DISCUSSION PAPER SERIES

No. 6634

FROM CITIES TO PRODUCTIVITY AND GROWTH IN DEVELOPING COUNTRIES

Gilles Duranton

INTERNATIONAL TRADE

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP6634.asp**

ISSN 0265-8003

FROM CITIES TO PRODUCTIVITY AND GROWTH IN DEVELOPING COUNTRIES

Gilles Duranton, University of Toronto and CEPR

Discussion Paper No. 6634
January 2008

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 6634

January 2008

ABSTRACT

From Cities to Productivity and Growth in Developing Countries*

This paper reviews the evidence about the effects of urbanisation and cities on productivity and economic growth in developing countries using a consistent theoretical framework. Just like in developed economies, there is strong evidence that cities in developing countries bolster productive efficiency. Regarding whether cities promote self-sustained growth, the evidence is suggestive but ultimately inconclusive. These findings imply that the traditional agenda of aiming to raise within-city efficiency should be continued. Furthermore, reducing the obstacles to the reallocation of factors and activities, and more generally promoting the movement of human capital and goods across cities may have significant positive dynamic effects as well static ones.

JEL Classification: O18 and R11
Keywords: cities in developing countries, growth and urbanisation

Gilles Duranton
Department of Economics
University of Toronto
150 Saint George Street
Toronto
Ontario M5S 3G7
CANADA
Email: gilles.duranton@utoronto.ca

* This paper was originally the Commission for Growth and Development. Comments and feedback from Gustavo Bobonis, Bob Buckley, Vernon Henderson, Frédéric Robert-Nicoud, Cam Vidler, and especially Patricia Annez were very much appreciated.

Submitted 19 December 2007

1. Introduction

Urban policy interventions in developing countries often have two objectives. The first is to make cities "work better" by improving their provision of local public goods, from sewage to public transport. The second is to limit urbanisation, the movement of people from rural areas to already crowded cities. This dual agenda is driven by the idea that the priority for policy should be to alleviate the grim life of urban dwellers in developing countries and slow down the growth of cities to prevent more misery. While there is no doubt about the abysmal conditions in the slums of Nairobi or Calcutta, is the gloomy outlook of many governments in developing countries about their cities justified? More precisely, we ask two related questions. First, do cities favour economic efficiency? Second, do cities and urbanisation bolster self-sustained growth?

To answer these two questions, an integrated and consistent theoretical framework is first developed. We start from the idea that the entire urban system is an equilibrium outcome (possibly one where politics and other institutional features play a fundamental role) and lay down a simple graphical device to describe the main feedbacks. The framework is then expanded to focus on a number of specific features of cities in developing countries. This highly tractable and flexible framework is also used to interpret the existing evidence about cities and urbanisation in developing countries.

To the first question about whether cities foster (static) economic efficiency, the answer from the literature is a resounding yes. Cities provide large efficiency benefits and there is no evidence that they systematically hurt particular groups. We show below that this result provides support for the first pillar of traditional urban policies (those that seek to improve the functioning of cities). The importance of efficiency benefits from cities also suggests that restricting urbanisation entails losses. Our theoretical framework also underscores key complementarities in urban policy and cautions us about a number of pitfalls.

The second question about the dynamic benefits generated by cities is more difficult to answer. The existing evidence suggests that cities can favour economic growth provided the largest city in a country does not grow too large compared to the others. While this evidence is not strong enough to provide the basis for radical new policy initiatives, it raises further doubts about policies that take a negative stance on cities and discourage labour mobility.

The priority for policy should be more to prevent or curb the worst imbalances in urbanisation rather than slow it down or reverse it. Broadening the focus from within-city efficiency to between-city efficiency even suggests that reducing the obstacles to the reallocation of factors and activities across cities is a highly desirable policy objective.

In conclusion, there is nothing wrong with the first traditional pillar of urban policy in

developing countries, although it may not be for the reasons that are commonly alleged. In addition, instead of restricting the influx of people into the cities, the second pillar of urban policies in developing countries should be to favour the mobility of resources across cities and regions while avoiding their concentration in only one primate city.

The rest of this paper is organised as follows. Our graphical framework is presented in section 2. This section also discusses the main policy issues in the framework. Section 3 reviews the empirical evidence about greater economic efficiency in cities. This section also expands the framework to discuss urban features that are salient in developing countries such as primate city favouritism and dual labour markets. Section 4 focuses on the evidence about the effects of cities on the dynamic of growth and development. Finally, section 5 provides further discussion of a number of policy issues and offers some conclusions.

2. A simple graphical framework to think about urban development

Modelling cities

Economic theories concerned with cities have a common underlying structure.[1] This structure contains three elements: A spatial structure, a production structure, and some assumptions about the mobility of goods and factors. These elements are necessary for any model of cities to be well specified.

Spatial structure. Since cities are located somewhere, some description of geography is obviously needed. It is often convenient to distinguish between the internal geography of cities and their external geography. Internal geography is concerned with land, housing, infrastructure, and internal transport. External geography is about the development of new cities and how cities are located relative to each other and to the location of natural resources.[2]

Production structure. It may be tempting to specify an aggregate production function that directly relates primary factors to the final output, as is customary in much of economic analysis. This standard simplification is often not adequate in our context because cities are characterised by increasing returns to scale and how such increasing

[1] The material in this subsection is adapted from Combes, Duranton, and Overman (2005).

[2] Depending on the focus of the analysis, some aspects need to be explained in great detail while others can be modelled in a very simple fashion. For instance, models that emphasise market access often propose a detailed modelling of the external geography of cities. On the contrary, models that focus on housing supply usually assume a very simple external geography but need to pay more attention to the internal geography of cities and the micro issues related to the operation of land markets. Furthermore, both the internal and external geography of cities are often taken as exogenous. This may be true in the short run, but this need not be the case in the long run as distances within and between cities can be modified following changes in policy or technology.

returns are generated has potentially important policy implications. In particular, detailed assumptions are needed about labour, the nature of products, the production function of individual firms, the input-output structure that links firms, and how the latter compete.

Three main mechanisms can be used to justify the existence of local increasing returns (Duranton and Puga, 2004). First, a larger market allows for a more efficient *sharing* of indivisible facilities (e.g., local infrastructure), risks, and the gains from variety and specialisation. For instance, a larger city makes it easier to recoup the cost of some infrastructure or, for specialised input providers, to pay a fixed cost of entry. Second, a larger market also allows for a better *matching* between employers and employees, buyers and suppliers, partners in joint-projects, or entrepreneurs and financiers. This can occur through both a higher probability of finding a match and a better quality of matches when they occur. Finally, a larger market can facilitate *learning* about new technologies, market evolutions, or new forms of organisation. More frequent direct interactions between economic agents in a city can thus favour the creation, diffusion, and accumulation of knowledge.[3]

Hence, the first general feature that emerges from the literature is that many different mechanisms can generate local increasing returns. The second main feature highlighted by the literature is that sources of local increasing returns are also sources of local inefficiencies.[4] For instance, specialist input producers in a model of input-output linkages may not be remunerated for increasing the choice of inputs in a city. In a matching framework, firms are not compensated for increasing the liquidity of their local labour market. With local learning spill-overs, workers are not rewarded for the knowledge they diffuse. More generally, private and social marginal returns do not in general coincide in a city. This means that urban production is inefficient, in the sense that it does not make the best possible use of local resources.

These two features have important implications. The pervasiveness of market failures hints at a strong role for policy. However, the appropriate corrective policies depend on the exact mechanism at play. The corrective policies associated with urban knowledge spill-overs are not the same as those correcting for imperfect matching on the labour market. Given that many mechanisms generate similar outcomes, identifying the pre-

[3]This typology differs from the traditional Marshallian 'trinity' (Marshall, 1890), which talks of spill-overs, input-output linkages, and labour pooling. In fact the two typologies complement each other. Marshall's is about 'where' those effects take place (market for labour, market for intermediates, and a mostly absent market for ideas) whereas the one used here is about the type of mechanism at stake (sharing, matching, learning). Arguably, these three mechanisms (and their associated market failures) can take place in different markets. Good policies will require knowing about both the type of market failures at play and where they take place.

[4]This is a deep property of any model of increasing returns with a non-degenerate market structure. Without any external effect, increasing returns would lead to a natural monopoly, i.e. a 'factory-town'. The latter certainly exist but are far from being the norm in the urban landscape.

cise sources of agglomeration and their associated market failures is extremely difficult (Rosenthal and Strange, 2004). In terms of policies, this suggests extreme caution when trying to 'foster agglomeration effects'. From a modelling perspective, the fact that a variety of mechanisms can generate increasing returns is very good news because we expect agglomeration effects to be a robust feature of cities. This also suggests that we can assume the existence of local increasing returns without having to rely on a specific mechanism.

Mobility of goods and factors. Assumptions about mobility, both within and between cities, play a crucial role. These assumptions need to cover the geographical mobility of goods, services, primary factors, ideas, and technologies. The extent to which material inputs and outputs are tradable clearly varies across sectors. Among primary factors, land is immobile, although its availability for different uses (e.g., housing versus production) is endogenous. Capital is often taken as highly mobile, with (roughly) the same supply price everywhere. As emphasised below, the (imperfect) mobility of labour, both geographically and sectorally, is a fundamental issue that warrants careful treatment. Finally, the mobility of ideas and technologies determines how production varies across space.

The '3.5-curve' framework of urban development

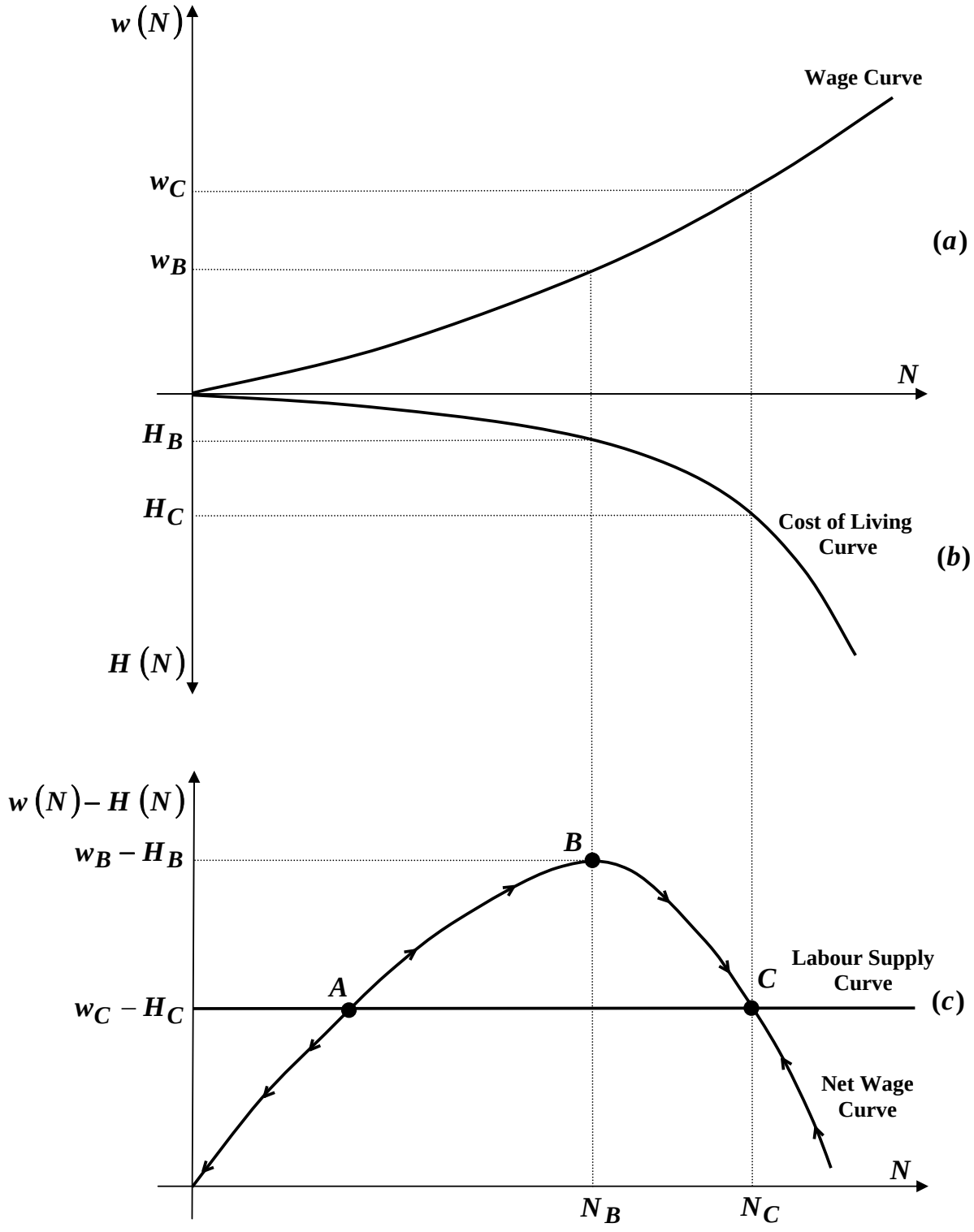
We now present a simple model of an urban system. This model, in the spirit of Henderson (1974), can be represented diagrammatically.

The wage curve. The first key relationship is the city aggregate production function relating total output in a city to the local inputs. If the three primitive factors of production are land, labour and capital and if furthermore land is perfectly immobile while capital is perfectly mobile, the focus of our attention needs to be on labour. Rather than considering output per worker as function of the size of the local workforce, it is technically equivalent, but more fruitful in terms of interpretation, to focus our attention on an inverse-demand for labour that relates the wage of workers to the size of the urban labour force. This curve is represented in figure 1 (a) and referred to as the wage curve in what follows.

In figure 1 (a), the wage in a city is increasing in the size of the urban labour force, reflecting the existence of local agglomeration externalities. The intensity of local increasing returns is measured by the slope of the wage curve.[5] Since the nature and intensity of increasing returns is expected to differ across activities so will the exact shape of the wage curve. This upward-sloping wage curve stands in sharp contrast with 'neo-classical' wage curves that slope downwards. Local increasing returns have received

[5]Whether the wage curve should be concave or convex depends on the specifics of the mechanism(s) that underpins increasing returns. For instance, an ever better match between workers and firms may have some bounded (i.e., concave) benefits whereas the entry of ever more specialised input producers might lead to 'snowballing' (i.e., convex) gains. The exact shape of the productivity curve is ultimately an empirical issue.

Figure 1. Baseline case



a considerable amount of theoretical attention. Modelling cities in this way is consistent with a fundamental stylised fact. Most, if not all, measures of productivity per capita increase with the size of the local population (see below for a discussion of the evidence in developing countries). In turn, a higher productivity in larger cities can explain why a disproportionate share of economic activity takes place in a small number of places rather than spreading uniformly over space as would be predicted by a neo-classical model.

If anything, the level of the wage curve (for any level of employment) is even more important than its slope. The concentration of employment fosters urban productive efficiency. However, this is not the only determinant of urban productive efficiency. The latter also relies on a broad range of productive infrastructures from roads and international airports to well-functioning rental markets for commercial property. This observation also suggests that the level of the wage curve can differ across cities of similar size because of differences in infrastructure and local institutions. Level differences for the wage curve can also occur because of natural endowments and a set of other factors discussed below.

The cost of living curve. The second relationship relates the costs of living in a city to its local employment size. The main components of the cost of living are the cost of commuting, housing and other consumption goods. It seems reasonable to assume that commuting costs increase with population because a larger population implies longer commutes and more congested roads. Similarly, one expects increasing population to drive up the cost of land and thus, of housing. Under some conditions to be clarified below, a larger city with a higher cost of land also implies higher retail costs and thus a higher price for consumption goods.

In figure 1 (b), the cost of living in a city is increasing in the size of the urban labour force, reflecting increasing urban crowding.[6] For reasons that will become obvious, this curve is drawn with a reversed Y-axis. The precise shape of the cost of living curve is driven by the details of the specific mechanisms that underpin it and is ultimately an empirical matter. However, that the cost of living should increase with population is intuitively obvious. As discussed below, the empirical literature strongly supports this notion.

Beyond its shape, the level of the cost of living curve is also of fundamental importance. First, just like with the wage curve the cost of living curve is also riddled with market fail-

[6] An important technical issue needs to be mentioned. An increase in productivity, which raises local wages, may be expected to have a positive effect on the demand for land and thus on its price. If commuting is paid in units of time, higher wages also lead to a higher shadow cost of commuting. Hence an upward shift in the wage curve implies a downward shift in the cost of living curve on the figure. We can ignore these two issues by assuming that the cost of living is paid in monetary terms only and that housing consumption per household is fixed. It is important to note that more formal modelling either ignores these effects or suggests they are second order and thus do not completely offset the direct effect of a shift to the wage curve. Hence, to keep the exposition simple, we ignore these effects in what follows. It would be possible, though cumbersome, to consider this type of link in a more formal model.

ures. For instance, un-priced urban congestion implies an inefficiently high cost of living for any level of population. Poorly defined property rights can also prevent the efficient densification of cities since investors may be reluctant to invest in property upgrading if they face a risk of expropriation, etc. Second, a low cost of living in a city also relies on a vast number of local public goods. In this respect, the provision of roads and public transport to ease commuting is important. The provision of many other public goods of a less capital intensive nature such as security or air cleanliness also matters. Like the wage curve, the cost of living curve is also expected to differ across cities because the latter differ in their shape, availability of land, etc.

The net wage curve. The difference between the wage curve and the cost of living curve is represented in Figure 1 (c) by the net wage curve.[7] On that figure, this difference is bell-shaped. This corresponds to the case where agglomeration economies dominate crowding costs for a small population, while the reverse occurs for a large population. For this to be the case, the wage curve must be steeper than the cost of living curve before a certain threshold and flatter beyond. At this threshold, net wages reach their peak (point B in the figure). This peak can be interpreted as identifying a 'pseudo-optimal' city size, which maximises net wages in the city. The reason this is only a 'pseudo-optimum' (also called a constrained optimum) rather than a true optimum is due to the existence of market failures in production and in the cost of living. These market failures imply that, on the figure, the wage and cost of living curves are not as high as they could be.

The labour supply curve. The second curve represented in figure 1 (c) is an inverse labour supply curve. For any level of net wage, it indicates the amount of labour supplied in the city. For simplicity, we assume that labour supply is a function of the total local population and ignore labour force participation decisions.[8] In that case, this curve essentially captures the migration response to local wages. A flat labour supply curve, as in the figure, implies perfect mobility. In a fully urbanised country, labour mobility takes place primarily across cities and the labour supply curve mainly reflects the conditions in other cities. In a country not yet fully urbanised, labour mobility mostly implies rural-urban migration and the labour supply curve mainly reflects the conditions of rural hinterlands. We return to this important issue below. Note finally that city-specific effects, such as amenities, shift this curve. More attractive cities face a labour supply curve that is below that of less attractive ones. This is because workers accept a lower net wage and are compensated by higher amenities.

Equilibrium. The equilibrium of the model in absence of any policy intervention can now be derived. The intersection between the labour supply and net wage curves deter-

[7]This curve is only a difference between two other curves and thus cannot count as a independent relationship. Hence, the '3.5 curve' name for this framework.

[8]We nonetheless distinguish between formal and informal sectors below.

mines the equilibrium. It corresponds to a situation where workers obtain the net wage they require to come to and stay in the city, given the local population. The intersection between these two curves may not be unique. In figure 1 (c), the two curves intersect twice (at points A and C). The labour supply curve first cuts the net wage curve from above (at point A) and then from below (at point C). Point A is not a stable equilibrium. It is easy to see that a small positive population shock raises the net wage. In turn, from the supply curve, this higher net wage attracts more workers, which again raises net wages and this process continues until the city reaches point C. By the same token, a negative shock if the city is at point A leads population and wages to fall to zero. Turning to the second intersection at point C, a similar argument verifies that this equilibrium is stable. From figure 1 (c), once we have established the equilibrium population, N_C, we can trace upwards to figures 1 (a) and 1 (b) to read off the equilibrium wage, w_C, and cost of living, H_C, respectively.

Before turning to welfare and policy issues, note that, to the extent that agglomeration effects take place within sectors, cities have a tendency to specialise. To see this, it is useful to consider two hypothetical activities in a city. These two activities are entirely unrelated and each has its own productivity curve and a given initial level of employment. Workers in both activities face the same cost of living since everyone is competing for the same land. On the other hand, the two activities offer, in general, different wages. Then, workers are expected to leave the activity with the lowest net wages and move to the other. This movement ends up only when the city is specialised in a single activity.[9] More generally, it is inefficient to have 'disjoint' activities in the same city since they bring no benefit to each other and crowd each other's land market. We thus expect the economic composition of cities to reflect this. Hence, should agglomeration effects take place mostly within sectors, cities should be specialised. If instead, agglomeration effects take place at a broad level of aggregation with strong linkages across sectors, more diversity should be observed.[10]

Finally, it is important to note that the analysis of cities is inherently a 'general equilibrium' problem, in which the researcher has to look beyond the direct effect of a change and assess the induced changes that follow. Doing this is possible only if there is a clear analytical framework within which the various effects interact.

[9]Should, for some unspecified reason, the two activities have the exact same returns, a small employment shock, positive or negative, in any of the two activities again creates a small asymmetry between the two activities and leads again to full specialisation.

[10]As made clear below, this is not an empirically empty statement able to rationalise anything. We can measure the strength of agglomeration effects within vs. between industries independently. Estimating strong agglomeration effects across sectors and, at the same time, observing very specialised cities would clearly be problematic.

Welfare in the 3.5-curve framework

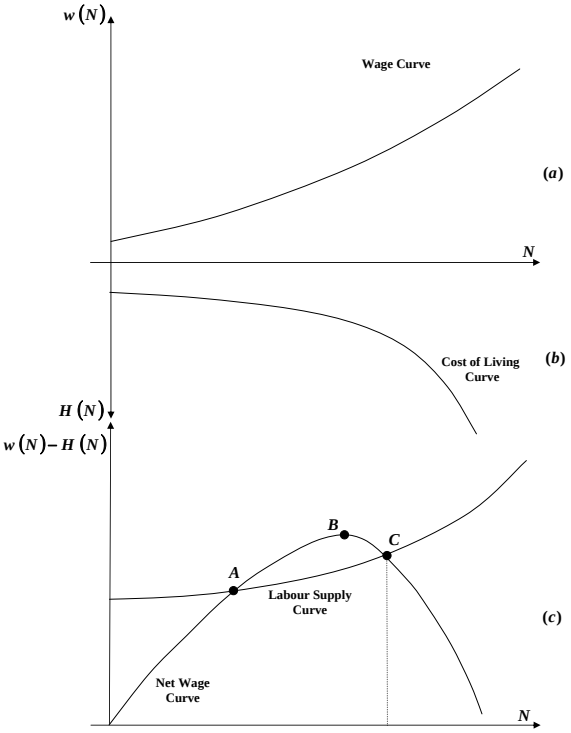
To discuss policy, we proceed in stages. This subsection discusses the main welfare issues. This discussion should be viewed more as way to reach a deeper understanding of our framework than a practical policy guide. General policy issues are addressed in the next subsection before turning to specific policy problems in a development context in section 3.

Uncompensated externalities in production. The first source of inefficiencies stems from the production structure itself. As argued above, the microeconomic foundations of the increasing returns operating inside cities are all associated with market failures. First, the indivisibilities at the heart of sharing mechanisms generate a number of inefficiencies. Like all indivisibilities, they imply that only a limited number of players enter the market. This results in imperfect competition and the (socially inefficient) exploitation of market power. If new entrants increase the diversity of, say, local inputs, they are unlikely to reap the full benefits of this increase in diversity. We also expect firms to make their entry decision on the basis of the profits they can make rather than the social surplus they create. Under imperfect competition, this is again inefficient. Second, with matching mechanisms, a different set of market failures is at play. For instance, firms neglect the positive effects of their vacancies on the job search of workers. Finally, there are also many possible market failures associated with learning mechanisms. Under imperfect intellectual property rights protection, firms are likely to invest too little in knowledge generation. In absence of rewards for knowledge diffusion, too little of it takes place. Firms in cities may also be reluctant to train their workers if they expect them to be poached by competition in the future, etc. These are only several of the inefficiencies that can occur when production takes place under increasing returns.

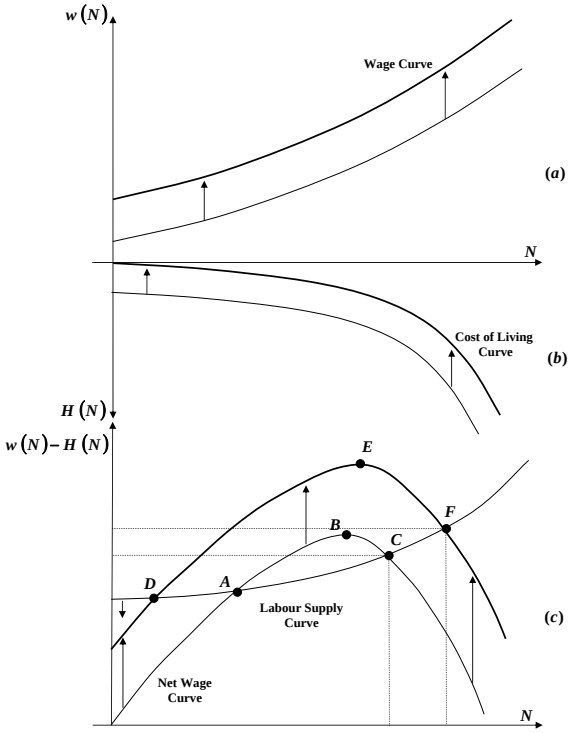
If these inefficiencies were suppressed, wages would increase in the city for any level of employment. Starting from the wage curve in part (a) of figure 2 (i), solving for the inefficiencies in production leads to the thick line in part (a) of figure 2 (ii).

Uncompensated externalities in cost of living. The second source of market failures is related to the cost of living curve. If the private marginal costs paid by residents were equal to social marginal costs (i.e., the costs to the economy), there would be no inefficiency in cost of living. With no congestion, a perfectly functioning land market, and redistribution of the land surplus, this equality between private and social marginal costs holds naturally. Empirically, we expect neither of these three assumptions to be satisfied: land markets are subject to significant frictions and are strongly regulated through planning and zoning regulations, increases in land values are not taxed away, and, as cities get more crowded, congestion becomes more important. About the latter, note that traffic congestion is a major form of congestion in cities, but by no means the only one. Most

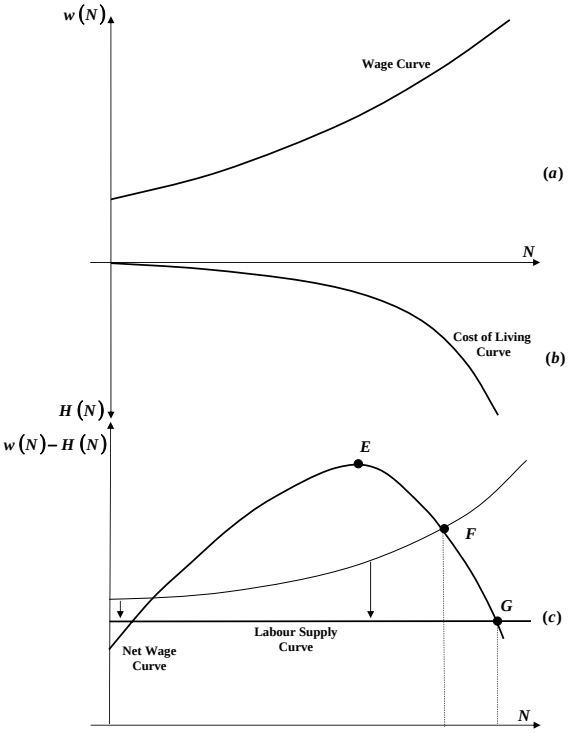
Figure 2. Welfare effects



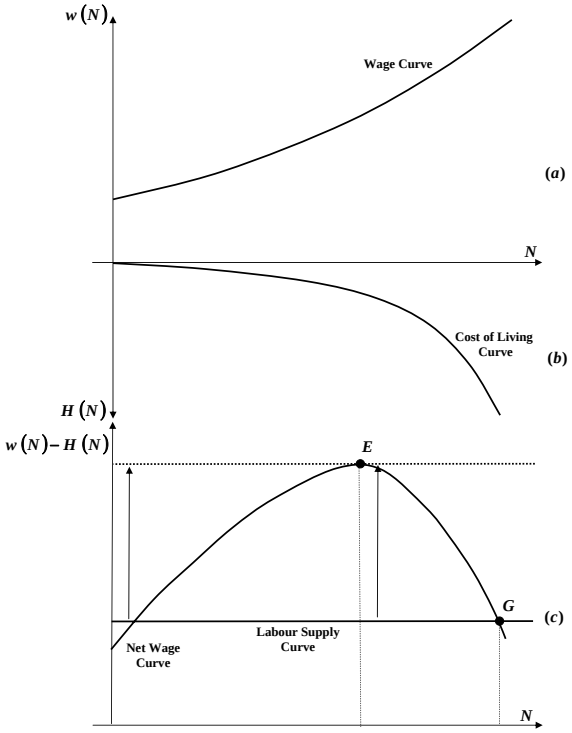
(i) Initial situation



(ii) Fixing market failures in production and cost of living



(iii) Making labour perfectly mobile



(iv) Solving for the city co-ordination failure

local public goods, from parks to cultural events, and many amenities are also subject to negative congestion externalities. Poorly defined property rights over urban land also constitutes a critical issue in many developing countries.

The main implication of congestion and frictions on the land market is that the cost curve in the absence of corrective policy is distorted. With proper corrective policies, it should be possible to reduce costs of living for any population level in the city. For instance, a congestion tax would reduce the level of traffic congestion in the city and can increase total surplus. Starting from the cost of living curve in part (b) of figure 2 (i), fixing the inefficiencies in cost of living leads to the thick line in part (b) of figure 2 (ii).

A higher wage curve and a lower cost of living imply a higher net wage curve in part (c) of figure 2 (ii). After curing the market failures in production and cost of living, the net wage curve and the labour supply curve intersect at points D and F (rather than A and C prior to the policy interventions). The net wage curve has its maximum at point E instead of B. Just like A, point D indicates an unstable equilibrium. The only stable equilibrium is in F. This new equilibrium offers a higher net wage than the one with no intervention at point C. Population is also higher. This is because solving for the inefficiencies in production and the cost of living makes the city more attractive. In turn, the labour supply response implies that workers migrate to the city.

The extent to which a higher net wage curve leads to a higher population versus a higher net wage depends on the slope of the labour supply curve. Perfect mobility (i.e., a flat supply curve) implies that all the gains from curbing the inefficiencies are translated into a higher population and more crowding. In absence of mobility, a vertical labour supply curve implies that the upward shift of the net wage curve leads only to higher net wages. It is also important to note that after solving all the inefficiencies associated with production and cost of living, the equilibrium in F does not coincide with the first best at point E.

Barriers to migration. The third source of inefficiencies is related to the labour supply curve and thus the migration process. The labour supply curve is driven by two different set of forces. First, it echoes the net wage in the rest of the economy. For many developing countries, we expect the labour supply curve to be mostly a reflection of rural earnings. In that case, a higher net wage in rural areas implies a higher labour supply curve. Second, barriers to migration are also reflected in the labour supply curve. More costly mobility implies a higher and steeper labour supply curve.

Eliminating obstacles to mobility in part (c) of figure 2 (iii) thus leads to a lower and flatter labour supply curve. As a result, the equilibrium shifts to point G. Interestingly, this new equilibrium implies a larger population and a *lower* net wage than the previous situation at point F. The net wage decreases because reducing barriers to mobility makes it easier for newcomers to settle in the city. Since the city is already in the region where the

marginal agglomeration gains are dominated by the marginal losses in urban crowding, an influx of newcomers lowers the welfare of existing residents.

This negative result underscores a fundamental policy issue. Urban economies are second-best economies. Nothing guarantees that fixing a market failure always brings the city closer to optimality. We already solved for the market failures in the wage and cost of living curves before removing barriers to migrations. Nonetheless, there is yet another market failure which prevents cities from reaching their optimal size. Unless this last market failure is also fixed, reducing the barriers to mobility need not improve local welfare.[11]

The city co-ordination failure. As made clear by part (c) of figure 2 (i), the equilibrium with no policy intervention (point C) is not efficient and is located to the right of the pseudo-optimum (point B). Without any corrective policy, existing cities are *too large* with respect to their pseudo-optimum. Put slightly differently, employment concentrates into too few cities that are too big.

The reason behind this inefficiency is a co-ordination failure. Fixing the inefficiencies embedded in the wage, cost of living, and labour supply curves changes nothing to the city co-ordination failure. In part (c) of figure 2 (iii), the equilibrium size, point G, is still inefficiently large compared to the first-best in E. It is easy to understand why this inefficient situation can be sustained. No one wants to move alone and develop a new city because it would mean forming a very small and thus very unproductive city. It is worthwhile to move to a new city only if it is already large enough or if a big enough group of workers and firms decide to co-ordinate their move. The creation of such a new city would be desirable for everyone since existing cities would become smaller and thus be able to offer higher net returns. The problem is of course that, in absence of corrective policy (or market for cities), there is no mechanism to co-ordinate the movement of workers to new cities.

To solve this governance problem and to reach the first-best in E, two solutions can be envisioned. First, the city under consideration may directly restrict its population size to reach point E. Doing so implies rejecting residents and sending them to places where they are worse off. Depending on where these rejected residents go, this can increase the cost of living in other cities or increase rural population and thus arguably lower agricultural earnings. Hence, this first solution is a partial equilibrium response to the city co-ordination failure which generates negative general equilibrium effects.

The second alternative is to create new cities and co-ordinate the move of residents to these new cities. This creation of new cities implies a reduction of population for previously oversized incumbent cities and thus an improvement in welfare for their remaining

[11] Even though *local* welfare may decrease, aggregate welfare increases. This point is made clear later.

residents. Should new cities be filled by rural migrants, this would also imply a decrease in rural population and thus arguably an increase in agricultural earnings. In turn, a higher welfare outside the city implies a higher labour supply curve. In this case, the general equilibrium effects are positive.[12] New cities can then be created until the labour supply curve hits the net wage curve at point E. At this stage the entire urban system is fully efficient.

Practical policy considerations

It is now time to take a more practical look at urban policies. A fundamental policy question should first be answered: Is it worth it for policy to bother about cities at all? Cities are clearly riddled with market failures: production is inefficient, congestion is rife, and overcrowding is expected to be the rule. The above welfare analysis also makes clear that full urban efficiency is extremely demanding to achieve. Hence, there is a strong temptation to view the 'urban problem' in developing countries as an unmanageable pathology and neglect cities. That would be wrong. Having numerous inefficiencies only implies that cities are much less efficient than they could be and that there are important gains from well-designed urban policies. Furthermore, existing urban inefficiencies do not imply that cities are less efficient than their rural alternatives. Actually, the very success of cities in developing countries points to the opposite. However suboptimal they may be, cities typically offer higher returns and better long-term opportunities. Neglecting cities and restricting their access can only have negative consequences: a worsening of urban inefficiencies and 'overcrowded' rural areas, which in turn implies low returns to agriculture and an exacerbation of rural poverty.

Going into the details, two important points need to be made about the wage curve. First, the wage curve reflects a considerable number of evolutions that are determined well beyond the city under consideration. To take a simple example, many developing countries have policies that distort agricultural prices relative to manufacturing prices. Since cities in developing countries are specialised into manufacturing and services, any increase in relative manufacturing prices is likely to translate into higher urban wages and thus a higher wage curve. In turn, this should lead to larger cities. More generally, technological evolutions and government policies are going to be reflected in the wage

[12] More generally, general equilibrium effects (i.e., what happens outside the city) matter and play a fundamental role. Changes taking place outside the city under consideration affect the labour supply curve and thus its equilibrium. These interdependencies can mean that a worsening of the situation outside the city (i.e., a lower labour supply curve) leads to an influx of new residents and a worsening of the welfare in the city as well. The importance of general equilibrium effects also implies that improving the functioning of only one city makes it grow but has ambiguous implications regarding local welfare. A better functioning city becomes attractive and the new residents can crowd out all the gains.

curve. These changes are expected to affect the level of the wage curve and, sometimes, its slope.[13]

The second issue with the wage curve is related to the market failures beneath it. To repeat, the existence and growth of cities is driven by a variety of mechanisms whose relative importance is extremely difficult to identify empirically. The market failures associated with these mechanisms then all require different corrective policies. For instance, corrective policies aimed at dealing with labour market matching problems have nothing to do with those aimed at fostering knowledge diffusion, etc. Put differently, we need some corrective policies for inefficiencies which we know close to nothing about. This suggests some caution.[14]

Given the limited possibilities for policy to raise the wage curve, the cost of living curve is a more promising area of action for city governments. The main reason is that many of the key determinants of the cost of living curve such as traffic congestion are reasonably well-identified problems which we know quite a bit about. From sewage to public transport, there are many components of the cost of living curve for which local governments can make a big difference. The second main policy issue related to the cost of living curve has to do with poorly defined property rights and the inefficient operations of the land market. This salient issue in many developing countries is dealt with at greater length below. Finally, note that many other policies of local governments such as the provision of public goods and amenities also get reflected into the cost of living curve. This only reinforces the point that the cost of living curve is the traditional area of expertise of city governments and should remain so.

Turning to labour mobility, it is clear that a flatter labour supply curve can potentially lead to important welfare gains by allowing workers to move from low net wage areas

[13]Regarding the latter, think for instance about progress in telecommunication technologies which may affect the intensity of agglomeration effects.

[14]Furthermore, it is also the case that these market failures are likely to occur in all cities. Creating a more efficient labour market or favouring the diffusion of knowledge is more appropriate for central rather than local governments. The main tool for local governments with respect to the wage curve should then be the provision of productive local public goods. However, a complete discussion of this issue, including of the qualifications that apply to the preceding statement, would take us well beyond the scope of this paper (see Helsley, 2004; Epple and Nechyba, 2004, for recent reviews).

to high net wage cities.[15] As hinted above, this increase in mobility is best carried out by central governments since any city that unilaterally increases labour mobility may decrease its local welfare. This prescription of greater labour mobility runs contrary to many policies in developing countries that aim instead at restricting internal migrations. Given the importance of labour mobility to improve efficiency in the short run, and possibly to foster economic growth in the long run as well, this issue is further developed in the next two sections.

The last prescription of the framework regards the fact that cities tend to be too large in equilibrium calling for the creation of new cities and the co-ordination of their settlement. This is of course a practical minefield and this recommendation should be taken with extreme caution. Past experiences of city creation, and more particularly of capital city creation, in developing countries have often led to mixed results (or worse). While in the United States new cities are often created by private developers (Henderson and Mitra, 1996), few developing countries appear to be able (or willing) to follow suit. Besides, developing countries already appear to host many very small cities. Hence, rather than creating new cities, the challenge is the lack of growth of many small cities.[16] These issues are discussed at greater length below.

3. What's special about cities in developing countries?

Empirical support for the framework

Before going deeper into policy issues, let us first discuss our framework in light of the empirical evidence in developing countries. In brief, the literature offers support for all its main building blocks: an upward-sloping wage curve, costs of living rising with city size, a bell-shaped net wage curve, and some labour mobility driven by net wage differentials.

[15] An important technical caveat applies here. In absence of pure externality in the wage and cost of living curves, it is always good from an efficiency (and welfarist) perspective to have workers move from low wage (rural) areas to high wage cities. This result holds even though cities are in a region of decreasing returns. This occurs because the difference between the net wage curve and the labour supply curve exactly measures the social marginal gain of one more migrant in the city. This is no longer true in the presence of pure externalities. Then, a new worker into the city can raise the wage of all other workers (through agglomeration effects) but also increase their cost of living. If the increase in cost of living associated with the externality is very large, the private gains from the move for the migrant and the higher wages for all workers can be more than offset by the cost of living loss of all the other inhabitants. Given how big spatial disparities can be in developing countries (Aten and Heston, 2005), the congestion externalities would need to be extremely large for migration from poor to rich areas not to raise overall output. This case remains to be made empirically.

[16] One may object to this and argue that all cities, small and big, are already expected to be oversized. There is no contradiction here if one acknowledges that cities should grow with their pseudo-optimal size. Note further that the growth and industrialisation of small cities is all the more important since developing countries often have their international comparative advantage in mature manufacturing sectors. Small and mid-size cities are natural locations for such activities (Henderson, 1997).

Starting with the wage curve, there is a large literature that documents the existence of agglomeration effects in developed economies (see Rosenthal and Strange, 2004, for a review). The main conclusion of this literature regards the existence of scale economies of 3 to 8% (i.e., the doubling of the size of an activity in a city raises its local productivity by 3 to 8%). These agglomeration effects take place both within sectors (localisation economies) and between (urbanisation economies). Although there is far less research about agglomeration effects in developing countries, the results are usually similar.

As in developed countries, studies of agglomeration effects in developing countries regress some productivity outcome in cities (and sectors) on city measures of economic activity within or across sectors. See Rosenthal and Strange (2004) and Combes, Duranton, Gobillon, and Roux (2007) for more details about this type of methodology. Following Henderson (1988)'s study of localisation economies in Brazil, several studies have found quantitative evidence of localisation effects. For instance, Henderson, Lee, and Lee (2001) find localisation economies for Korean industries, more particularly traditional industries. Lall, Shalizi, and Deichmann (2004*b*) for India, and Deichmann, Kaiser, Lall, and Shalizi (2005) for Indonesia provide similar evidence, albeit less strong. Further evidence about localisation effects can be found in a number of case-studies looking at a wide variety of countries and sectors (see Overman and Venables, 2005, for references).

There is also evidence of urbanisation economies in developing countries. Henderson *et al.* (2001) show that they matter for advanced sectors in Korea. There is also evidence of urbanisation effects for India. It is rather weak in Lall *et al.* (2004*b*) but much stronger in Lall, Koo, and Chakravorty (2003). Deichmann *et al.* (2005) also find mild evidence of urbanisation effects in Indonesia for a number of sectors. The results of Au and Henderson (2006*a*) and Au and Henderson (2006*b*) about Chinese cities are also consistent with a mix of localisation and urbanisation effects. The evidence is further discussed in Henderson (2005) and Overman and Venables (2005) who provide detailed reviews of agglomeration findings for developing countries.

Strong localisation economies are expected to foster the growth specialised cities while strong urbanisation economies foster that of diversified cities. Evidence of both localisation and urbanisation economies is consistent with the existence of diversified cities and specialised cities in developing countries.[17]

Two main criticisms can be made to these studies. First, they usually do not control for the individual characteristics of workers (observed and unobserved). It could be that measured agglomeration effects only reflect the sorting of more productive workers in bigger and more specialised cities rather than true agglomeration economies. Using French data,

[17]Despite strong evidence about localisation economies, it seems that there are few specialised cities in many developing countries relative to the US. Other factors such as high transport costs must thus be invoked to explain these weak patterns of urban specialisation. See below for more on that.

Combes, Duranton, and Gobillon (2008*a*) show that such sorting is empirically important and goes a long way towards accounting for observed spatial disparities. Nonetheless, controlling for sorting does not make agglomeration effects vanish. Second, most of the available findings concern primarily the formal sector. Hopefully household surveys that cover both the formal and informal sectors will be used more widely in the future. At this stage, we can only note that the linkages between formal and informal sector firms are often intense which suggests that agglomeration effects are generated within both the formal and informal sectors with benefits that accrue to both. It is also worth mentioning that the case-study evidence that supports the existence of agglomeration effects also strongly supports the idea that the informal sector is a strong contributor.

Turning to the cost of living curve, the evidence is scarce. Early works by Thomas (1980), Henderson (1988), and Richardson (1987) show a fast rise in the cost of living with city size. These cost of living findings are confirmed by more recent work from Henderson (2002*a*) who looks at a broader cross-section of cities. He finds the elasticities of various cost of living measures to cities size to be between 0.2 and 0.3. Finally, Timmins (2006) develops a novel methodology to infer the 'true cost of living' from widely available data using a model of location choice. He implements his approach on Brazilian data and finds that the cost of living increases with city size above a certain threshold.[18]

The evidence about the net wage curve is extremely thin. The main difficulty here is that, with sufficient labour mobility, we expect all cities to be on the decreasing portion of the net wage curve following the stability argument exposed above. Having most cities on the decreasing portion of the net wage curve is, for instance, consistent with the Brazilian findings of da Mata, Deichmann, Henderson, Lall, and Wang (2007). Direct evidence about net returns to size being bell-shaped is provided by Au and Henderson (2006*a*) and Au and Henderson (2006*b*) for cities in China. They exploit the fact that the Chinese government has imposed very strong barriers to labour mobility. As a result, a very steep labour supply curve is expected in China. Provided it is steep enough, some cities can be too small in equilibrium and a bell-shaped net wage curve can be estimated. Interestingly, Au and Henderson (2006*a*) and Au and Henderson (2006*b*) find that Chinese cities tend to be significantly undersized. This results in large income losses. The other finding of Au and Henderson (2006*b*) is that the net wage curve is quasi-flat after its maximum. This suggests that cities may become grossly oversized under free mobility but that the costs of being oversized are small (unlike the costs of being under-populated).

The migration mechanism that underlies the labour supply curve has been widely studied. Greenwood (1997) proposes a general survey of internal migrations in developed and developing countries while Lall, Selod, and Shalizi's (2006) review focuses on developing

[18]Interestingly, he also finds that the cost of living also decreases with population below the threshold. This suggests that the cost of living is high in large cities and in small and isolated places.

countries. A key finding of the literature is that internal migration flows in developing countries are consistent with an upward-sloping labour supply curve. Among many, Brueckner (1990) and Ravallion and Wodon (1999) find that the direction of migration flows is consistent with existing differences in net wages. In their work in Bangladesh, Ravallion and Wodon (1999) also address the slope of the net wage curve by showing that there are persistent differences in living standards across areas despite the absence of formal barriers to mobility.

Closer to the spirit of the labour supply curve in our framework, da Mata *et al.* (2007) estimate a population supply function for Brazilian cities and find the elasticity of population to income per capita to be between 2 and 3. This is quite elastic, but still far from perfect mobility. Barrios, Bertinelli, and Strobl (2006) show that, in sub-Saharan Africa, there is a direct link between climate, which directly affects living standards in rural areas, and urban growth. This type of finding is consistent with an important role for shocks that shift the labour supply curve, up or down. It also suggests that in less advanced countries, the labour supply curve is mostly driven by living conditions in the countryside rather than in other cities.[19] This, in itself, is consistent with the traditional notion of surplus labour (Lall *et al.*, 2006). The final conclusion that can be drawn from Barrios *et al.* (2006) is more subtle. There is a negative correlation between urban growth and the welfare of urban dwellers. This negative correlation may explain why many developing country governments attempt to restrain urbanisation. However, this correlation is not causal. Negative agricultural shocks lower the labour supply curve and workers flock to the cities thereby lowering urban welfare. Hence, cities can still offer efficiency benefits despite a negative correlation between urban growth and urban net wages. On the contrary, preventing rural dwellers from moving to the cities can make them worse off.

We now turn to the development of new cities. Although the theoretical literature has recently made progress on efficient city development (Henderson and Venables, 2006), little is known empirically about it. Using data about world cities spanning several decades, Henderson and Wang (2007) use a 100,000 population threshold to track the entry of 'new cities'. Several interesting findings emerge. First, in a typical country, the rate of growth in the number of cities is not statistically different from that of population growth. This suggests that new cities do indeed rise and the rough proportionality between the entry of new cities and population growth is reassuring. Nonetheless, this does not say much about the efficiency of the process of city creation beyond ruling out the notion that it is entirely dysfunctional. Henderson and Wang (2007) also show that the emergence of new cities is favoured by democratisation and government decentralisation while it is slowed down by having a large fraction of educated workers. With world urban population

[19]The corollary of this is that worsening rural conditions, which lower the labour supply curve, lead to 'urbanisation without growth', as documented for instance in Fay and Opal (1999).

growing by about a hundred million per year, there is no doubt that those issues deserve further attention.

Finally, there is scant evidence about cities being too large in general. Very strong barriers to labour mobility have made Chinese cities too small according to Au and Henderson (2006*a*,*b*). There is unfortunately no other study that attempts to look at this question without making heroic assumptions about what optimal city size is. In light of the framework exposed above and its predictions about cities being oversized, casual observation of cities in developing countries raises some apparently puzzling facts with respect to city size. While many mega-cities in developing countries such as Karachi in Pakistan with a population well above 10 million are arguably 'too big', most cities in developing countries are much smaller. In Thailand for instance, there is only one 'large' city with a population above 300,000. How is it that Bangkok could be too big with a population nearing 6 million while the fifth largest city in Thailand, Chiang Mai, could also be too large with a population only around 150,000? To solve this puzzle, we can extend our framework into two directions to consider the issue of primate city favouritism and that of market access.

Primate city favouritism

Urban primacy is a well-known feature of urbanisation in developing countries (e.g., Henderson, 2005). The causes of why in so many developing countries the largest city is often disproportionately larger than the second largest are still disputed. A careful reading of the evidence nonetheless suggests two fairly simple answers. Urban primacy is sometimes attributed to protectionist trade policies. In the model of Krugman and Livas Elizondo (1996), trade liberalisation reduces urban primacy because it allows all cities to import differentiated goods from abroad. In turn, this reduces the tendencies for the agglomeration of manufacturing in a single core city. Although correct, their model is arguably very particular. Rather than equalising market potential, it seems more reasonable to assume that trade liberalisation gives privileged market access to coastal cities or to cities close to trading partners. Then, inland primate cities can obviously see their dominance reduced by trade liberalisation. Mexico City since NAFTA, which served as motivating example to Krugman and Livas Elizondo (1996), may be an illustration of this. On the contrary, coastal primate cities can see their dominance reinforced by trade liberalisation. Think of Buenos Aires in Argentina whose primacy remained unabated despite trade liberalisation. The effects of trade policy are thus theoretically ambiguous.

More generally, the empirical support for trade-based explanations of urban primacy is weak. Studies that find a negative effect of trade on primacy often do so because they fail to control properly for other channels that can influence primacy and are correlated with

trade policy (e.g., Moomaw and Shatter, 1996). The better and more recent studies (Ades and Glaeser, 1995; Nitsch, 2006) suggest that trade plays no systematic role with respect to urban primacy.

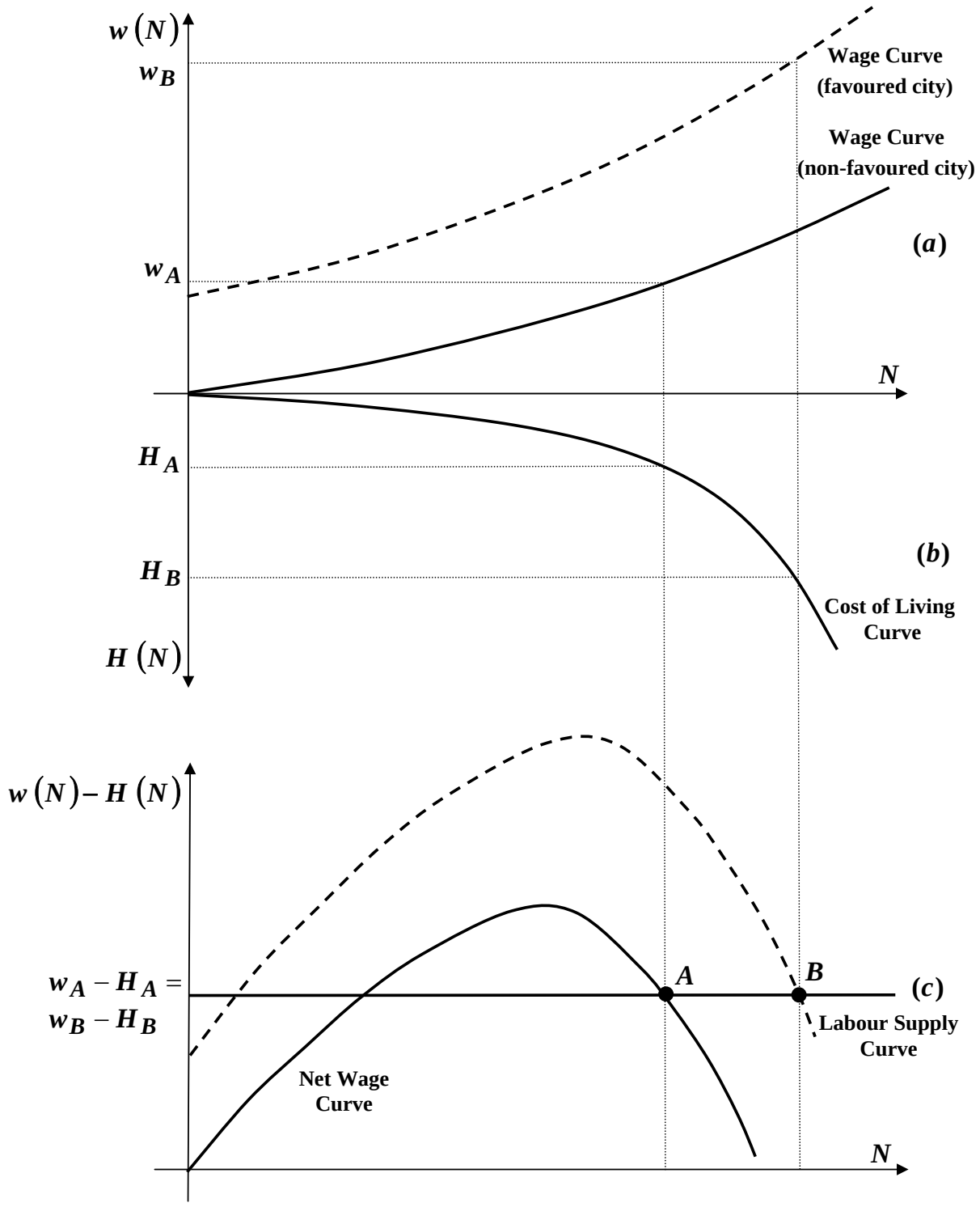
Instead, political and institutional factors appear to be at the root of the primacy phenomenon. There is strong evidence of a positive association between unstable and undemocratic regimes and urban primacy (Ades and Glaeser, 1995; Davis and Henderson, 2003). The exact underlying mechanism(s) is nevertheless not fully elucidated. The story is often told in terms of dictatorial regimes bribing the residents of the primate city because they are afraid of being overthrown by social unrest. Direct evidence about this mechanism is lacking. Furthermore, this type of explanation appears to assume fairly strong state institutions able to tax their countryside and redistribute the proceeds to the primate city. On the contrary, it may be argued that undemocratic and unstable regimes are weak and favour primate cities 'by default'. Primate city favouritism can work through a myriad of small decisions from underpriced gasoline and better provision of local public goods to better business opportunities for government cronies in the primate city (Henderson and Becker, 2000; Henderson, 2002*b*,*a*). In this respect, the many regulations and permits that govern economic activity in most developing countries could play an important role. Being close to a centre of power makes it easier to obtain permits or to circumvent their necessity. A complementary explanation points at a better road infrastructure linking the primate city to the rest of the country (Saiz, 2006). We can thus speak of primate city favouritism, but keeping in mind that this favouritism takes place in many different ways.

Primate city favouritism can readily be incorporated into our framework. We assume for simplicity that favouritism (or the lack thereof) affects primarily wages. Earnings are higher than they would otherwise be in the favoured cities. They are also lower than they would otherwise be in the other cities since favouritism can only come at a direct cost for them. Part (a) of figure 3 represents the wage curve for the favoured city and the lower wage curve for a non-favoured city. As a result, with the cost of living curve being the same in both cities on part (b) of the same figure, the net wage curve of the favoured city is above that of the non-favoured city. It is then easy to see that the equilibrium size of the favoured city is larger than that of the non-favoured city in part (c) of figure 3.[20] Because of general equilibrium effects, the labour supply curve is also lower than it would be in absence of favouritism in part (c) of figure 3.

The potentially large misallocation of resources suggests that some effective policies

[20]On the graph, the favoured city is larger but not disproportionately larger than the non-favoured city. This for clarity only. With a flatter downward sloping portion of the net wage curve, this difference can be made much bigger. According to Au and Henderson (2006*b*), the net wage curve is empirically rather flat in China.

Figure 3. Primate cities favouritism



to reduce urban primacy are needed.[21] However, dealing effectively with this problem is going to be hard. First, primate city favouritism manifests itself in many different ways and there is no definite evidence at this stage about which channel(s) matters most. Interestingly, the Korean experience hints that administrative deregulation may be a powerful tool to reduce urban primacy (Henderson *et al.*, 2001). Red-tape may be costly for all businesses but more so for those located far away from the main centre so that deregulation is more beneficial to them. Second, the political economy associated with urban primacy may be very difficult to break. Cronies who benefit handsomely from their proximity to the political power are unlikely to easily accept a levelling of the playing field.

Internal market access

The proposition that a good access to markets matters can be traced back at least to Harris (1954). It was recently revived by Krugman (1991) and the ensuing work. This body of work is referred to as the New Economic Geography and is summarised in Fujita, Krugman, and Venables (1999), Baldwin, Forslid, Martin, Ottaviano, and Robert-Nicoud (2004), and Combes, Mayer, and Thisse (2008*b*).

Krugman's (1991) model considers two regions (rather than cities) and two sectors. Agriculture produces a homogenous good under constant returns. For simplicity, this good is assumed to be perfectly tradable and is produced by immobile workers. Manufacturing firms operate under increasing returns. Each monopolistically competitive firm employs mobile manufacturing workers to produce a separate variety of differentiated product, which is demanded by consumers in both regions. Manufacturing varieties are costly to transport between regions so that firms' sales have a 'home-market bias'.

The wage of manufacturing workers is determined as follow. Consider a 'high' level of transport costs. This is a reasonable assumption for most developing countries.[22] Due to high transport costs, local manufacturing producers are partly insulated from imports from the other region. Local producers can thus charge high prices, which in turn imply high local manufacturing wages. If manufacturing expands, the local market gets more crowded. This happens because, although the expansion of manufacturing also implies a larger local market, the size of the latter does not increase proportionately (remember the fixed agricultural sector). Furthermore, with high transport costs, very little of the

[21] One may argue that such policies have been attempted for a long time. This is true but many of these policies, such as the relocation government activities, did not provide the right incentives for residents to relocate and took place in a framework of highly controlled labour mobility.

[22] Further details about this case and a complete explanation of the low transport cost case can be found in Combes *et al.* (2005). Interestingly, the tradeoff between the two main forces described below is resolved differently when transport costs are low.

increase in manufacturing output gets exported. Basically, when transport costs are high manufacturing wages decrease with the size of the local manufacturing workforce.

This alone would lead to a downward sloping wage curve and a dispersion of manufacturing. However it seems difficult to completely write off the urban agglomeration effects described above. This suggests that the wage curve is determined by opposing forces: market access vs. agglomeration economies.

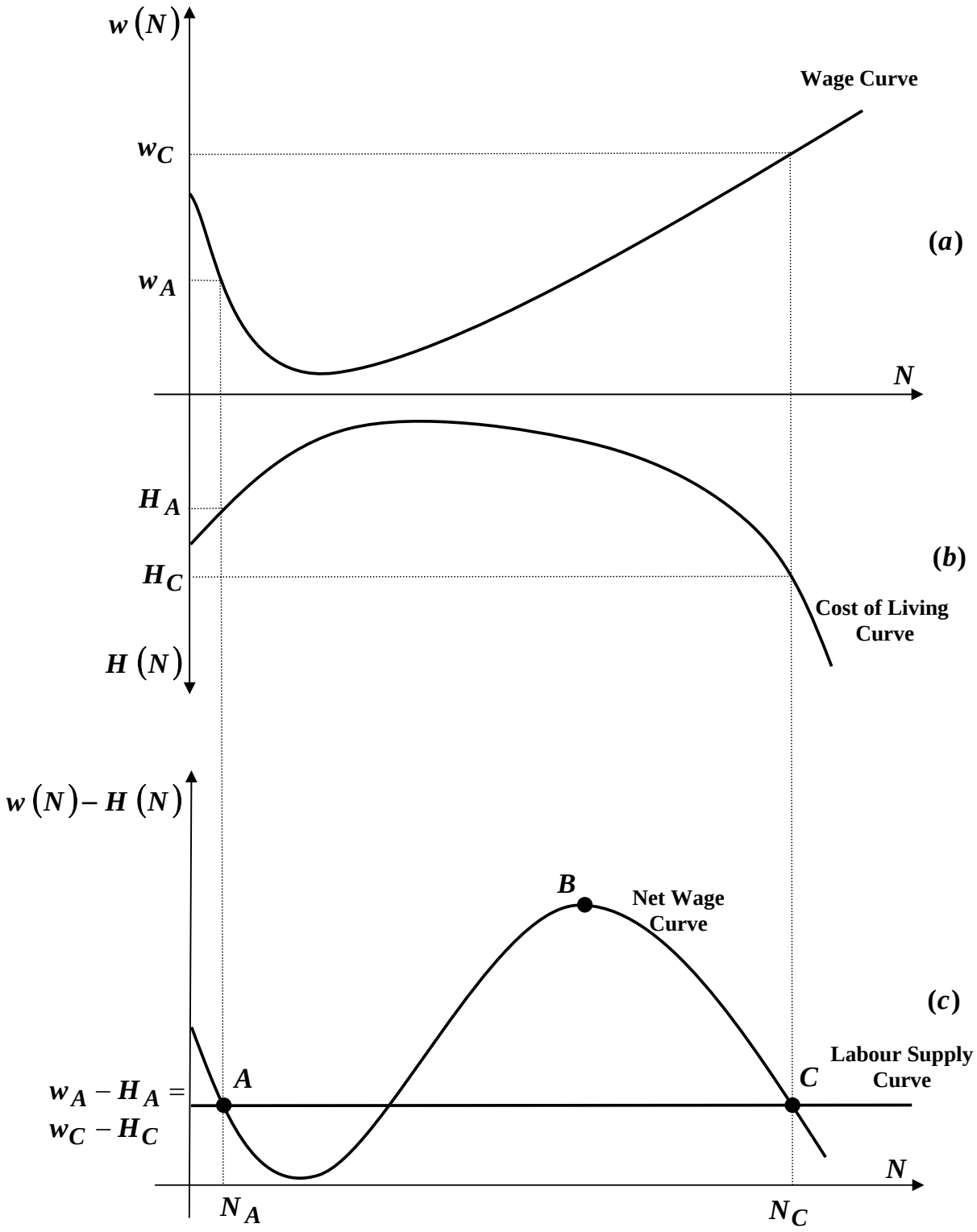
Assume that market access effects dominate agglomeration effects in small markets while the reverse holds in large markets. This implies a wage curve that first slopes downwards and then upwards. This wage curve is represented in figure 4 (a). We concentrate on this case because it has more interesting implications than its opposite. To defend it, one could also argue that negative market crowding effects can be very strong at the margin in a very small market while they are going to be much milder if many firms are already operating in a market. Furthermore, it could also be argued that a minimum city size is needed for agglomerations economies to take place.

It is important to understand that transport costs affect not only the wage curve (through the production of goods) but also the cost of living curve (through their consumption). A small isolated city may face very low housing and commuting costs. However, consumption goods can be very expensive because most of them are produced elsewhere and need to be shipped to this city at a very high cost. As the city grows and produces more manufacturing, the price of goods declines since a smaller proportion of them need to be imported. On the other hand, other components of the cost of living such as housing and commuting increase with city size. Again, we have forces pushing in opposite directions. It seems reasonable to assume that higher housing and commuting costs eventually dominate when cities grow very large. This suggests that the cost of living first decreases and then increases as cities grow. The cost of living curve corresponding to this situation is represented in figure 4 (b).

Subtracting the cost of living from the wage implies that the net wage first decreases, then increases, before decreasing again with city size.[23] In figure 4 (c), the net wage curve and the labour supply curve intersect three times. Ignoring the unstable equilibrium, we are left with two stable equilibria. Cities are either very small, at point A, or much bigger, at point C. The optimal city size (point B) is somewhere in between. Compared to figures 1 and 3, the novelty in figure 4 is the existence of small cities in A whose growth is limited by strong local crowding on the product market and insufficient agglomeration effects.

[23]This requires the wage curve to decrease faster initially than the cost of living as the city grows. We expect this to happen because manufacturing wages are expected to decline proportionately to the price of local manufacturing goods whereas the local cost of living is expected to decrease less than proportionately since the prices of the agricultural good and imported manufacturing are unchanged while the other components of costs of living increase.

Figure 4. Internal market access



This crowding is in turn caused by high transport costs and the difficulty for the cities to export their output.

The representation in figure 4 is important because it provides a strong rationale for the co-existence in many developing countries of small stagnant cities and large primate cities. High costs of trade between cities may also explain why cities in developing countries may not be as fully specialised as in developed countries. This is because urban specialisation makes little sense when the costs of inter-city trade are very high.

The literature offers strong empirical support regarding the importance of market access for cities in developing countries. Using two different approaches, Lall *et al.* (2004*b*) and Lall *et al.* (2003) underscore the importance of market access in India. Strong effects of market access are also found in Brazil (Lall, Funderburg, and Yepes, 2004*a*; da Mata *et al.*, 2007) and Indonesia (Deichmann *et al.*, 2005; Amiti and Cameron, 2007). This within-country evidence is complemented by a large literature that looks at the importance of market access at the country level (Redding and Venables, 2004; Head and Mayer, 2004). Regarding the shape of the cost of living curve, the evidence is much thinner. In large part, this is due to the general paucity of research on this issue. However, the paper that currently defines the frontier on the topic (Timmins, 2006) finds strong evidence for Brazilian cities of non-linear cost of living curves taking the shape hypothesised above.

Let us now turn to policy implications. Improving market access implies better access to other markets but it is also synonymous with a loss of protection for local firms. Depending on which effect dominates, the wage curve can shift upwards or downwards. Better market access for small isolated cities also implies a less steeply decreasing wage curve so that a flatter wage curve (at least in its early part) is expected. With better access, we also expect a lower cost of living. The cost of living curve could flatten as well. On balance, for small cities better market access implies a flatter and possibly higher net wage curve. In turn, this implies that the small city equilibrium at point A should shift to the right (city growth) or even disappear entirely keeping point C as the only stable equilibrium. With broad-based gains from better market access, we can also expect a higher labour supply curve through general equilibrium effects. As a result of a higher labour supply curve, the equilibrium size of large cities would decrease. The final outcome could be smaller big cities but a higher number of them.[24]

In practice, market access is improved by two sets of policies. The first is about building and developing roads and other transport infrastructure such as airports or high-speed train lines. The second is about removing impediments to trade across regions, from

[24]Even assuming a reduction in urban primacy, sustaining much larger cities may only come from population growth, rural-urban migration, or the disappearance of some smaller cities. Hence having larger cities may be synonymous with a smaller equilibrium number of viable cities. It may not be as much of a problem as it is in Europe where population growth is minimal and the urbanisation process completed but it is still worth a thought.

administrative hurdles to cartelised distribution networks. A number of caveats must be kept in mind. First, much remains to be done because most existing specifications in the empirical literature are not derived directly from theory (Head and Mayer, 2004, 2006). Put differently, the importance of market access is established but it is still unclear how it precisely works. Next, the development of road networks may have perverse effects. Linking small cities to large economic centres increases the market potential of the former but it may increase that of the latter even more and thus reinforce primacy instead of reducing it.[25] The US experience nonetheless suggests that there can be large productivity gains associated with the development of an integrated transport network (Fernald, 1999). Finally, improving market access may also have some effects at a geographical scale greater than cities. A key prediction of modern regional economics is that lower levels of transport costs can lead first to increased regional agglomeration, and then possibly decreased regional agglomeration for even lower levels of transport costs (Fujita *et al.*, 1999; Combes *et al.*, 2008*b*). However, with better transport infrastructure, there is a possibility of a group of winning cities in core regions and a group of cities left behind in the periphery.[26] One could think of coastal Chinese cities vs. hinterland cities or high-plateau cities in Colombia vs. cities on the Colombian Caribbean coast, etc.

In summary, urban primacy is often attributed to a dysfunctional political economy leading to primate city favouritism. There is a lot of empirical support for this. A complementary explanation points at high internal trade costs leading to either large or small cities. A lot of the evidence is consistent with this explanation as well. In both cases a reduction in urban primacy is desirable. Doing it through a reduction in primate city favouritism may be effective but is hard to implement politically. Improving market access for isolated cities may be politically easier to achieve but the precise effects of better access are more difficult to predict since improved access may reinforce primacy. Furthermore, as made clear below the exact policy prescription is likely to be country-specific. For instance, increased political decentralisation may be effective to reduce primate city favouritism in some countries but not in others.

Migration and dual labour markets

The framework developed above assigns a positive (and equalising) role to internal migrations and labour mobility. This is in contrast with some of the academic literature and much of the policy reality in developing countries. From internal passports in China and the 'nativist' policies of Indian States to the resettlements policies carried on in Africa

[25] These theoretical ambiguities about the effects of transport costs on agglomeration are analysed in depth by Baldwin *et al.* (2004).

[26] See Fujita and Mori (2005) for a systematic study of how transport costs can affect cities when regions are explicitly considered.

and Latin America, there is a strong bias against free labour mobility in many developing countries.

As shown above, restricting the movement of labour is not the right answer if cities become too big as in the framework exposed above. Another justification for anti-mobility policies rests on the existence of dual labour markets. The argument was first developed by Harris and Todaro (1970) and has been extremely influential in policy circles. Theoretically, it works as follows. There is a formal sector with a fixed number of urban jobs that pay a high wage, w_A, in figure 5 (a). In rural areas, workers get lower earnings represented by the labour supply curve in figure 5 (c). This earnings gap between the rural and the formal urban sector causes workers to move to the city.

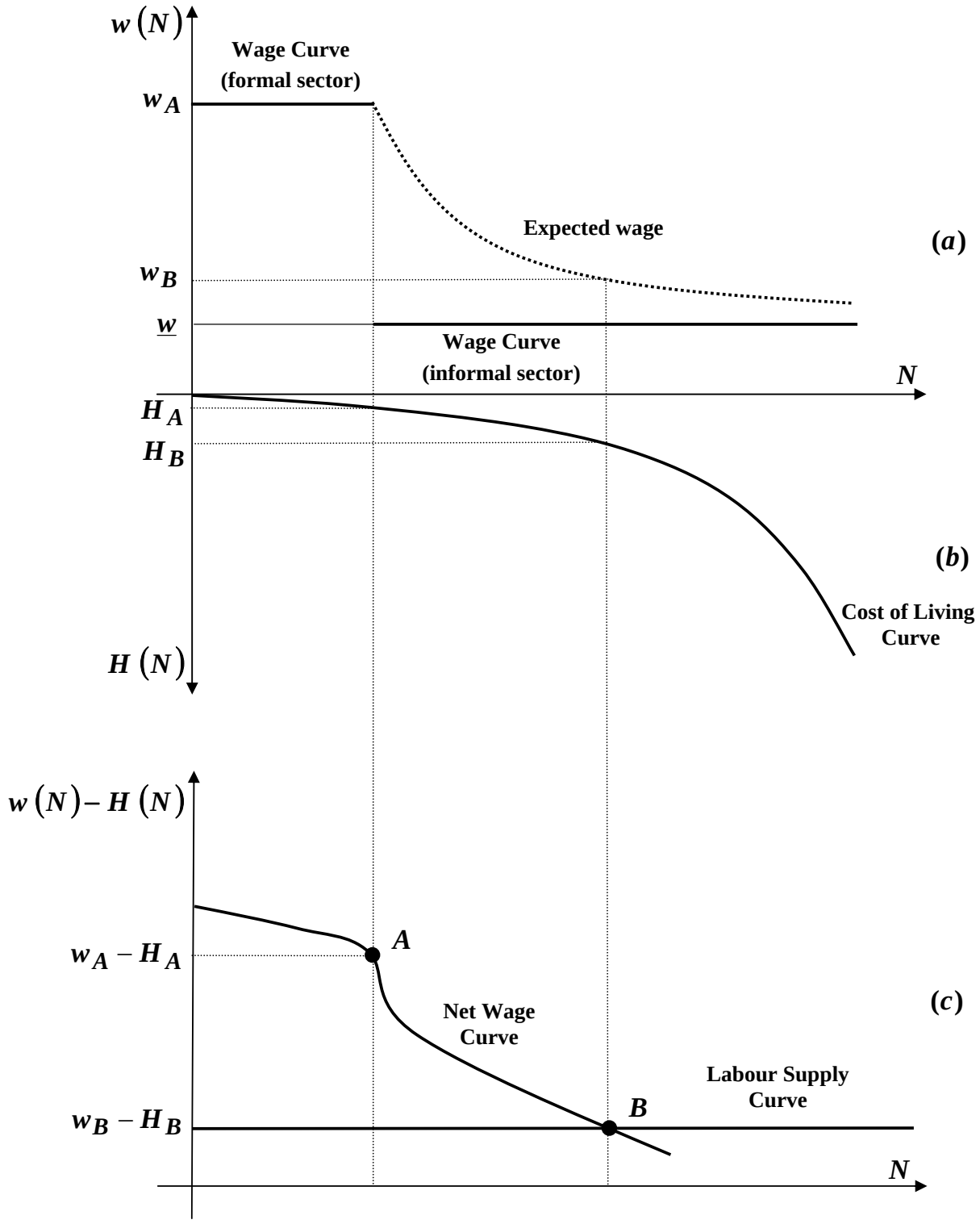
Should there be only so many migrants to the city as there are jobs in the formal sector, the city would end up at the social optimum, point A in figure 5 (c). However, when the city is at point A it cannot be in equilibrium because the net wage is above that received in rural areas. If there are more workers than available jobs in the formal sector, the model assumes that jobs are randomly attributed to the city residents. The lucky ones get a job in the formal sector while the unlucky ones get a job in the informal sector of the same city. This informal sector offers a low wage, $\underline{w}$. In this case workers keep moving to the city until the *expected* wage they receive minus the cost of living (i.e., their expected net wage) intersects with the labour supply curve.[27] This occurs at point B. It is easy to see that this equilibrium entails cities that are too large. The main difference with the baseline case explored above is of course not that cities are too large – that was the case in the baseline as well – but that it makes sense to curtail entry into the city.

Although appealing, the Harris-Todaro argument can be criticised on a variety of grounds (see Lall *et al.*, 2006, for an in-depth analysis and empirical references). It must first be said that workers end up in the informal sector because of wage rigidities in the formal sector. Trying to solve a problem that occurs in the labour market by restricting the mobility of workers is not the most direct solution and is likely to have a number of unwanted side effects.[28] The stark assumptions of the Harris-Todaro model also bias it towards generating over-migration to the cities. For instance, workers are risk-neutral and formal sector job are randomly allocated. However, workers are arguably risk-averse and know that formal sector jobs are not randomly allocated so that only those with high chances of getting one are expected to move. The fact that the formal and informal sectors appear quite segmented in most developing countries only reinforces the point.

[27] The graph follows the approach of Brueckner and Zenou (1999) who explicitly consider a land market. This already reduces the tendency of cities to become too big compared to the most basic versions of the Harris Todaro model.

[28] This rigidity is also partly attributable to a very large and spatially concentrated public sector. The same argument nonetheless applies. Restricting urbanisation is not the way to deal with a dysfunctional public sector.

Figure 5. Harris-Todaro migrations



Furthermore, the downward sloping wage and net wage curves predicted by figure 5 do not receive any empirical support as made clear above. All this suggests that the main argument used to restrict labour mobility is relatively weak.

To go beyond a mere rebuttal of Harris and Todaro (1970), we need ask ourselves why restrictions on labour mobility are so widespread in developing countries. A first possibility is to point at an overzealous application of Harris and Todaro (1970) by policy makers. In such a case, policies can change after showing the weakness of their underpinnings. Instead, restrictions on labour mobility may be part of a political-economy equilibrium. In this latter case, better policies would then be much more difficult to implement. We need to know more about this issue to understand the nature of the challenge for labour mobility and how it may be overcome.

Dual housing markets

The last key feature of cities in developing countries is the existence of a dual housing sector with a division between the formal sector and squatter settlements (also referred to as slums, invasions, shanty-towns, etc). In some large developing country cities, more than half the population live in squatter settlements and face very poor public services provision (if at all), insalubrious living conditions, and a number of constraints associated with the precariousness of their housing.

Squatter settlements are often associated with the idea of low-cost and low-quality housing. If it was only this, they would be simply a reflection of the general poverty of some urban dwellers opting out of the formal housing sector because they cannot afford it. Policy decisions regarding what to do with squatter settlements would mostly be choices about how much redistribution to do (or not to do) and whether it is best to do this redistribution through subsidised housing and public services or by other means.

These issues are important but there is more to squatter settlements than this. First, it has been widely argued that poorly defined or poorly enforced property rights over urban land, which make squatter settlements possible, could also affect a wide range of other economic outcomes. De Soto (2000) argues that a lack of effective, formal property titles prevents residents of squatter settlements from using their housing as collateral and is thus a major barrier to enterprise development. Although the evidence about the existence of these financial constraints is disputed, Di Tella, Galliani, and Schargrodsky (2007) find that a lack of titles has important effects on the beliefs of people and thus their economic behaviour and Field (2007) finds that it also matters for female labour supply.

Next, squatter settlements may be the outcome of policy distortions. Henderson (2007) argues that binding minimum size lots is responsible for the growth of squatter

settlements in Brazilian cities. One could also mention the prevalence of rent controls (Malpezzi, 1999) that limit the expansion of the rental market, etc.

Finally, once the absence of public services or their very poor quality is taken into account, squatter settlement may not be so cheap. For instance, water in slums often needs to be bought at a very high price from local water distributors. Without titles, squatters must also often pay a steep price for some form of protection, etc.

Taking the last two ideas seriously about exclusionary zoning by the formal sector and the relatively high costs of squatter settlements, it is possible to expand our theoretical framework to gain some insights about dual housing markets. In figure 6 (a), we assume a standard upward-sloping wage curve that applies to all city residents.[29] In figure 6 (b), there are three cost of living curves. The dotted curve represents the cost of living in the formal sector in absence of exclusionary zoning. Exclusionary zoning (e.g., minimum lot size in Brazil) raises the cost of living in the formal sector, yielding the solid cost of living curve of the graph. The alternative to the formal sector is a squatter settlement. The cost of living in a squatter settlement is represented by the dashed line on figure 6 (b). Because of the high cost of the substitutes for missing public services and other expenses, the cost of living in squatter settlements is higher than in the formal sector in absence of exclusionary zoning. The cost of living in squatter settlements is also higher than the cost of living in the formal housing sector with exclusionary zoning for small cities but lower for large cities. The main justification for this is that the higher cost of 'public' services in squatter settlements is rather insensitive to city size while the economy in land rent made by squatting is more likely to increase with city size.

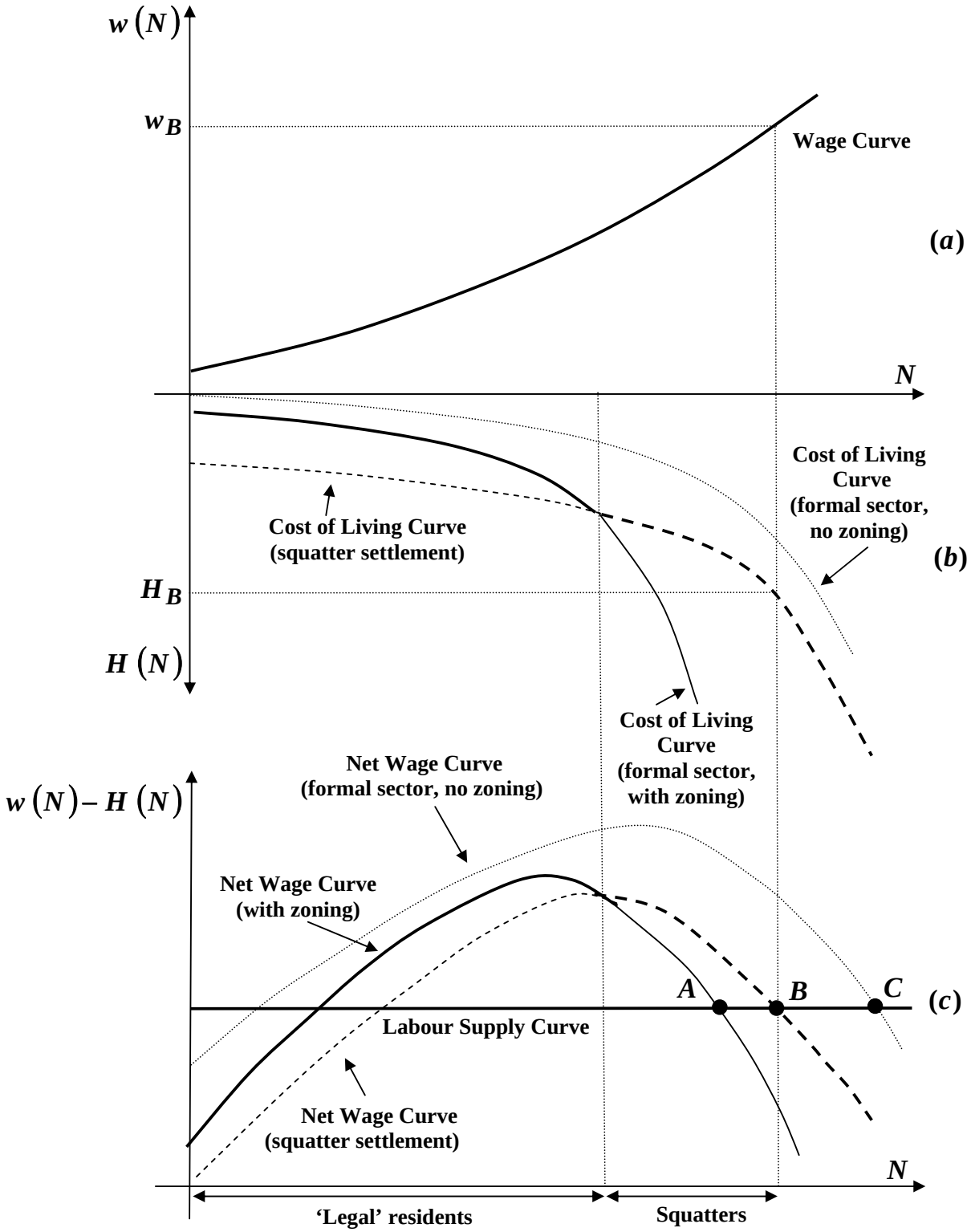
In absence of exclusionary zoning, the cost of living is always lower in the formal sector and no one would choose to live in a squatter settlement. The net wage curve corresponding to this situation is the dotted thin curve figure 6 (c). The city equilibrium is reached at point C.

With exclusionary zoning, it is cheaper to live in the formal sector than in squatter settlements when the city is small but it gets more expensive when the city grows. The thick line in figure 6 (b) then represents the minimum cost of living under exclusionary zoning. This thick line is solid (i.e., formal sector) for small cities and dashed for larger cities since they expand though squatter settlements.

Under exclusionary zoning, the net wage curve is the maximum of the net wage offered

[29] In many cities with squatter settlements, a significant proportion of slum dwellers have a job in the formal (production) sector. Although the duality of the labour market is related to that of the housing market, the two need to be distinguished. As a first-order approximation, it is nonetheless reasonable to assume that all workers benefit from agglomeration effects. A more refined version of the graph would take into account the fact that slums are often located far from the main places of work and are poorly served by public transport. Hence, slum dwellers are less likely to benefit from agglomeration effects. We acknowledge this but this is not core to the argument here which focuses on the housing market.

Figure 6. Dual housing sector



either by the formal housing sector or by squatter settlements. It is represented by the thick 'continuous and dashed' curve in figure 6 (c). The equilibrium for the city is at point B. Below the X-axis, we can read the city population that resides in the formal sector and in squatter settlements.

This analysis suggests a number of policy implications. Imposing regulatory constraints in the formal housing sector may reduce the equilibrium size of the city but a good fraction of this reduction is crowded out by the growth of squatter settlements. The equilibrium for the city is at point B, and not at point A, as originally intended. Removing unnecessary constraints in the formal housing sector is socially desirable since it lowers the cost of living, hence raises the net wage curve and eliminates squatter settlements. Furthermore, improving the situation in one city only is not enough to raise net wages since a higher net wage curve may keep hitting the same labour supply curve, only at a larger city population. Once more, general equilibrium effects matter.

'Titling' policies are also desirable because, as argued above, poorly defined property rights have a range of other negative side-effects. After solving disputes over land ownership and the financing of the titles handed-out, the main issues with titling policies are, first, how to avoid further preemptive invasions driven by the expectation of a future title and, second, how to make sure that there is a local tax counterpart to legalised titles. In some respect, the problems in dealing with illegal settlements are the same as those of urban favouritism. There are many dimensions associated with this phenomenon and it is not clear yet which are those that matter most empirically. As with urban favouritism, there is also a political economy of illegal settlements with vested interests that benefit from slums, either directly by charging their residents, or indirectly by providing expensive substitutes for missing public services. These vested interests often pose formidable challenges.

4. Does all this matter for growth?

Big effects of urbanisation on growth?

So far, the discussion has focused on the efficiency gains associated with cities. It could be argued that such gains are mostly static in nature and that the urban policies discussed above are, to a large extent, one-time improvements. This is true, although it should be kept in mind that some of the improvements discussed above might be large and take place over relatively long periods of time. For instance, Fernald (1999) estimates that the US interstate system generated about one point of annual GDP growth during nearly 20 years.

With this in mind, it is now worth asking whether cities and urbanisation can affect the long-run rate of economic growth. This question is fraught with difficulties and the empirical growth literature has not been particularly successful at isolating the causes of long-run growth (see Durlauf, Johnson, and Temple, 2005, for a recent review). Sadly, this literature has also barely paid any attention to cities and urbanisation as possible causes of growth.

There is indeed only one study, Henderson (2003), that uses a reasonable cross-section of countries and sound econometric methods to look at the dynamic aggregate effects of cities and urbanisation.[30] The two main conclusions of that study are the following. First, urbanisation per se does not affect economic growth. Second, urban primacy has large effects on economic growth. The first conclusion is rather unsurprising and confirms a broad consensus. Urbanisation is a benign transition that, to a large extent, follows the process of development but does not profoundly affect it. The second conclusion is more provocative. Henderson (2003) finds that an increase in urban primacy by one standard deviation (or 15%) from the mean (of 31% of the urban population living in the largest city) reduces the rate of GDP growth by about 1.5% per year. These are large effects. They are also interesting from a policy perspective since urban primacy can evolve relatively quickly.

The question is of course how seriously we should take these estimates. The main issue in any such investigation regards the direction of causality. A strong negative statistical association between urban primacy and growth may not be surprising. A strong causal effect is more so. To deal with causal issues, one ideally needs to find some good instruments for urban primacy, i.e., variables that determine urban primacy but are not otherwise correlated with economic growth. In this case, one can use the variation in urban primacy caused by these exogenous variables to assess the causal effect of primacy on growth.

Unfortunately, it is hard to think of any variable that would determine primacy and be otherwise uncorrelated with economic growth. In particular, the key determinants of urban primacy, political variables, are expected to have a strong independent effect on economic growth. Instead, Henderson (2003) proceeds as follows. He takes the first difference of all his variables to get rid of any permanent country effects that would be correlated with both economic growth and urban primacy. Then, using a GMM estimation technique, he instruments changes in urban primacy by lagged primacy levels from 10 or 15 years before. This estimation technique yields large effects of urban primacy on economic growth.

[30] Using the same data, Bertinelli and Strobl (2003) replicate and confirm some of the findings of Henderson (2003) using non-parametric techniques.

How convincing is it? When it comes to urban primacy, past levels are good predictors of current changes. Put differently, past levels are relevant instruments for contemporary changes in primacy. That past levels of primacy are otherwise uncorrelated with changes in the rate of growth is much harder to show. To make the case for the exogeneity of his instruments, Henderson (2003) shows that over-identification tests are easily passed. This suggests that if the instruments used in the regression are invalid, it can only be because they bias the results in the same way. It can also be argued that after first-differencing, one controls for all static explanations whereby institutions (a possible missing variable) would explain both long-run growth and primacy. Unfortunately, this does not rule out more dynamic explanations where institutions and urban primacy could interact at higher frequency. Such a story nonetheless remains to be written and empirically assessed.

All this puts us in a very uncomfortable situation. On the one hand, Henderson's (2003) work uses the best available data and methods so that we should certainly not dismiss his findings. On the other hand, his findings are not corroborated by other findings (for lack of other work on this issue). To deal with this dilemma, the rest of this paper takes these findings seriously but remains cautious in its policy recommendations and only advocates policies that are not going to be harmful should the findings of Henderson (2003) turn out to be spurious.

Explaining large dynamic effects of primacy

The simplest explanation behind potentially large dynamic effects of urban primacy on growth relies on the idea developed above that primacy makes production less efficient. With less output, less is being saved. In turn, less accumulation can lead to sluggish growth. This explanation is problematic for two reasons. The first is that the *level* of output does not seem to have a large effect on the *rate* of accumulation of capital and knowledge (Caselli, 2005; Durlauf *et al.*, 2005). The other problem is that *any* one-time change in static efficiency should also have large growth effects. There is not much evidence of such effects in the empirical growth literature.

Another line of explanation is that the spatial system could be understood as the spatial imprint of economic growth. As an economy develops, so does the way it is organised spatially. Duranton and Puga (2001) argue that modern systems of cities experience a division between cities where innovation takes place ('nursery cities' with a very diverse production structure) and cities that are more specialised into the production of one particular set of goods. In developed economies, the last 50 years have also seen a growing separation between business centres, which host headquarters and business services, and production cities, which host production plants (Duranton and Puga, 2005). Finally, Duranton (2007) shows that, within countries, sectors tend to change location quite fast

following technological change. If the urban landscape is a reflection of economic growth, the corollary may be that constraining the geography of cities hinders growth. More generally, preventing urban dispersion by favouring the primate city or by preventing the development of secondary cities is going to entail costs.[31] For instance, favouring a primate city can prevent the efficient division between nursery cities and production cities and thus slow down innovation. In the same vein, favouritism can also prevent the development of new ideas and new productions in secondary centres. This type of claim is of course hard to evaluate empirically. There is good evidence from Korea (Henderson, 2002*b*, 2005) that mature manufacturing quickly moved out of Seoul and relocated to secondary cities. Brazil appears to follow a similar path, albeit more slowly (da Mata, Deichmann, Henderson, Lall, and Wang, 2005). In many other countries, this process of urban change appears to be even slower, if it takes place at all.

A number of policy implications are associated with this type of explanation. First, heavy-handed policies to relocate economic activity away from primate cities are unlikely to be successful because of the difficulty of replicating the subtle alchemy of nursery cities. On the other hand, policies that reduce primate city favouritism are of course desirable. Removing obstacles to the relocation of production plants in secondary cities is also desirable. Finally, it should be added that production will move away from primate cities only if it can access their markets back from wherever it chooses to relocate. This suggests that transport and infrastructure policies are also important in this respect.

A third explanation is that cities may play a direct role in the accumulation process and in innovation (and not only in production as emphasised by our framework). Back to Jacobs (1969) and more recently Lucas (1988), there is a strong tradition in urban economics that views cities as engines of growth (e.g., Eaton and Eckstein, 1997; Black and Henderson, 1999; Glaeser, 1999; Rossi-Hansberg and Wright, 2007).

The model of Eaton and Eckstein (1997) assumes that individual human capital accumulation is driven by three factors: the fraction of time spent learning, the current level of human capital, and the 'knowledge base' of the city. In turn, the knowledge base of the city is taken to be the sum of the human capital in the city, possibly adding the discounted sum of human capital of other cities. Because of these accumulation effects, larger and more educated cities are desirable to foster growth. However, there are also costs to city size with respect to innovation and factor accumulation. For instance, the physical crowding of a city is time-consuming and thus implies that there is less available time to learn and accumulate human capital. The opportunity cost of direct interactions in the city also increases as it grows in size. In other words, there is an optimal city size that maximises

[31] This process of differentiation between nursery and headquarter cities (which need not be the same) and production cities is strongly related to the older idea of a Kuznets curve for urban primacy, first argued by Williamson (1965).

the rate of growth. This optimal dynamic size need not be the same as the static optimal size. It can be bigger or smaller. Unfortunately, serious evidence on this issue is lacking.[32]

Beyond their immediate intuitive appeal, these models of urban growth are useful to rationalise a number of important stylised facts from the concentration of human capital in large cities to the existence of spatial disparities in productivity. These models assign a key role to local dynamic externalities, and in particular dynamic human capital externalities. But do we know whether such externalities exist and, if yes, how much they matter quantitatively? There is a large literature in developed countries about human capital externalities in cities. Although not undisputed, the current consensus is that human capital externalities in cities exist and might be large. Estimates of social returns to education being of the same magnitude as its private returns are not uncommon.[33] It is also noteworthy that dynamic externalities are taken seriously in macroeconomics to provide a quantitative explanation to a number of puzzles about economic growth (Klenow and Rodríguez-Clare, 2005). On the other hand, it should be kept in mind that externalities are notoriously hard to identify empirically and the distinction between static and dynamic externalities is still far from settled.

A key limitation of the urban growth literature is that it tends to view each city as an 'island of growth'. More precisely, it typically assumes that each city can generate economic growth by itself and for itself.[34] Although very little is known about this, especially in a development context, it is fundamental to understand how knowledge flows across places within (developing) countries and between each country and the rest of the world.[35] Of course, this issue of knowledge flows goes well beyond the scope of this paper since it raises issues such as the protection of intellectual property rights and how to set up the right incentives for the creation and diffusion of knowledge independently of the spatial context.

To open the black-box of knowledge flows within countries, a first possibility is to assume that knowledge is embedded in people and is acquired by direct contact with 'those that know'. There are two facets to this issue. First, there is a compelling argument that cities are places where workers learn. This was articulated first by Glaeser (1999) and later adapted to a development context by Lucas (2004). In particular, the latter

[32]See Bertinelli and Black (2004) for further discussion of these issues. The fact that many young professionals accept low net wages in New York or London might suggest that the dynamically optimal size of cities is larger than the statically optimal size.

[33]This literature is surveyed in Moretti (2004). See also Duranton (2006) for a less technical introduction to the topic. While most of the evidence is about Europe and North America, Conley, Flyer, and Tsiang (2003) find evidence of localised human capital externalities in Malaysia.

[34]It is true that Eaton and Eckstein (1997) consider the possibility of the knowledge base of one city to depend on that of other cities. However, they model this in a ad-hoc manner by setting an exogenous spillover function across cities.

[35]About knowledge flows across countries, see Keller (2004).

shows that rural workers may optimally migrate to cities and then spend some time accumulating human capital before becoming more productive. This suggests that what is interpreted negatively as urban unemployment of rural migrants in a Harris and Todaro (1970) framework may actually be a time of adjustment and learning.

Existing empirical findings about learning in cities are very suggestive, though mostly limited to the US. Glaeser and Maré (2001) show that there is an urban wage premium, which workers retain when they move back to smaller cities or rural areas. Among a number of papers, Peri (2002) and Wheeler (2006) document that wage growth is stronger in cities, particularly for young educated workers. This could be due to the self-selection of workers with fast career progressions in cities for reasons unrelated to learning. However, Freedman (2007) shows that this type of result is found even when controlling for the fact that some workers may experience higher wage growth independently of their location. Although this has not yet been investigated in a developing country context, there are thus strong reasons to think that cities bolster workers' learning. Restricting migration to cities may then have negative dynamic consequences.

The counterpart to the learning-in-cities argument is that flows of people are also flows of knowledge. Knowledge gets disseminated by people. This argument has been modelled in the context of the mobility of employees between firms (Combes and Duranton, 2006; Franco and Filson, 2006) but not yet between cities. Empirically, Møen (2005) and Freedman (2007) show that technological progress is indeed associated with the movement of skilled workers between firms. Job-hopping appears to be beneficiary to the job-hoppers and to their industry, if not to their employers. In addition, Almeida and Kogut (1999) show that long distance flows of knowledge, as tracked though patent citations in the US semiconductor industry, coincide with the movement of star scientists across firms in different cities. Interestingly, Agrawal, Cockburn, and McHale (2006) also show that the scientists who leave a city keep being cited there. They are gone but not forgotten.

To the extent that these findings about highly skilled US workers also apply to highly skilled workers in developing countries, we can draw a number of policy conclusions. First, the general working of the labour market and more specifically the covenants that restrict labour mobility can play an important role to hinder the diffusion of knowledge within and between cities. A lack of labour mobility, especially in the most skilled segments of the labour market, between the main city and secondary cities may be an important contributor to both urban primacy and the backwardness of secondary cities. With limited labour mobility between cities, nearly all skilled labour may go to the main city and stay there. The main city then becomes an island of more advanced knowledge with a much higher wage curve. As highly skilled workers remain in the primate city, their knowledge does not percolate to other cities. These other cities then stay behind

technologically and remain small because of their low wage curve. A key issue is that, even in absence of formal impediments to labour mobility, this situation can remain since the technological backwardness of small cities may provide little incentive for skilled workers to relocate there.

If the two-way mobility of skilled labour between cities seems important to foster the diffusion of technologies across places, it may not be the only channel. Although the existing evidence mostly concerns countries and not regions within countries, there is a good case to be made that more trade in goods is also associated with higher growth and convergence across places. In a cross-country setting, Wacziarg and Welch (2003) show that increased openness has large positive effects on growth and investment. Alcalá and Ciccone (2004) also show that the positive growth effects of trade work through total factor productivity. The effects found by Alcalá and Ciccone (2004) and much of the prior literature in a cross-country setting are relatively large. For instance, moving from the twentieth percentile of openness to the median raises productivity by 160% according to Alcalá and Ciccone (2004). With no evidence of weaker effects when openness is already high, this suggests that there are potentially large dynamic gains from removing impediments to trade within developing countries.[36]

Finally, there is very strong evidence that productivity growth is linked to the process of creation and destruction at the firm level (Davis, Haltiwanger, and Schuh, 1996; Foster, Haltiwanger, and Krizan, 2001; Bartelsman, Haltiwanger, and Scarpetta, 2004). In particular resources need to flow from less to more productive firms and allow new entrants to rise and challenge incumbents. An analysis of this process of reallocation would, of course, go much beyond the scope of this paper. Nonetheless it is important to remember that in developed countries there is a strong spatial dimension to this process as industries tend to change location when their technology evolves (Duranton, 2007). An important conclusion here is that hindering the movement of factors across firms, including across firms located in different cities, may have large dynamic costs.

5. Policy conclusions

It is now time to summarise our policy conclusions and consider a number of practical issues regarding their implementation.

The first issue is whether a growth agenda leads to urban policy recommendations that differ from those of a traditional agenda, usually more concerned with urban efficiency. The urban efficiency agenda is explored in the first part of the paper. Its main recommendations are the following: eliminate primate city favouritism; improve urban efficiency

[36]These gains could be all the bigger since after physical impediments to trade are removed, trade is much easier to conduct within countries than between.

so as to lower the cost of living curve by dealing with urban crowding, and public good provision; solve the biases that lead to squatter settlements with a reasonable titling policy and urban deregulation; improve market access between cities by developing transport infrastructure and lowering impediments to trade; and do not discourage internal migrations which foster an efficient allocation of the population and have an equalising effect across places.

By underscoring the need for better public service delivery and the importance of housing and commuting issues, this set of recommendations is consistent with some of the objectives of many existing urban policies. The main difference is that our baseline framework also emphasises labour mobility. This emphasis is strongly at odds with existing urban policies that often seek, on the contrary, to reduce labour mobility and more generally to promote some form of stability. Another novelty of the static framework explored above is to underscore the possible effects that technological, institutional, or policy-driven changes can have on cities. The urban equilibrium is determined by the interplay of the wage, cost of living, and labour supply curves. In turn, these curves are determined by a wide array of forces, which can all affect cities indirectly.

Taking a more dynamic perspective does not fundamentally alter the recommendations of more static approaches. It leads us to put even more emphasis on the mobility of people and goods across places. This emphasis on mobility and flexibility in factor allocation and reallocation should also arguably be part of any modern growth agenda. Hence even though, at some fine level of detail, static and dynamic approaches to urban policy might conflict (e.g., optimal city size may not be the same to maximise static vs. dynamic efficiency), these divergences are minor from a practical perspective. It is also important to note that an urban perspective on economic growth does not appear to conflict with any broader growth agenda.

This being said, implementing a broad-ranging urban agenda aimed at bolstering economic growth raises a number of problems. The first is that such agenda is rather demanding since it includes raising the efficiency of public good provision, lowering barriers to mobility, improving market access to allow secondary cities to develop, etc. The second difficulty is the political economy of many of these issues is often a formidable obstacle to change. Hence, politics and other more mundane feasibility constraints, such as the limited capabilities of many governments, require establishing priorities. On the other hand, the framework used above shows clearly that cities operate in second-best world where fixing one problem may not result in any tangible improvement locally. Hence we face a policy dilemma where doing all at once may not be possible but a step-by-step approach may not be effective.

Furthermore, growth agendas often identify a number of 'growth drivers' that need to be fostered. Rather than drivers, it may be more fruitful in practice to think about

constraints and bottlenecks to be removed. In this respect, the theoretical framework developed above can be useful to identify constraints to harmonious urban development.[37] Since constraints and bottlenecks are likely to differ across countries, so will the diagnostic and, ultimately, the urban strategy. The main caveat with this diagnostic approach is that static constraints to urban development such as a grid-locked city are for all to see, while dynamic constraints are much more difficult to identify.

The final question relates to who should be in charge of implementing any 'cities and growth' agenda. The emphasis given here to the mobility of goods and factors between cities suggests that central governments should have a prominent role in promoting labour mobility, developing infrastructure, and removing impediments to internal trade. However, cities have also an important part to play to improve the life of their residents and minimise their cost of living. This division of labour between central and local governments is nonetheless unlikely to remain free of tensions. First, there is a fundamental asymmetry between primate cities and secondary cities. No secondary city can alone have an effect on the entire urban system whereas primate cities do. There is also considerable heterogeneity in the capabilities of secondary cities to design and implement local policies that would be consistent with a national growth agenda.

[37] Although this would go beyond the scope of this paper, the framework used above and its extensions could be further developed as a diagnostic tool in the spirit of Hausmann, Rodrik, and Velasco (2005).

References

Ades, Alberto F. and Edward L. Glaeser. 1995. Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* 110(1):195–227.

Agrawal, Ajay, Iain Cockburn, and John McHale. 2006. Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography* 6(5):571–591.

Alcalá, Francisco and Antonio Ciccone. 2004. Trade and productivity. *Quaterly Journal of Economics* 119(2):613–646.

Almeida, Paul and Bruce Kogut. 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management Science* 45(7):195–227.

Amiti, Mary and Lisa Cameron. 2007. Economic geography and wages. *Review of Economics and Statistics* 89(1):15–29.

Aten, Bettina and Alan Heston. 2005. Regional output differences in international perspective. In Ravi Kanbur and Anthony J. Venables (eds.) *Spatial Inequality and Development*. New York: Oxford University Press.

Au, Chun-Chung and J. Vernon Henderson. 2006*a*. How migration restrictions limit agglomeration and productivity in China. *Journal of Development Economics* 80(2):350–388.

Au, Chun-Chung and J. Vernon Henderson. 2006*b*. Are Chinese cities too small? *Review of Economic Studies* 73(3):549–576.

Baldwin, Richard E., Rikard Forslid, Philippe Martin, Gianmarco I. P. Ottaviano, and Frederic Robert-Nicoud. 2004. *Economic Geography and Public Policy*. New Jersey: Princeton University Press.

Barrios, Salvador, Luisito Bertinelli, and Eric Strobl. 2006. Climatic change and rural-urban migration: The case of sub-Saharan Africa. *Journal of Urban Economics* 60(3):357–371.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta. 2004. Microeconomic evidence of creative destruction in industrial and developing countries. Processed, University of Maryland.

Bertinelli, Luisito and Duncan Black. 2004. Urbanization and growth. *Journal of Urban Economics* 56(1):80–96.

Bertinelli, Luisito and Eric Strobl. 2003. Urbanization, urban concentration and economic growth in developing countries. Research Paper 03/14, CREDIT, University of Nottingham.

Black, Duncan and J. Vernon Henderson. 1999. A theory of urban growth. *Journal of Political Economy* 107(2):252–284.

Brueckner, Jan K. 1990. Analyzing Third World urbanization: A model with empirical evidence. *Economic Development and Cultural Change* 38(3):587–610.

Brueckner, Jan K. and Yves Zenou. 1999. Harris-Todaro models with a land market. *Regional Science and Urban Economics* 29(3):317–339.

Caselli, Francesco. 2005. Accounting for cross-country income differences. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1A. Amsterdam: North-Holland, 680–741.

Combes, Pierre-Philippe and Gilles Duranton. 2006. Labour pooling, labour poaching, and spatial clustering. *Regional Science and Urban Economics* 36(1):1–28.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2008*a*. Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63(forthcoming).

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. 2007. Estimating agglomeration effects: History and geology. University of Toronto.

Combes, Pierre-Philippe, Gilles Duranton, and Henry G. Overman. 2005. Agglomeration and the adjustment of the spatial economy. *Papers in Regional Science* 84(3):311–349.

Combes, Pierre-Philippe, Thierry Mayer, and Jacques Thisse. 2008*b*. *Economic Geography*. Princeton (NJ): Princeton University Press. Forthcoming.

Conley, Timothy G., Frederick Flyer, and Grace R. Tsiang. 2003. Spillover from local market human capital and spatial distribution of productivity in Malaysia. *Advances in Economic Analysis and Policy* 3(1):Article 5.

da Mata, Daniel, Uwe Deichmann, J. Vernon Henderson, Somik V. Lall, and Hyoung Gun Wang. 2005. Examining the growth patterns of Brazilian cities. Policy Research Working Paper 3724, World Bank.

da Mata, Daniel, Uwe Deichmann, J. Vernon Henderson, Somik V. Lall, and Hyoung Gun Wang. 2007. Determinants of city growth in Brazil. *Journal of Urban Economics* 61(forthcoming).

Davis, James C. and J. Vernon Henderson. 2003. Evidence on the political economy of the urbanization process. *Journal of Urban Economics* 53(1):98–125.

Davis, Steven J., John C. Haltiwanger, and Scott Schuh. 1996. *Job Creation and Destruction*. Cambridge (Mass.): MIT Press.

De Soto, Hernando. 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. New York: Basic Books.

Deichmann, Uwe, Kai Kaiser, Somik V. Lall, and Zmarak Shalizi. 2005. Agglomeration, transport, and regional development in Indonesia. Policy Research Working Paper 3477, World Bank.

Di Tella, Rafael, Sebastian Galliani, and Ernesto Schargrodsky. 2007. The formation of beliefs: Evidence from the allocation of land titles to squatters. *Quarterly Journal of Economics* 122(1):209–241.

Duranton, Gilles. 2006. Human capital externalities in cities: Identification and policy issues. In Richard J. Arnott and Daniel P. McMillen (eds.) *A Companion to Urban Economics*. Oxford: Blackwell, 24–39.

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

Duranton, Gilles and Diego Puga. 2001. Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review* 91(5):1454–1477.

Duranton, Gilles and Diego Puga. 2004. Micro-foundations of urban agglomeration economies. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2063–2117.

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

Durlauf, Steven N., Paul A. Johnson, and Jonathan W. Temple. 2005. Growth econometrics. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1A. Amsterdam: North-Holland, 555–677.

Eaton, Jonathan and Zvi Eckstein. 1997. Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27(4–5):443–474.

Epple, Dennis and Thomas Nechyba. 2004. Fiscal decentralization. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2423–2480.

Fay, Marianne and Charlotte Opal. 1999. Urbanization without growth: A not-so-uncommon phenomenon. Policy Research Working Paper 2412, World Bank.

Fernald, John G. 1999. Roads to prosperity? Assessing the link between public capital and productivity. *American Economic Review* 89(3):619–638.

Field, Erica. 2007. Entitled to work: Urban property rights and labor supply in Peru. *Quarterly Journal of Economics* 122(4):forthcoming.

Foster, Lucia, John C. Haltiwanger, and C. J. Krizan. 2001. Aggregate productivity growth: Lessons from microeconomic evidence. In Charles R. Hulten, Edwin R. Dean, and Michael J. Harper (eds.) *New Developments in Productivity Analysis*. Chicago: National Bureau of Economic Research and University of Chicago Press, 303–372.

Franco, April M. and Darren Filson. 2006. Spin-outs: Knowledge diffusion through employee mobility. *Rand Journal of Economics* 37(forthcoming).

Freedman, Matthew. 2007. Location decisions in a changing labour market environment. Processed, University of Maryland.

Fujita, Masahisa, Paul R. Krugman, and Anthony J. Venables. 1999. *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, MA: MIT Press.

Fujita, Masahisa and Tomoya Mori. 2005. Transport development and the evolution of economic geography. *Portuguese Economic Journal* 4(2):129–156.

Glaeser, Edward L. 1999. Learning in cities. *Journal of Urban Economics* 46(2):254–277.

Glaeser, Edward L. and David C. Maré. 2001. Cities and skills. *Journal of Labor Economics* 19(2):316–342.

Greenwood, Michael J. 1997. Internal migrations in developed countries. In Mark R. Rosenzweig and Oded Stark (eds.) *Handbook of Population and Family Economics*, volume 1B. Amsterdam: North-Holland, 647–720.

Harris, Chauncy D. 1954. The market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers* 44(4):315–348.

Harris, John R. and Michael P. Todaro. 1970. Migration, unemployment and development: A two-sector analysis. *American Economic Review* 60(1):126–142.

Hausmann, Ricardo, Dani Rodrik, and Andrés Velasco. 2005. Growth diagnostics. Processed, Kennedy School of Government, Harvard Univserity.

Head, Keith and Thierry Mayer. 2004. The empirics of agglomeration and trade. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2609–2669.

Head, Keith and Thierry Mayer. 2006. Regional wage and employment responses to market potential in the EU. *Regional Science and Urban Economics* 36(5):573–594.

Helsley, Robert W. 2004. Urban political economics. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2381–2421.

Henderson, J. Vernon. 1974. The sizes and types of cities. *American Economic Review* 64(4):640–656.

Henderson, J. Vernon. 1988. *Urban Development: Theory, Fact and Illusion*. Oxford: Oxford University Press.

Henderson, J. Vernon. 1997. Medium size cities. *Regional Science and Urban Economics* 27(6):583–612.

Henderson, J. Vernon. 2005. Urbanization and growth. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1B. Amsterdam: North-Holland, 1543–1591.

Henderson, J. Vernon. 2007. Exclusion through informal sector housing development. Processed, Brown University.

Henderson, J. Vernon and Randy Becker. 2000. Political economy of city sizes and formation. *Journal of Urban Economics* 48(3):453–484.

Henderson, J. Vernon and Arindam Mitra. 1996. The new urban landscape: Developers and edge cities. *Regional Science and Urban Economics* 26(6):613–643.

Henderson, J. Vernon and Anthony J. Venables. 2006. The dynamics of city formation. Processed, Brown University.

Henderson, J. Vernon and Hyoung Gun Wang. 2007. Urbanization and city growth: The role of institutions. *Regional Science and Urban Economics* 37(3):283–313.

Henderson, Vernon. 2002*a*. Urban primacy, external costs, and the quality of life. *Resource and Energy Economics* 24(1):95–106.

Henderson, Vernon. 2002*b*. Urbanization in developing countries. *World Bank Research Observer* 17(1):89–112.

Henderson, Vernon. 2003. The urbanization process and economic growth: The so-what question. *Journal of Economic Growth* 8(1):47–71.

Henderson, Vernon, Todd Lee, and Yung Joo Lee. 2001. Scale externalities in Korea. *Journal of Urban Economics* 49(3):479–504.

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Random House.

Keller, Wolfgagng. 2004. International technology diffusion. *Journal of Economic Literature* 42(3):752–782.

Klenow, Peter and Andrés Rodríguez-Clare. 2005. Externalities and growth. In Philippe Aghion and Steven N. Durlauf (eds.) *Handbook of Economic Growth*, volume 1A. Amsterdam: North-Holland, 817–861.

Krugman, Paul R. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99(3):484–499.

Krugman, Paul R. and Raúl Livas Elizondo. 1996. Trade policy and the Third World metropolis. *Journal of Development Economics* 49(1):137–150.

Lall, Somik V., Richard Funderburg, and Tito Yepes. 2004*a*. Location, concentration, and performance of economic activity in Brazil. Policy Research Working Paper 3268, World Bank.

Lall, Somik V., Jun Koo, and Sanjoy Chakravorty. 2003. Diversity matters: The economic geography of industry location in India. Policy Research Working Paper 3072, World Bank.

Lall, Somik V., Harris Selod, and Zmarak Shalizi. 2006. Rural-urban migration in developing countries: A survey of theoretical predictions and empirical findings. Policy Research Working Paper 3915, World Bank.

Lall, Somik V., Zmarak Shalizi, and Uwe Deichmann. 2004*b*. Agglomeration economies and productivity in Indian industry. *Journal of Development Economics* 73(3):643–673.

Lucas, Robert E. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22(1):3–42.

Lucas, Robert E. Jr. 2004. Life earnings and rural-urban migration. *Journal of Political Economy* 112(1, part 2):S29–S59.

Malpezzi, Stephen. 1999. Economic analysis of housing markets in developing and transition countries. In Edwin S. Mills and Paul Cheshire (eds.) *Handbook of Regional and Urban Economics*, volume 3. Amsterdam: North-Holland, 1791–1864.

Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.

Møen, Jarle. 2005. Is mobility of technical personnel a source of R&D spillovers? *Journal of Labor Economics* 23(1):81–114.

Moomaw, Ronald L. and Ali M. Shatter. 1996. Urbanization and development: A bias towards large cities? *Journal of Urban Economics* 40(1):13–37.

Moretti, Enrico. 2004. Human capital externalities in cities. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2243–2291.

Nitsch, Volker. 2006. Trade openness and urban concentration: New evidence. *Journal of Economic Integration* 21(2):340–362.

Overman, Henry G. and Anthony J. Venables. 2005. Cities in the developing world. Report, Department for International Development (UK).

Peri, Giovanni. 2002. Young workers, learning, and agglomerations. *Journal of Urban Economics* 52(3):582–607.

Ravallion, Martin and Quentin Wodon. 1999. Poor areas, or only poor people? *Journal of Regional Science* 39(4):689–711.

Redding, Stephen and Anthony J. Venables. 2004. Economic geography and international inequality. *Journal of International Economics* 62(1):63–82.

Richardson, Harry W. 1987. The costs of urbanization: A four-country comparison. *Economic Development and Cultural Change* 35(3):561–580.

Rosenthal, Stuart S. and William C. Strange. 2004. Evidence on the nature and sources of agglomeration economies. In Vernon Henderson and Jacques-François Thisse (eds.) *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: North-Holland, 2119–2171.

Rossi-Hansberg, Esteban and Mark L. J. Wright. 2007. Urban structure and growth. *Review of Economic Studies* 74(2):597–624.

Saiz, Albert. 2006. Dictatoship and highways. *Regional Science and Urban Economics* 36(2):187–206.

Thomas, Vinod. 1980. Spatial differences in the cost of living. *Journal of Urban Economics* 8(2):108–122.

Timmins, Christopher. 2006. Estimating spatial differences in the Brazilian cost of living with households location choices. *Journal of Development Economics* 80(1):59–83.

Wacziarg, Romain and Karen Horn Welch. 2003. Trade liberalization and growth: New evidence. Working Paper 10152, National Bureau of Economic Research.

Wheeler, Christopher H. 2006. Cities and the growth of wages among young workers: Evidence from the NLSY. *Journal of Urban Economics* 60(2):162–184.

Williamson, Jeffrey G. 1965. Regional inequality and the process of national development: A description of the patterns. *Economic Development and Cultural Change* 13(4, part 2):1–84.