playing a similar role in developing countries? Clearly, the kind of R&D and innovation undertaken by developing country entrepreneurs differs from the way these terms are commonly used in the developed country context. Nevertheless, entrepreneurs in low-income countries must also engage in a process of innovation and learning. Their focus is on what Rodrik (2004: 9) calls 'cost discovery': 'What is involved is not coming up with new products or processes, but discovering that a certain good, already well established in world markets, can be produced at home at low cost'. Rodrik suggests some developing country examples: cut flowers in Colombia, t-shirts in Bangladesh, soccer balls in Pakistan, and software in India. The urban nature of these cost discovery processes remains largely unexplored. However, Hausmann and Rodrik's (2002) emphasis on the importance of tacit knowledge in the self-discovery process suggests that, just as for their developed country counterparts, this process of cost

discovery will be easier in the information rich environment of large diverse urban areas.

2.2 Rent seeking

The arguments discussed turned on real efficiency gains from the scale effects of urban centres. Other arguments are based on the idea that urban–rural differentials are due to transfer payments – urban dwellers benefit not by creating resources, but by extracting them from the rest of society. These arguments have been developed by a number of researchers, including Lipton's view of 'urban bias' (Lipton 1976, 1993) and Hoselitz's 'parasitic city' (Nash 1977).

The main mechanism is political access. In many developing countries, starting a business, hiring and firing workers, registering property, enforcing contracts, getting credit, protecting investors, and closing a business are subject to extensive regulation (World Bank 2005). The probability of getting permits and licences could be enhanced by proximity to the administrative centre. Furthermore, the political power of an urban proletariat might mean that the government acts to raise their real incomes. This will attract workers to the town, although its effect on employment is ambiguous; food subsidies might reduce the wages that firms need to pay, creating jobs, while minimum wage legislation will have the opposite effect.

The urban bias argument is also made in terms of overall patterns of import protection and relative prices. Import substituting governments have typically raised the prices of manufactures relative to agricultural goods, and this is sometimes argued to be a source of urban bias. For present purposes, however, it is important to keep separate the question of what is produced and where it is produced. Supporting manufactures is supporting cities only if – for some other reason – manufactures are produced in cities.

2.3 Urban costs

The forces outlined in the previous section were to do with cities creating income (productivity effects) or transferring it (rent seeking). However, cities also destroy income – they create costs for urban dwellers.

A concentrated urban structure, particularly one that has a high degree of city specialization, involves large volumes of trade between cities and, hence, incurs high transport costs. This cost is internalized by firms but is nevertheless a cost of urban concentration to be placed against the benefits of economies scale. Within the city, large urban areas incur high costs of travel and commuting. A standard urban economics model assumes that jobs are clustered in one (or several) 'central business districts' to which workers have to commute. This, in turn, generates a land rent gradient; rents are high in the centre where commuting costs are low, and low on the edge where commuting costs are high. Commuting costs are the resource and time costs of running and using an urban transport system, and can be amplified by congestion costs.

Conceptually, the costs outlined in the preceding paragraph are of three quite distinct types. The first is the direct cost of trade between cities or uncongested commuting within a city; a resource cost, but one that is not necessarily associated with any sort of market failure. The second is the cost of paying urban rents. This is a transfer payment,

not a resource cost; it is paid by urban dwellers to urban landowners. The third is the addition to commuting costs created by congestion, a negative reciprocal externality between those travelling within the city. Of course, this is not the only negative externality associated with urban areas. Others include air and water pollution and, in some places, costs of crime and urban violence.

A further source of cost that has been the subject of much attention in the development literature is that a city might attract a number of workers who are either un- or under-employed. The economics underlying this dual labour market structure are illuminated by the Harris–Todaro model. The model supposes that the urban real wage is above real earnings in agriculture. This might be because of institutional rigidities supporting a high urban wage (minimum wage legislation, union power or price support), or for efficiency wage reasons (wage reductions are unprofitable, as they reduce the quality of labour through nutritional, effort, or selection effects). The high urban wage attracts labour to the city in search of these 'formal sector' jobs. Given the number of such jobs, equilibrium migration is attained when the probability of a migrant getting a job is low enough that the expected wage from migration equals earnings in agriculture. Migrants who fail to obtain a formal sector job are unemployed, or work for a much lower wage in the urban 'informal sector'.[5] The strength of the Harris–Todaro model is its simplicity, but this comes at the cost of abstracting from many important aspects of the problem. The dual structure of formal and informal urban labour markets is complex, and rural–urban migration occurs for many reasons. Nevertheless, the model makes the point that a possible cost of urbanization is the associated development of a mass of low productivity urban informal sector labour.

3 Urban economics: empirics

The previous section outlined the benefits and costs of urbanization. How large are these in practise, what is the net balance between them, and how does this vary with city size?

3.1 Productivity

The main areas of enquiry revolve around the scope, sources, and magnitudes of productivity effects (Rosenthal and Strange 2004). For geographical scope, data requirements mean that analysis usually takes as given some broadly defined metropolitan statistical area. Studies that have focused on geographical scope generally find productivity effects operating over quite short geographical distances. For example, Patacchini *et al*. (2006) find that, in the UK, productivity effects operate largely within a 45-minute driving time.

With regard to industrial scope, empirical work has usually resorted to a dichotomy between two types of externalities: localization and urbanization. Localization economies exist if firms benefit from the presence of firms in the same industry. Urbanization economies exist if the benefit arises from simply being in large urban

5 See Becker and Morrison (1999), for discussion of rural urban migration.

areas. Some authors, but not all, define urbanization economies as arising from large *diverse* urban environments.

Several stylized facts emerge from the developed country literature. There is consistent evidence that total factor productivity increases with city size. The survey by Rosenthal and Strange (2004: 2133) suggests that 'doubling city size seems to increase productivity by an amount that ranges from 3–8 per cent'. When studies attempt to distinguish between urbanization and localization economies, the strength of these economies can vary substantially across industries. For example, higher technological and service activities appear to benefit from urbanization economies, while more standardized production appears to benefit from localization economies. Within the standardized production activities, some industries show very strong localization externalities, while the effects for other industries are much weaker. Generally, the most robust findings concern the existence of localization economies.

Most of the econometric evidence that is available refers to developed countries but, as we have stated, there is nascent literature on productivity effects in developing countries. Table 61 presents an overview of the available papers and highlights their findings. The econometric literature summarized in Table 61 is still in its infancy. Our reading is that, with the exception of two of the studies on India, the findings for developing countries are broadly in line with those from developed countries. Unfortunately, in the particular case of India, the finding of localization diseconomies in some sectors sits uneasily with that of significant spatial concentrations of particular industries reported in Lall *et al.* (2004). Hopefully, further work on the detailed Indian data used in Lall *et al*. (2003, 2004) will throw further light on this matter. Results for Brazil, China, Indonesia, and Korea suggest that there is some evidence of urbanization economies, but not in all industries. Interestingly, little of the developed or developing country literature considers how these externalities change with city size. In fact, nearly all specifications are log-linear, which implies constant elasticity of productivity with respect to own industry concentration or diversity, so the externalities created per worker are basically independent of city size. We also lack good studies in sectors outside manufacturing, in particular in services.

These findings on urbanization and localization appear to be reflected in the pattern of economic activity across cities in developing countries. Activities that are subject to diversification economies tend to be found in the largest cities. Those that are subject to large localization economies are discovered in a few medium-sized cities, while those subject to smaller localization economies are less concentrated across a number of small size cities. Thus, Lall *et al*. (2004) find evidence of high spatial concentration for the leather and metals sectors, and moderate concentration in food products, textiles, mechanical machinery and computing and electronics, while firms in the paper products and chemicals sectors do not exhibit patterns of spatial concentration.

Table 1: Developing country evidence on productivity effects

Country	Paper	Findings
Brazil	Henderson (1988)	Localization
China	Chen (1996)	Localization economies for 2 out of 2 industries (machinery and food); does not consider urbanization economies
India	Shukla (1996)	Localization and urbanization economies; urbanization stronger than localization in 11 out of 13 industries
	Mitra (2000)	Urbanization economies in 11 out of 17 industries; does not consider localization economies.
	Lall *et al*. (2003)	Urbanization: Food processing, textiles, leather, paper, chemical, basic metals, mechanical machinery, electrical
		Localization diseconomies
	Lall *et al*. (2004)	No localization or urbanization
Indonesia	Henderson and Kuncoro (1996)	Localization: Apparel (inc. textiles), non-metallic minerals, machinery (including transport and electrical)
		Urbanization: Wood, furniture, publishing
Korea	Lee and Zang (1998)	Localization not urbanization economies (19 industries)
	Henderson *et al*. (2001)	Localization: Traditional, heavy, transport, machinery
		Urbanization: High tech

Evidence on localization is also provided by numerous case studies of spatial clusters of firms.[6] This literature makes it clear that the clustering of firms to benefit from some sort of agglomeration externalities is a widespread occurrence in developing countries. There are often strong market-based input–output linkages between these firms, and some evidence that the non-market exchange of goods, information, and people also occurs. Many of these case studies also stress the importance of common cultural and social background in generating particular norms of behaviour, and of local private and

6 Examples include the Sinos Valley, Brazil (shoes); the Gamarra region of Lima, Peru (clothing); Guadalajara and Leon, Mexico (shoes); Eastlands, Kenya (garments); Kamukunji, Kenya (metal products); Ziwani, Kenya (vehicle repair); Lake Victoria (processed fish); Suame, Ghana (vehicle repair and metal work); Western Cape, South Africa (clothing); Tiruppur and Ludhiana, India (knitwear); Agra, India (knitwear) and Sialkot, Pakistan (surgical instruments).

public institutions that might reinforce those norms. However, identifying the true benefits to firms located in these clusters will involve much more analysis, focusing in particular on the identification of a set of firms outside the cluster that provide a suitable set of controls for comparison (see Humphrey 1995; Schmitz and Nadvi 1999; and Visser 1999).

3.2 Productivity effects in the informal sector

The evidence on agglomeration effects reported in Table 61 comes exclusively from data concerning the 'formal' sector. How should the existence of a large pool of unemployed/underemployed workers (in accordance with Harris–Todaro) change our thinking on agglomeration economies? It might be expected to reduce the benefits of city scale, yet there is no evidence that this is the case (see, for example, Au and Henderson's [2006] results for China). There are two possible explanations. One such explanation is that the existence of an informal sector drives up urban costs and crowds out the formal sector, but not sufficiently quickly to offset the positive productivity effects of increasing city size. The second possibility is that the informal sector also contributes to agglomeration economies. There is evidence pointing to the existence of networks of small firms that benefit considerably from the productivity effects of the concentration of employment. In developing countries, authors such as Mukherjee (1990) emphasize the vitality of the informal sector. The informal sector often plays an important and visible role in the case study literature on clusters in developing countries. For example, Chari (2004), in his work on the knitwear cluster in Tiruppur, paints a vivid picture of the journey taken by cotton thread through the various milling operations, dyeing firms, and fabrication units. Formal and informal sectors play their part along the way as independently owned bullock carts shuttle yarn and knitted fabric between knitwear companies and fabrication units. To assume that no agglomeration externalities exist for Tiruppur's informal sector, and for informal firms more generally, is surely inappropriate. Unfortunately, our reading of the econometric studies is that more formal evidence on this issue is simply unavailable. Clearly, this is an important area for future work.

3.3 Costs

It seems uncontroversial to assert that costs rise with city size, although there is surprisingly little systematic developing country evidence on the subject. Evidence from Latin America (see Thomas 1980; Henderson 1988) finds that costs of urban living increase with city size. Moving from a small urban area to a large urban area at least doubles the cost of living. Richardson (1987) finds that the per family marginal investment cost is three times larger in urban than in rural areas for four developing countries, Bangladesh, Egypt, Indonesia, and Pakistan.

What about the non-economic costs of living in cities? Environmental problems are often worse in cities than in rural areas, but it is not clear how these externalities change with city size. Glaeser (1998) presents evidence for developed countries that suggests that levels of sulphur dioxide and ozone are not related to city size but, rather, that particulate concentration increases with city size. Shukla and Parikh (1992) suggest that sulphur dioxide has a slight tendency to increase with city size for developing countries. They do not consider the impact on ozone levels. Their results for particulates in less-

developed countries suggest an inverted-U, first rising and then falling with city size. It is very likely that this reflects the absence of industry or income controls (larger cities are richer and undertake less manufacturing, concentrating more on services). In addition to environmental costs, there are also likely to be a variety of social and health costs (and some benefits) associated with urban dwelling; more work is needed to integrate these effects in measures of real income in cities.

3.4 Real incomes

Real incomes give the difference between productivity effects and urban costs, and can be observed directly. Au and Henderson (2006) use Chinese data to study the relationship between real wages and city size. Their results provide several insights that have important policy implications. First, there is an inverted-U between real income and city size.7 The optimal point on this inverted-U depends on what kind of activity is located in the city. Second, the exact level of optimal city size is very imprecisely determined. This suggests that policy based on any notion of optimal city size will probably face insurmountable difficulties in deciding what that optimal city size actually is. Third – at least, from an economic viewpoint – it is much more costly to be under-sized than over-sized. This point is so important that it is worth quoting at length: 'For [a city with manufacturing to service ratio of 1] from a …peak size of 1.27 million if one subtracts 1.22 million people so city size is 50,000, real output per worker falls by 83 per cent; while, if one adds 1.22 million so size is 2.49 million, it only falls by 26 per cent. Real output per worker has a fairly flat portion near the peak, and real output per worker initially drops slowly past the peak' (Au and Henderson 2006: 568). Finally, the very flat peak has implications for city sizes when workers are allowed to be mobile. In particular, cities with slightly better amenities or market access could end up becoming very large compared with cities with slightly worse amenities or market position.

3.5 Conclusions

While more research work is needed, it seems clear from the evidence we have that there are substantial productivity advantages to urban centres and, as such, the development of these centres is a key part of countries' economic growth. The remainder of the paper considers this relationship between city size and productivity, and explores some of its implications.

4 Trade and urban structure

In this section we explore the determinants of urban systems in a simple analytical model. Central to the model is the presence of localization economies creating an incentive for specialized cities to form. Pulling in the opposite direction are the costs of shipping goods, creating an incentive for production to be dispersed rather than concentrated in specialized cities. We use the model to explore the trade-off between

---

7 One reason for using Chinese data is the presence of restrictions on labour mobility. Free labour mobility means that one should not observe cities on the left of the peak.

these forces. In particular, we investigate the impact of falling trade costs – globalization – on urban structures. What impact do we expect falling trade costs to have on urban structure, and what world and national urban systems are likely to evolve?

To capture the trade-off between the productivity benefits of urbanization and the cost of shipping goods between cities, we make the following assumptions. First, localization economies mean that each city is specialized in a particular sector (or a clearly defined and linked set of sectors), an approach developed by Henderson (1974). Second, countries are small enough that each country has, at most, one city operating in a particular sector, (the reader might wish to interpret 'countries' as regions or provinces in a large country). And third, international trade means that a particular city might supply its goods to several countries, and the extent to which this occurs is determined from the trade-off between the strength of the localization economies and the costs of shipping goods between countries. Details of the model are given in the Appendix and, here, we go straight to describing results.

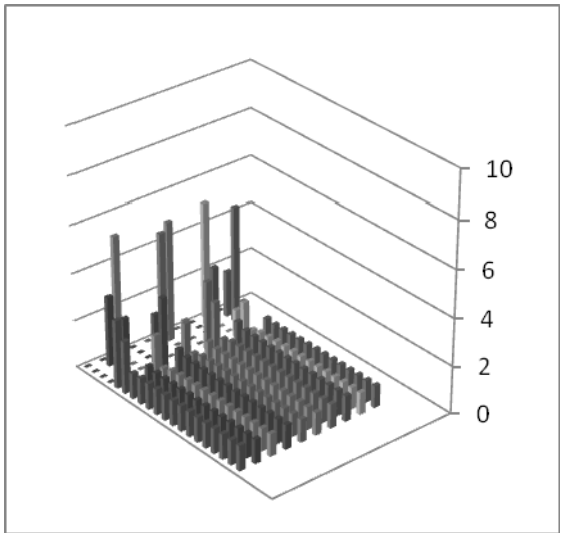
In Figure 1, the front right-hand axis represents countries. There are ten of them, and in the model they are arranged around a circle, with transport costs proportional to the (shortest) distance between them around the circle. The left-hand axis represents sectors, 20 of them ranked by the degree of returns to scale, those with highest returns at the back, and lowest returns at the front, the front three having diminishing returns to scale. The vertical axis is the scale of operation of each sector in each city.

The first panel (top left, $t = 3.2$) gives the outcome when trade costs are very high (t is the trade cost factor for a journey between the two most distant cities). Most city/sectors are present in all countries, as illustrated by the large mass of columns of equal height, indicating that each country has an equal size city specializing in each of those sectors. However, even at these high trade costs a few sectors have sufficiently strong increasing returns to concentrate and supply the world from a few locations. To be more exact, there are six sectors in which large 'regional cities' have formed, supplying the world market from either two, three, or four locations. These are the bars at the back of the figure. They are tall because the fewer cities there are operating in a sector, the larger those cities will be.

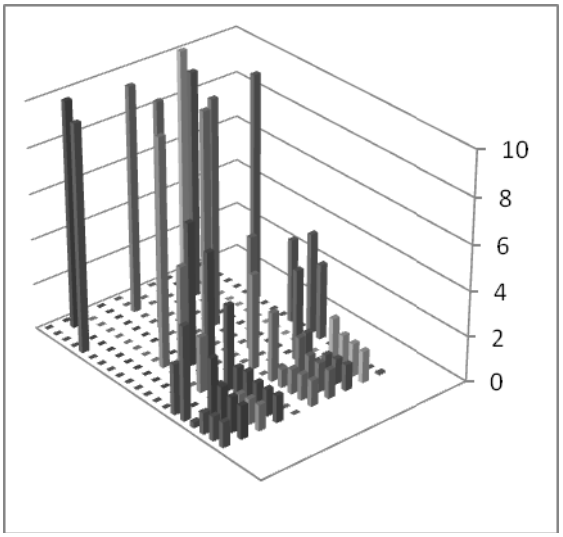
The second panel in the left-hand column gives the situation once trade costs have fallen to 2.2. Then, we find the following structure: one sector has concentrated all its production into a single world city (with output value = 10). The five sectors with the next highest increasing returns are each operating from two cities, each with output value ≈ 5. There are then four sectors operating from three cities, with output value ≈ 3, and three sectors, each of which is concentrated in four cities, with output value ≈ 2.5. Other sectors operate in all countries.

Further reductions in trade costs are illustrated in the remaining three panels. As trade costs fall, we see that that more sectors come to supply the world from a few global hubs, and average city size increases. In the last figure, with perfectly free trade, the only sectors that do not operate from a single world city are those in which there are diminishing, rather than increasing, returns to scale.

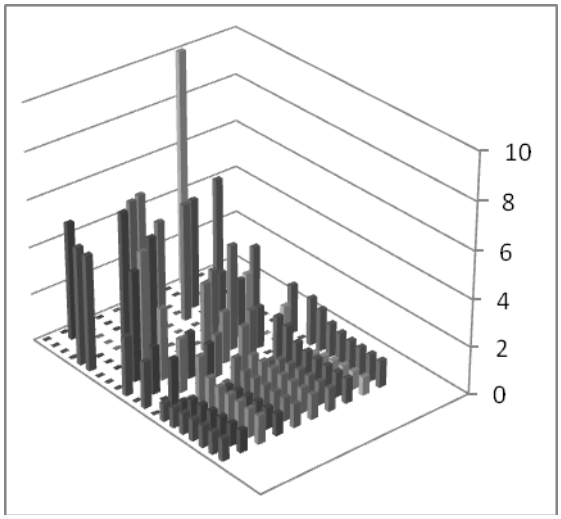
Figure 1: City sizes as trade costs fall: city/sectors and countries



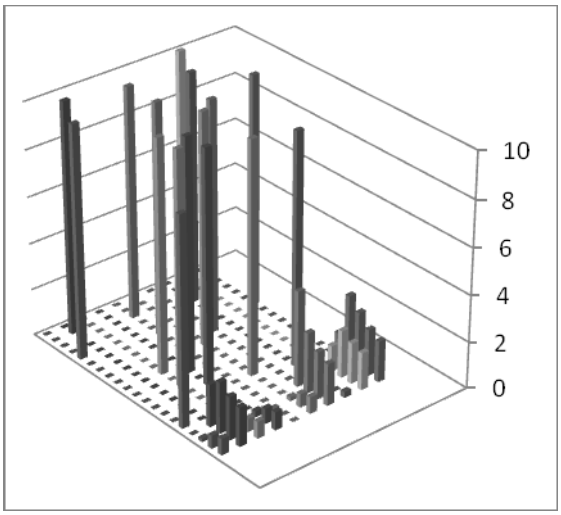
Trade costs = 3.2



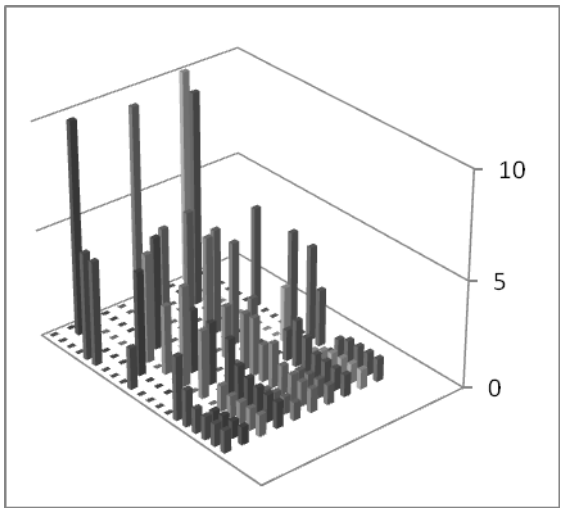
Trade costs = 1.4



Trade costs = 2.2



Trade costs = 1.0



Trade costs = 1.8

Several points come out of this analysis. The first is that declining trade costs and globalization are forces that will promote the development of large cities. Cities specialized in activities with the most increasing returns to scale are the first to become regional centres (supplying two or three countries), and then to become world cities. As this happens, so cities become larger and – given our assumption that total urban population is fixed – there are fewer cities. The second point is the relationship of this approach to the extensive literature on the size distribution of cities and, in particular, to Zipf's law (see Fujita *et al*. 1999). In the example presented here, the size distribution of cities depends on openness to trade, and there is a compelling power series logic underlying the structure. The largest cities supply the entire world market in the sector; the next largest supply half of it; the next largest one third, then one quarter, and so on. This does not map directly into the size distribution of cities in any country, as a country will not have a presence in all these sectors, but Table 2 reports the estimated coefficient on a power law relationship between city size and within-country city rank. We see that the city size distribution goes from being much flatter than Zipf's law (coefficient of -1) when trade costs are high, to being steeper when trade costs are low. Of course, these numbers are specific to the parameters of the example developed here (considering that all sectors take the same share of expenditure) but, as noted, there is a fundamental power series logic about the way that trade allows the world to be supplied from integer numbers of cities.

Table 2: Estimated elasticity of city size with respect to city rank

t_{max} =3.2	t_{max} =3.0	t_{max} =2.8	t_{max} =2.6	t_{max} =2.4	t_{max} =2.2	t_{max} =2.0	t_{max} =1.8	t_{max} =1.6	t_{max} =1.4	t_{max} =1.2	t_{max} =1.0
-0.35	-0.46	-0.51	-0.56	-0.65	-0.77	-0.79	-0.89	-0.99	-1.27	-1.5	-1.6

5 Dynamics: threshold effects and growing an urban structure

The dynamic in the preceding section comes from falling trade costs. Many developing countries also face their own urban dynamic arising from natural population increase and internal migration. How is this additional urban population distributed across cities, and what does economic analysis have to say about the equilibrium and socially optimal pattern of urban development?

The productivity effects that we described earlier are largely externalities, either pecuniary or technological. Such external economies of scale make it hard to start new cities. Small cities do not benefit from urban scale economies: they are therefore unattractive to firms and to migrants and, as a consequence, fail to grow. Instead, migration flows into existing cities, which might grow to become mega-cities, growing well beyond their optimum scale and, possibly, to the point where, at the margin, diseconomies such as congestion outweigh the positive economies of scale. Such an outcome is clearly inefficient, but raises difficult policy issues. There are likely to be

two quite distinct types of market failures here. One is that increasing returns to scale give rise to externalities, so that the benefits *created* by a single economic agent (a migrant to the city or a relocating firm) are not internalized. The other is that the benefits *received*[8] by a single economic agent accrue over time and might be highly uncertain. These two issues require different policy responses: let us deal with the second issue first.

5.1 Expectations of future benefits

When does it become worthwhile for a single small firm or individual to enter a new or secondary level city, rather than an existing established centre? Such entry decisions typically require very substantial investment in sunk capital – the physical structures of housing, office and factory construction, as well as public infrastructure.[9] The answer therefore depends on the confidence that investors have in the future development of the city, on their ability to capture the future economic benefits, and on their ability to finance investment expenditures. Investors here include both entrepreneurs establishing productive activities, and also house-builders meeting residential demand. Clearly, if investors do not know which cities are likely to develop, are not able to buy land as a way of accessing future capital gains, or cannot borrow to fund their investments, then the growth of new urban centres will be postponed, and the growth of mega-cities exacerbated.

These points suggest a fairly clear policy agenda. Government should be aware that urban infrastructure investment is of value, both in its own right and as a way of signalling to investors that this particular city (as compared with the numerous other potential city sites) is one in which there is commitment to growth. There is a need for long-term property rights in urban land markets, both to provide security and to give investors access to expected future capital gains. Well-functioning long-run credit and mortgage markets are also particularly important, given the highly durable nature of urban capital stock.

5.2 Internalizing externalities

Adopting these measures increases the incentives to be an early mover from an existing mega-city to a new secondary city, but does not remove all market failures. Investors are investing in the expectation of *receiving* the external benefits of a dynamic growing city, but they are not capturing the benefits of the externalities that they are themselves *creating*.

There are two textbook solutions to the problem of internalizing urban externalities. One is to internalize effects by creating ‘large developers’, who buy up the land in the city, attract firms and immigrants (using subsidies), and then take all the land rents. The other is for the public sector to offer Pigovian subsidies for the creation of external benefits (and taxes for the dis-benefits of congestion). In practice, neither of these

---

8 Notice that these are reciprocal externalities, so firms and migrants receive as well as transmit benefits.

9 This section draws on Henderson and Venables (2009).

solutions is likely to be satisfactory. Developers play this role in shopping malls and office developments, but are unlikely to be large enough to capture more than a fraction of the benefits of a city. What, then, is the scope for public intervention through subsidy? As we have seen, there are many different channels through which urban external economies operate, and there remains great uncertainty about their magnitude. It is neither feasible nor desirable to seek to identify and to subsidize every possible source of positive externality in production. This is particularly true, since the subsidy should depend on the present value of future externalities created by an investment. The theory of the second-best warns us to the dangers of piecemeal policy – the possibility that when there are multiple distortions correcting some, but not all of them, this does not necessarily raise welfare. And notions of targeting city size as a whole are fraught with danger. At least conceptually, it should be possible to identify an optimal (or efficient) city size. The available evidence suggests that this is *extremely unlikely* to provide a good policy target in practice. This reflects that fact that there are very large margins of error associated with attempting to identify optimal city size.

6 Conclusion

What do these arguments mean for policy towards the evolving urban structure? There are four points:

- First, institutional reform is needed to remove the most egregious sources of bias towards the primate city. This covers deregulation and measures against corruption in order to reduce the attractiveness of the primate city as a source of rents.
- Second, new and secondary cities can develop only if there are properly functioning land and capital markets. Of course, both these arguments are important in many contexts, and are simply reinforced by their importance in shaping urban structure.
- Third, there is coordination failure in developing new centres, and this inevitably means that government has to play a strategic role. It is an argument for developing infrastructure with a view to facilitating deconcentration, and to signalling which cities or regions are likely to grow next. Notice that this is ‘indicative’, signalling the areas where growth is most likely to occur, not dirigiste, seeking to move activity to unprofitable regions.
- Finally, having economic agents anticipate the benefits of future city growth is probably more important than seeking to internalize the externalities that they create. Pigovian subsidies are therefore not an attractive route to follow. The practical, informational, political-economic, and fiscal costs of such policies are large. Most of the policy gains can, instead, be achieved by shaping expectations, by signalling, and by developing land and capital markets.

Appendix

Each city contains a single sector, and we refer to 'city/sectors', labelled by subscript i = 1, 2 ... S. Countries are labelled by j = 1, 2 ... C, so X_{ij} is the output of city/sector i in country j. We assume $S > C$, and that all countries have the same endowments and technologies.

The unit cost of producing good i in country j, c_{ij}, is:

$$c_{ij} = w_j f\left(X_{ij} : i\right), \qquad (1)$$

where w_j is the wage in country j and the function f is sector specific. The wage rate w_j is the same for all city/sectors in country j as there is perfect within-country mobility of labour and goods. Following Dixit–Stiglitz (1977), there are n_{ij} firms (varieties) in city/sector i in country j and their equilibrium scale of production is $\bar{x}$, the same in all sectors.[10] Total output is $X_{ij} \equiv n_{ij}\bar{x}$ and total production costs in city i country j are

$$n_{ij}\bar{x}c_{ij} = w_j X_{ij} f\left(X_{ij} : i\right). \qquad (2)$$

The function f might have increasing returns, $f' < 0$, external to the firm but internal to the city/sector.

On the demand side, we use the usual combination of Cobb–Douglas and CES preferences. Total income in country j is Y_j, and expenditure is divided equally between sectors; so, country j expenditure on city/sector i is Y_j/S. Within sector, the CES price index for varieties of sector i sold in country k from all countries (j) is

$$G_{ik} = \left[\sum\nolimits_j n_{ij}\left(p_{ij}t_{ijk}\right)^{1-\sigma}\right]^{1/(1-\sigma)}. \qquad (3)$$

where t_{ijk} are trade costs between countries j and k for goods of sector i and the producer price of a unit from country j is $p_{ij} = c_{ij}$. The number of varieties adjusts such that demand for each variety equals the equilibrium scale of production, so

$$\bar{x} \geq p_{ij}^{-\sigma}\left[\sum\nolimits_k \left(t_{ijk}/G_{ik}\right)^{1-\sigma} Y_k/S\right] \qquad n_{ij} \geq 0. \qquad (4)$$

To close the general equilibrium of the model, we assume a perfectly tradable 'agricultural' sector in each country, taken as the numéraire. The wage equals the marginal product of labour, which increases as labour is pulled out of agriculture into the urban sectors. Income is the wage bill plus agriculture rent.

For the numerical example, we set the number of city/sectors at S = 20, and technologies take the form

$$c_{ij} = w_j\left(1 + X_{ij}\right)^{\alpha_i} \qquad (5)$$

---

10 We simply assume that firms operate at this size. The Dixit–Stiglitz model shows how this fixed size assumption can be made a consequence of fixed price marginal cost mark-ups.

where parameters α_i measure the strength of increasing returns to scale and vary across industries in the interval [-0.16, 0.007]. The number of countries is $C = 10$, arranged around a circle, with iceberg transport costs between each pair. We simulate the model, starting with initial conditions in which $t_{max} = 3.2$ (t_{max} is the trade cost factor between the two most distant locations), and there is a small random perturbation of n_{ij} from a uniform distribution. This produces a non-uniform equilibrium. Trade costs are then reduced in small steps, and the new equilibrium found at each step. Further details of the programme are available on request.

References

Arnott, R. J., and Gersovitz, M. (1986). 'Social Welfare Underpinnings of Urban Bias and Unemployment'. *Economic Journal*, 96: 413–24.

Au, C.-C., and Henderson, J. V. (2006). 'Are Chinese Cities Too Small?'. *Review of Economic Studies*, 73(3): 549–76, July.

Bairoch, P. (1988). *Cities and Economic Development: From the Dawn of History to the Present*, translated from French by Christopher Braider. Chicago: University of Chicago Press.

Becker, C. M., and Morrison, A.R. (1999). 'Urbanization in Transforming Economies'. In P. Cheshire and E.S. Mills (eds), *Handbook of Regional and Urban Economics*, Vol. 3. Amsterdam: North Holland.

Chari, S. (2004). *Fraternal Capital: Peasant-Workers, Self-Made Man and Globalization in Provincial India*. Delhi: Permanent Black.

Chen, Y. (1996). 'Impact of Regional Factors on Productivity in China'. *Journal of Regional Science*, 36: 417–36.

Dixit, A., and Stiglitz, J. (1977). 'Monopolistic Competition and Optimum Product Diversity'. *American Economic Review*, 67: 297–308.

Duranton, G., and Puga, D. (2001). 'Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products'. *American Economic Review*, 91(5): 1454–77.

Duranton, G. and Puga, D. (2004). 'Micro Foundations of Urban Agglomeration Economies'. In V. Henderson and J. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4. Amsterdam: North Holland.

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

Glaeser, E. (1998). 'Are Cities Dying?'. *Journal of Economic Perspectives*, 12(2) 139–60.

Hausmann, R., and Rodrik, D. (2002). 'Economic Development as Self-Discovery'. NBER Working Paper 8952. Cambridge, MA: National Bureau of Economic Research.

Henderson, V. (1974). 'The Sizes and Types of Cities'. *American Economic Review*, 64: 640–56.

Henderson, V. (1988). *Urban Development: Theory, Fact and Illusion*. Oxford University Press: New York.

Henderson, V., and Kuncoro, A. (1996). 'Industrial Centralization in Indonesia'. *World Bank Economic Review*, 10: 513–40.

Henderson, V., and Thisse, J. (eds) (2004). *Handbook of Regional and Urban Economics*, Vol. 4. Amsterdam: North Holland.

Henderson, V., Lee, T., and Lee, Y.-J. (2001). 'Scale Externalities in Korea'. *Journal of Urban Economics*, 49: 479–504.

Henderson, V., and Venables, A.J. (2009). 'The Dynamics of City Formation'. forthcoming, *Review of Economic Dynamics*.

Humphrey, J. (1995). 'Introduction'. *World Development*, 23(1): 1–7.

Lall, S., Koo, J., and Chakravorty, S. (2003). 'Diversity Matters: The Economic Geography of Industry Location in India'. WB Policy Research Working Paper 3072. Washington, DC: World Bank.

Lall, S. Shalizi, Z., and Deichmann, U. (2004). 'Agglomeration Economies and Productivity in Indian Industry'. *Journal of Development Economics*, 73: 643–73.

Lee, Y.-J., and Zang, H. (1998). 'Urbanization and Regional Productivity in Korean Manufacturing'. *Urban Studies*, 35: 2085–99.

Lipton, M. (1976). *Why Poor People Stay Poor: Urban Bias in World Development*. Cambridge, MA: Harvard University Press.

Lipton, M. (1993). 'Urban Bias: Of Consequences, Classes and Causality'. *Journal of Development Studies*, 29: 229–58.

Mitra, A. (2000). 'Total Factor Productivity Growth and Urbanization Economies: A Case of Indian Industries'. *Review of Urban and Regional Development Studies*, 12: 97–108.

Mukherjee, R. (1990). 'Conditions of Labour in the Small Scale and Unorganised Sectors in Calcutta and its Neighbourhood'. WEP 2-43/WP.36. Geneva: ILO.

Nadvi, K. (1999). 'The Cutting Edge: Collective Efficiency and International Competitiveness in Pakistan'. *Oxford Development Studies*, 27(1): 81–107.

Nash, M. (ed.) (1977). *Essays on Economic Development and Cultural Change: In Honor of Bert F. Hoselitz*. Chicago: University of Chicago Press.

Patacchini, E., Rice, P., and Venables, A.J. (2006). 'Spatial Determinants of Productivity: Analysis for the UK Regions'. *Regional Science and Urban Economics*, 36: 727–52.

Richardson, H. (1987). 'The Costs of Urbanization: A Four-country Comparison'. *Economic Development and Cultural Change*, 33: 561–80.

Rodrik, D. (2004). *Industrial Policy for the Twenty-first Century*, CEPR Discussion Paper 4767.

Rosenthal, S., and Strange, W. (2004). 'Evidence on the Nature and Sources of Agglomeration Economies'. In J.V. Henderson and J. Thisse (eds), *Handbook of Regional and Urban Economics*, Vol. 4. Amsterdam: North Holland.

Schmitz, H., and Nadvi, K. (1999). 'Clustering and Industrialization: Introduction'. *World Development*, 27(9): 1503–14.

Shukla, V. (1996). *Urbanization and Economic Growth*. Delhi: Oxford University Press India.

Shukla, V., and Parikh, K. (1992). 'The Environmental Consequences of Urban Growth: Cross-national Perspectives on Economic Development, Air Pollution, and City Size'. *Urban Geography*, 12(5): 422–49.

Tacoli, C. (2004). 'The Role of Small and Intermediate Urban Centres and Market Towns and the Value of Regional Approaches to Rural Poverty Reduction Policy'. Paper prepared for the OECD DAC POVNET Agriculture and Pro-Poor Growth Task Team.

Thomas, V. (1980). 'Spatial Differences in the Cost of Living'. *Journal of Urban Economics*, 8: 108–22.

Visser, E. (1999). 'A Comparison of Clustered and Dispersed Firms in the Small-Scale Clothing Industry of Lima'. *World Development*, 27(9): 1553–70.

World Bank (2005). *Doing Business in 2005: Removing Obstacles to Growth*. Washington, DC: World Bank, International Finance Corporation and Oxford University Press.