Agglomeration and the adjustment of the spatial economy[§]

Pierre Philippe Combes[♣]
GREQAM

Gilles Duranton[♦]
University of the Andes and London School of Economics

Henry G. Overman[♥]
London School of Economics

Revised: 12th May 2005

ABSTRACT: We consider the literatures on urban systems and New Economic Geography to examine questions concerning agglomeration and how areas respond to shocks to the economic environment. We first propose a diagrammatic framework to compare the two approaches. We then use this framework to study a number of extensions and to consider several policy relevant issues.

Key words: Urban systems, New Economic Geography, Urban and regional policy, diagrammatic exposition
JEL classification: R00, R58

[§] Financial support from the UK Office of the Deputy Prime Minister is gratefully acknowledged. We are also very grateful to Tony Venables who helped develop our thinking on these issues while working jointly on a related project focused on intercity linkages. Comments from an anonymous referee and seminar participants in Cartegena de Indias also helped improve the paper.

[♣] GREQAM, 2, rue de la Charité, 13236 Marseille cedex 2, France, combes@ehess.univ-mrs.fr. Also affiliated with Paris-jourdan Sciences-Economiques and the Centre for Economic Policy Research.

[♦] Facultad de Economía, Universidad de los Andes, Carrera 1 #18A-70, Bogotá, Colombia, gduranto@uniandes.edu.co. Also affiliated with the Centre for Economic Policy Research.

[♥] Department of Geography and Environment, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom, h.g.overman@lse.ac.uk. Also affiliated with the Centre for Economic Policy Research and the Centre for Economic Performance at the London School of Economics.

1. Introduction

The last fifteen years have seen a renewal of interest from economists in spatial issues. What started as an attempt to re-invigorate regional economics by Paul Krugman and his associates has led to a flurry of theoretical work, which culminated with the publication of Fujita, Krugman and Venables' *Spatial Economy* (1999). In turn, these theoretical developments – the 'New Economic Geography' – have triggered a wave of empirical research and also inspired more policy-oriented work (e.g., Baldwin, Forslid, Martin, Ottaviano and Robert-Nicoud, 2003).

Despite this surge in interest, several crucial issues still hinder the development of spatial economics and its application to a range of policy issues. *First*, despite some claims to the contrary, the New Economic Geography (NEG) is not the undisputed framework of reference on all things spatial. Instead, it cohabits somewhat uneasily with the older approach of urban systems theory that builds on Henderson's (1974) pioneering work. There have been some attempts to reconcile these two frameworks (Fujita, Krugman and Mori, 1999, Tabuchi, 1998) but at this point we are still missing a clean comparison between the two approaches. *Second*, despite a very large number of extensions, many important dimensions of these models remain under-explored. This problem is particularly acute from a policy perspective, where the most policy-relevant issues – imperfect labour mobility, local regulations, labour force participation etc – have attracted the least attention. *Third*, both urban systems and NEG models are very difficult to work with and are thus still very poorly understood beyond their narrow circles of contributors, two circles which, somewhat surprisingly, do not exhibit much overlap. Possibly as a result of these three problems, these approaches have yet to make big headway in influencing regional and urban policy.

In an attempt to address these issues, this paper proposes a unified diagrammatic framework, which encompasses both urban systems and NEG. This diagrammatic framework helps highlight the key similarities and differences between the approaches and is also amenable to extension to give insights on a range of policy issues. The rest of the paper is structured as follows. In section 2, we first highlight the key ingredients of any consistent theory of agglomeration and the adjustment of the spatial economy to shocks. We then introduce our diagrammatic framework. In section 3, we use it to compare the urban systems approach to NEG and discuss their common ground and highlight their differences. In section 4, we show that our diagrammatic framework is flexible enough to accommodate a number of existing extensions to both approaches. In section 5, we apply our diagrammatic framework to a number of new extensions focusing our attention, in particular, on the most policy-relevant issues.

Before turning to the next section, we want to make it clear that we would not advocate the wholesale replacement of formal models by graphs in this, or any, field of research. Ultimately, any fully-fledged unified approach to regional and urban economics will have to rely on formal models.[1] We only view our diagrammatic approach as a first step, a step that is particularly useful for the preliminary exploration of well-articulated problems that can be very difficult to model more formally. In addition, diagrams are, of course, often very efficient tools with which to explain economic theories to students and policy-makers.

2. A diagrammatic analytical framework

Most economic theories concerned with agglomeration and how places respond to shocks have a common underlying structure even if this is not always apparent at first sight. Our objective is first to identify this common structure and consider the assumptions of any consistent theoretical framework. Any model of a spatial economy has to contain three elements: A spatial structure, a production structure, and some assumptions about the mobility of goods and factors.

Spatial structure: Evidently, the geographical space must be divided into some number of units, covering cities or regions in the country (we will call both of these 'areas').[2] The number of areas is often taken as given but in practice this need not be the case. New areas can be created either by private land developers (as in the US) or through local or central government action (as in most of the EU). Given a selection of areas, the model must define both an external and an internal geography. The external geography is about how the areas are related to each other, while the internal geography is concerned with the internal structure of the area (land, housing, infrastructure etc). If the focus is on linkages between places, the external geography is obviously crucial, as it will determine the channels through which a change in one area affects others. It is often taken to be exogenous, but this need not be the case as the distance between areas can change, for example as a result of changes in policy or technology. A number of factors may also change the internal geography of areas.

Production structure: Assumptions on the local availability of primary and intermediate inputs play a crucial role. Of the primary inputs, most important is labour.[3] Detailed modelling of labour is required, both to capture different skill levels and to allow for labour force participation decisions. Intermediate inputs may also play a crucial rule. It is possible to specify directly an aggregate production function

[1] Among many reasons, we would underscore that checking the consistency of formal models is much easier than that of diagrams. Formal models also lend themselves well to quantitative exercises, which can only be sketched with diagrams.

[2] Our focus is only on sub-national applications of these models.

[3] The role of land for production has received surprisingly little attention. It is ignored in NEG, while in the urban systems literature it plays a key role in housing, but not production.

relating primary and intermediate inputs to some final output. Often this is not very enlightening. Instead, assumptions about the nature of products (particularly the degree of product differentiation), the input-output structure and the degree of firm level increasing returns will determine the extent and nature of any aggregate increasing returns.[4] Imperfect competition, resulting from firm level indivisibilities, usually looms large in the analysis.

Mobility of goods and factors: Assumptions about mobility, both between and within areas, play a crucial role in determining the spatial structure of the economy. These assumptions should cover the geographical mobility of goods, services, ideas, technologies, and primary factors. The extent to which material inputs and outputs are tradable clearly varies across sectors. Some activities are 'footloose' while others are necessarily tied closely to the markets that they serve. The mobility of ideas and technologies will determine how the production function varies across space. Finally, assumptions on primary factors are crucial. Land is immobile, although its availability for use in different activities (e.g., housing versus agriculture) is endogenous. Capital is often taken as highly mobile, with the same supply price in all areas. The (imperfect) mobility of labour, both geographically and sectorally, is a fundamental issue that warrants careful treatment. The conjunction of some form of increasing returns with imperfect inter-area mobility is the main driver of the clustering of economic activity and a key determinant of the way in which linkages between areas operate. Assessing the strength of such clustering, its range across space, and its scope across industrial sectors is of key importance, as will be clear from our detailed discussion below.

Before continuing, we make three remarks. First, the analysis of linkages between cities and regions 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. Second, the exact level of spatial disaggregation selected cannot be pre-determined and depends on the questions being addressed. Third, the usefulness of different levels of sectoral disaggregation for production is also a context specific matter of modelling judgement. Put differently, the levels of spatial and sectoral disaggregation are contingent to the problem under scrutiny.

Equilibrium and its perturbation: With these ingredients in place the general equilibrium of the system can be characterised, in which location choices are made and wages, rents, and income levels determined. The equilibrium may be constrained in the short run because of supply rigidities or factor immobilities that in the longer run are removed.

[4] See Duranton and Puga (2004) for more on the micro-foundations of urban agglomeration economies.

To understand the workings of these models, it is often insightful to worry about how, given some initial situation (that may either correspond to a short run or a long run equilibrium), the system adjusts to ‘shocks’ to the economic environment. Thinking about the effects of such shocks is also fundamental from a policy perspective. There are three sorts of shocks to be considered each of which may result in different kinds of adjustment to the urban or regional system. The first we term ‘location specific’ shocks. That is, some change, which has its direct impact on a single area; for example, the construction of new houses or the closure of a factory. The second are ‘common’ shocks. These are shocks, the direct impact of which is felt in many areas. Examples are technological or institutional changes. The third sort of shock is ‘integrative’. This occurs where the inter-area geography changes, as when a road is built or telecommunications improved. Such a change will have its own effects, possibly causing relocation of some activities, as well as changing the economy’s response to other shocks.

Channels of adjustment: An economic shock changes relative prices and induces changes in quantities – relocations of activity – between areas. There are potentially many channels of adjustment. Following, say, a positive productivity shock somewhere, firms and workers may relocate, wages and/or rents may go up, the price of some final and intermediate goods may change, etc. As will become clear below, the theoretical literature often makes fairly extreme assumptions about the mobility of goods and workers and the determination of prices. Such simplifications are warranted when the objective is to clarify some aspect of the workings of the urban/regional system or replicate some stylised facts. However these simplifications also imply that the channels of adjustment after a shock are reduced to one or two, whilst most of the others are arbitrarily shut down by the extreme nature of the assumptions. For policy purposes, however, these channels of adjustment are everything.

For instance, the welfare effect of a negative location specific shock may differ substantially if the adjustment takes place through a reduction in labour force participation as opposed to out-migration. In this respect, we do not know of any theoretical model of system of cities or economic geography in which the participation decision of workers is endogenous. The only channel through which labour supply is usually assumed to adjust is in and out-migration to and from the area. When trying to understand why some cities are specialised while others are diversified, this is an acceptable simplification. When trying to understand how declining industrial cities from Pittsburgh to Liverpool have adjusted after negative shocks, it is important to consider a variety of channels of adjustment, including out-migration but also participation decisions, etc.

The main problem when considering a variety of channels of adjustment is that the algebra becomes very quickly intractable. As we discussed in the introduction, our choice here is to go instead for a graphical exposition. As will become clear below, a graphical exposition can allow for a much more flexible set of assumptions without neglecting the key relationships that govern what happens within areas and between areas. We first sketch our graphical device before turning to its main applications.

Labour demand: The first key relationship of both urban systems and NEG is the area level aggregate production function relating total output in an area to the local inputs. This relationship has received a considerable amount of theoretical attention. The main spatial feature that the literature has attempted to replicate is “agglomeration”, i.e., the concentration of a disproportionate share of economic activity in a small set of areas. As will become clear below, many models generate agglomeration through the existence of (aggregate) increasing returns at the area level. Modelling agglomeration in this way is also consistent with a second major stylised fact: the increase of most measures of productivity per capita with the size of the local population.

Deriving a local aggregate production function, which exhibits increasing returns without the market structure being degenerate (e.g., a single producer per area), is a significant challenge. Despite this, the literature has been fairly successful at proposing microeconomic foundations for such local increasing returns. These microfoundations typically rely on one of three key mechanisms: sharing, matching or learning. Sharing mechanisms emphasise how small indivisibilities (like the fixed costs associated with the production of a new local variety) can be aggregated to generate increasing returns because a larger local population allows every fixed cost to be spread over a larger number of customers. Matching mechanisms show how larger markets increase the quality of matches between economic agents and/or the probability of finding a match. Matching issues are important in a wide variety of contexts from the local labour market (employer-employee) to the local markets for intermediate goods. Finally, learning mechanisms aim to show how more frequent direct interactions between economic agents in denser environments favour the creation, diffusion and accumulation of knowledge (see Duranton and Puga, 2004, for a review).

A key feature that emerges from the literature on the microeconomic foundations of agglomeration is that many different mechanisms all lead to a local production function that exhibits increasing returns. For our purposes, this is a very positive result because one can assume some form of local increasing returns without having to rely on a specific mechanism. The negative counterpart of this result is that identifying the precise sources of agglomeration will be difficult (Rosenthal and Strange, 2004). When concerned with policy, the second caveat is that these local increasing returns derive from specific market failures. Thus, production is in general inefficient, in the sense that it does not make the best possible use of local resources. This suggests a role for policy, but the appropriate corrective policies will depend on the exact mechanism at play: the corrective policies associated with urban knowledge spill-overs are not the same as those stemming from imperfect matching on the labour market. However since the state of our knowledge about the exact sources of agglomeration economies is still very imprecise, it is legitimate to start from such local aggregate production functions and take them as given.

If we assume that 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 relating the wage of workers to the size of the local labour force. This curve is represented in figure 1a and we will refer to it as a 'wage curve' in what follows. Consistent with our discussion above, the wage in area *i* as a function of the local labour force, $w(N_i)$, is assumed to be increasing in the size of the labour force reflecting the existence of local agglomeration externalities.

The intensity of the local increasing returns is captured by the slope of the wage curve. Of course, a neo-classical inverse wage curve would be downward sloping since the amount of land is fixed so land per worker, and thus the marginal product of labour, decreases with city size. Whether one should consider a concave or convex upward sloping curve for $w(N_i)$ depends on the specifics of the microeconomic foundations.[5] However, this is ultimately an empirical issue. Our concern here is how changes in the wage curve affect the equilibrium and how the slope of the curve determines adjustment. Note that, at this stage, it is easier to assume that workers are either identical or are horizontally differentiated (i.e., all equally productive but specialised in different activities). Vertical differentiation (i.e., a workforce where some workers are more productive than others) is ignored for the time being but its main consequences will be explored later. Note finally that at this stage we rely only on internal characteristics of an area to generate the upward-sloping wage curve. As will become clear below, external geography (i.e., linkages across areas) can also play a key role.

Area crowding: The second crucial relationship concerns the various costs associated with having a significant number of households living in the same area and so we will talk of a 'cost of living curve'. The main components of the cost of living are the cost of commuting, housing and other consumption goods. The key question regards the shape of the cost of living curve as a function of the size of the local population. 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. However, the impact of area size on the cost of consumption goods is more complicated. A higher cost of land will have

[5] For instance, the popular two-stage production functions derived from the Dixit and Stiglitz (1977) model of monopolistic competition generate a convex wage curve while those derived from Salop (1979) imply a concave curve.

implications for the price of consumption goods throughout the local economy (higher retail costs etc). On the other hand, a larger market offers a wider variety of suppliers without having to import goods from elsewhere.

Two additional factors, which do not depend on local market size, may also be important in determining the cost of living. First, external geography plays a role here because importing goods from other areas may be costly and these costs will be captured by this curve as well. Second, and particularly crucial from a policy perspective, land-use and planning regulations will also have an impact since they affect the supply of land.

We represent this curve in figure 1b where we assume that the cost of living increases with population (the alternative case will be discussed further below).[6] For reasons that will become obvious, this curve is drawn with a reversed Y-axis. To be able to use this curve in a simple and tractable way, we make two simplifying assumptions when drawing it. The first one is that the cost of living is paid in monetary terms only. The second is that housing consumption per household is fixed. These two simplifications imply that changes to the wage curve will leave the cost of living curve unchanged. In reality, commuting costs have both a monetary and a time component so an increase in the wage rate increases the shadow cost of commuting and thus shifts the cost of living curve downwards on the figure. By the same token, an increase in the wage rate tends to increase demand for housing and thus land, making land more expensive. Again, this would shift the cost of living curve downwards. 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 shock to the wage curve. Hence, to keep the exposition simple, we ignore these effects in what follows. It would, however, be easy (though cumbersome) to consider them.

Like the wage curve, the position and slope of the cost of living curve reflects a range of market failures, which means that outcomes may be inefficient. For instance, un-priced urban congestion will imply an inefficiently high cost of living for any level of population. We will treat these inefficiencies as given, just like those associated with the microeconomics of production, since the object of this paper is not their detailed study. This does not, however, prevent the analysis of the broader implications of, for instance, a reduction in congestion in the centre of a large capital city due to congestion charging (e.g., London) or the construction of a mass-transit system (e.g., Bogotá). For such changes, we consider that the main direct effect is to shift the cost of living curve for these cities

[6] For the moment, we assume that everyone living in a city works so that we can use the terms population and labour force interchangeably. Later, when considering issues of labour force participation we will need to distinguish between the two.

upwards. That is, to reduce the cost of living for a given population size. Changes to the supply of land for housing as, say, local planning regulations are relaxed will shift the curve similarly.

The difference between the wage curve and the cost of living curve is represented in figure 1c. This curve represents the net disposable wage for the area and thus we call it the '*net* wage curve'. In the case of figure 1c, the curve is bell-shaped. This corresponds to the traditional case of urban economics 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 slope of the wage curve must be larger than that of the cost of living curve below a certain threshold, while it is smaller above this threshold. At this threshold, the net wage reaches its peak (point B in the figure). This peak can be interpreted as identifying an optimal area size since it maximizes the net wage of its population.[7]

Labour supply: The second curve represented in figure 1c is an inverse labour supply curve. It indicates for any level of net wage, the amount of labour supplied in the area, which is clearly increasing with net wage. For simplicity, we temporarily assume that labour supply is a function of the total local population and ignore labour force participation decisions. In that case, this curve essentially captures the migration response to local wages. A low level of mobility between areas will be captured by a very steep labour supply curve (i.e., even large wage differences do not impact much on migration flows), while perfect inter-area mobility will imply a flat labour supply curve (i.e., small wage differences imply large migration flows). Area specific effects, such as amenities, will shift this curve, with a more attractive area facing a labour supply curve that is below that of a less attractive area (workers accept a lower net wage but are compensated by higher amenities).

Equilibrium: The intersection between the labour supply and net wage curves determines the equilibrium of the model. The intersection between these two curves may not be unique. In figure 1c, the two curves intersect twice (at A and C). The labour supply curve first cuts the net wage curve from above (at A) and then from below (at C). Point A is not a stable equilibrium. It is easy to see that a small positive population shock will raise the net wage. In turn, from the supply curve, we see this higher net wage attracts more workers, which again raises net wages and this process continues until the area reaches point C. By the same token, a negative shock at point A will lead population and wages to fall to zero(a degenerate equilibrium which we do not consider further). Turning to the second intersection at C, a similar argument verifies that this equilibrium is stable. From figure 1c,

[7] More accurately, this is the constrained optimal area size. Optimality here is conditional on the local wages being optimal given the population. As shown by Duranton and Puga (2004), unconstrained optimality is unlikely to occur since it requires an efficient production process in the area while, as we discussed earlier in the text, the assumptions needed to generate agglomeration usually require some market failure. Hence agglomeration is in general associated with inefficient production and the optimality of area size is constrained by that inefficiency.

once we have established the equilibrium population (N^*) at point C, we can trace upwards to figures 1a and 1b to read off the equilibrium wage (w^*) and cost of living (H^*), respectively.

3. A diagrammatic comparison of systems of cities and NEG

In this section, we show that our diagrammatic framework can be used to explain the workings of both the urban systems and NEG literature. We begin with Henderson’s (1974) system of cities model before turning to NEG.

Urban systems

Henderson (1974) assumes that there are local increasing returns taking place within industries (localisation economies) in a city. This implies that the wage curve is as in figure 1a but specific to each sector. Different sectors may show different degrees of increasing returns. Figure 2a represents the wage curve for two different industries, 1 and 2, *as a function of local industry employment*. In the figure, industry 1 exhibits slightly higher increasing returns than industry 2 for low employment, but for high employment, this is reversed and industry 2 exhibits much stronger increasing returns than industry 1.[8] These final goods are assumed to be freely tradable so that there is no issue of market access.[9] We assume that the cost of living depends on the *total* workforce in the city and this relationship is drawn in figure 2b.

The first result of the model, that all cities will be fully specialised in equilibrium, follows from the assumption that increasing returns are industry specific, but the cost of living depends on total workforce. The reasoning behind this result is the following. Imagine that instead of being fully specialised, a city has positive employment in both industry 1 and 2. With workers being perfectly mobile between industries, the equilibrium requires wage equalisation between the two industries. The only possible interior equilibrium must be where the two curves intersect. Assume then a small shock, which moves employment to industry 1 at the expense of industry 2. Such a shock will imply a higher wage in industry 1 and a lower wage in industry 2. This in turn will imply more workers moving from industry 2 to industry 1, etc. The only stable equilibrium must thus involve the full specialisation of cities in either industry 1 or industry 2.

[8] Note that in figure 2a, the wage curves for the two industries cross. Assuming that both goods are indispensable then, in equilibrium, prices of the two goods must adjust to ensure that this is the case. If the two curves did not cross, one activity would always pay more and thus attract all the workers and the other good would not be produced.

[9] Trade costs, which imply that market access matters, introduce further complications, which are dealt with in the exposition of the NEG framework.

As discussed above, provided that at some point the marginal increase in cost of living begins to dominate the marginal increase in wages in industries 1 and 2, the net wage curve will be bell-shaped for both industries. These two curves are represented in figure 2c. Note that under our assumptions on returns to scale, optimal city size is larger for industry 2. Assume that labour is perfectly mobile across cities so that the labour supply curve is flat. This implies that the short-run equilibrium will be given by points A and B, with net wage the same in both types of cities and the population N_A and N_B for cities specialised in industry 1 and 2, respectively.

At this stage, two important properties must be noted. First, all cities will be too large with respect to their constrained optimal city size following the same stability argument as previously.[10,11] Second, all cities with the same specialisation must be of the same size. It is interesting to note that this prediction receives mild empirical support (Henderson, 1988).

The crucial issue with regard to long run equilibrium concerns the mechanism for city creation. In the long run, the number of cities is not fixed but instead can vary depending on the action of land developers. The model assumes that these developers can create cities of a size of their choosing and can use fiscal instruments to appropriate rent to finance their activity. The optimal strategy of a competitive land developer is thus to create a city of optimal size and tax the workers the difference between the net wage that the city offers and the net wage that can be obtained elsewhere. As workers move to the newly created cities, they leave old established cities. A decline in the size of the latter will increase the net wage they can offer to their workers. In turn, this dampens the incentive for further new cities to be developed. In equilibrium, the free-entry of competitive developers will imply that all cities will reach their optimal size.

This is not the end of the story, however. Note that in figure 2c the peak of the industry 2 curve (the maximum net wage payable in an optimally sized industry 2 city) lies above the peak of the industry 1 curve. That is, cities specialised in industry 2 can offer a higher wage at their constrained optimal size.

[10] It is important to note that cities are oversized only with respect to their constrained optimal size (i.e., the "optimal" size, which ignores inefficiencies in both production and housing/commuting). If for instance production is inefficient in figure 2a, it may well be the case that the first-best wage curve is above the equilibrium wage curve with a steeper slope. This steeper first-best wage curve would imply an unconstrained optimal city size larger (and possibly much larger) than the constrained optimal size. Then the comparison between the fully optimal city size and the equilibrium city size is ambiguous since there are two distortions pushing in opposite directions. On the one hand, without a market for cities, equilibrium city size is too large. On the other hand, all the benefits from agglomeration are unlikely to be exhausted, which implies equilibrium city size being too small. As discussed in the text, in reality, we also expect important inefficiencies regarding the determination of the cost of living. Such inefficiencies are likely to increase the gap between the constrained and unconstrained optimal city size.

[11] Note that this result is to some extent an artefact of perfect inter-area labour mobility. With an upward sloping labour supply curve, equilibrium city size can be stable in the region of increasing returns. See below for more on this.

However, these differences cannot persist in equilibrium with competitive land developers because they imply higher returns for building industry 2 cities. These higher returns induce the development of relatively more new cities specialised in industry 2. Thus output in industry 2 increases more than output in industry 1 and this causes the relative price of industry 2's final good to decline. This decline in price, pushes down the demand for labour in industry 2 and so the wage curve moves downwards and conversely for the wage curve in industry 1. This process continues until optimal city size is the same for both types of cities[12]. This long run equilibrium is shown in figure 2d. Developers make zero profit for both types of cities, all types of cities reach their optimal size and workers receive the same net wage everywhere. Finally, cities export the good they produce and import all other goods.

This model is still a landmark in the literature, providing the first consistent model of an urban system in which cities arise endogenously from a tension between agglomeration economies and urban crowding while interacting together through trade. However, the usefulness of this model for policy purpose remains limited. Consider for instance a negative productivity shock in a given city (which reduces the marginal productivity of labour by a given proportion). The net wage at the optimal city size will end up below that offered in other cities. Workers will move out of this city and a land developer will choose an empty site to develop another city of optimal size. Alternatively, a negative common productivity shock affecting all cities of a given type will lead to cities of smaller size and, depending on the substitutability between goods, more or less cities of that type. These examples clearly show that the ultimate margin of adjustment in the model is the number of cities. In turn, adjusting the number of cities depends crucially on the assumption of perfect labour mobility. Clearly, cities do not enter and exit the urban landscape like this, particularly in European countries where the activity of land developers is severely restricted.[13]

Despite its drawbacks for policy analysis, this model has served as a starting point for much of the subsequent theoretical work on urban systems (see Duranton and Puga, 2000, for a survey of this literature). Unfortunately, most of the extensions of this framework are concerned with only two issues: providing more detailed or alternative microeconomic foundations for the wage curve and generating more realistic predictions for the composition of economic activity in cities, since full specialisation is obviously too extreme. These extensions, however, do not tackle the problem of the entry and exit of cities (Helsley and Strange, 1997, and more recently, Henderson and Venables, 2004,

[12] Note that price changes ensure that the wage curves for both city types are tangential at this optimal city size.

[13] In contrast, in countries with fast growing populations, city creation is of fundamental importance to avoid the concentration of population in a few grossly oversized cities. The exit of cities remains a much more contentious issue. First, European governments are highly reluctant to let failing cities experience sustained population declines. Second, even in the absence of policies to prevent exits, real exits are rare occurrences. This is because an existing stock of housing capital is very slow to depreciate. Low rents, possibly falling below the rental cost of capital, allow declining cities to retain some population. A complete analysis of the appropriateness of these policies is beyond the scope of this paper.

are rare exceptions) or the issue of imperfect labour mobility. Later in the paper, we consider the issue of labour mobility as well as a number of policy relevant extensions, but for now we leave the urban systems literature and turn our attention to NEG.

New Economic Geography

Despite very significant modelling differences, the core models of NEG can be explained with the same graphical device. This is because, as we argue below, a wage curve and price index (i.e., cost of living curve) lie at the heart of NEG models. As above, the difference between the nominal wage and the cost of living curve then defines a net wage curve and assumptions on labour mobility between areas allow one to derive the long-run equilibrium.[14] The use of our graphical device is, however, made more difficult in NEG because in some core models (e.g., Krugman, 1991), the wage equation and the price index do not admit closed-form solutions. To get round this difficulty, we can either rely on numerical solutions for specific parameter values or instead use variants of these core models that can be solved analytically (Ottaviano, Tabuchi and Thisse, 2002, and Forslid and Ottaviano, 2003).

Models in the line of Krugman (1991) consider two areas and two sectors.[15] For ease of exposition in what follows, we appeal to NEG's roots in international trade and refer to the first area as 'Home' and the second as 'Foreign'. The first sector produces some homogenous good under constant returns. To simplify the derivation of the model, this good is assumed to be perfectly tradable (so that its price is equal in both areas and can be normalised). This sector is often referred to as 'agriculture' or identified with some traditional good for which workers are immobile both geographically and sectorally (i.e., they always work in this homogenous good sector). Given these mobility assumptions, the wage in this sector can then be derived directly from the normalised price of the good it produces. For empirical and policy purpose, it is enough to assume the existence of some autonomous or residual demand in each area (e.g., civil servants, pensioners, etc) so that this sector can remain in the background, while attention focuses on 'manufacturing'. Manufacturing consists of a number of firms, each operating with internal economies of scale and producing their individual variety of differentiated product. This sector is monopolistically competitive as in Dixit and Stiglitz (1977). Assumptions on the substitutability between goods ensure that all consumers demand all varieties and so each firm sells its output in both areas. Thus if manufacturing is operating in both areas, there is 'intra-industry

[14] Note that for the ease of presentation the price index is subtracted from the wage, instead of dividing the wage by this index. For this to be technically correct, both the wage curve and price index should be drawn on a log scale.

[15] The properties below are those of Forslid and Ottaviano's (2003) model, which differ only minimally from Krugman (1991).

trade'. However, the presence of transport costs means that firms' sales have a 'home-market bias'.[16] In contrast to agricultural workers, manufacturing workers are mobile and go to the area that offers them the highest utility.

The wage of the workers in the manufacturing sector in an area is represented by the wage curve in figure 3a. The shape of the wage curve is the result of a subtle trade-off between two opposing forces: a crowding effect on the product market and a home-market effect. The resolution of this trade-off and hence the slope of the curve, depends on the level of transport costs. For high transport costs, the wage curve is downward sloping (the plain curve), while for low transport costs it is upward sloping (the dashed curve).

To better understand this, consider Home (symmetric arguments hold for Foreign) and assume that employment in the manufacturing sector is arbitrarily small in Home (and thus most manufacturing activity takes place in Foreign). With high transport costs, manufacturing goods produced in Foreign are expensive in Home. Hence, not only does Home have relatively few producers but these producers are also protected from imports by high transport costs. They can thus charge high prices, which imply high wages for the workers employed in manufacturing in Home. Hence, when the manufacturing workforce in Home is small and transport costs are high, Home pays high manufacturing wages. Now, continue to make the same assumption on the shares of employment in the two different areas, but assume that transport costs are very low. With low transport costs, the protection enjoyed by manufacturing in Home is eroded. In turn, this leads to lower manufacturing prices and thus lower wages in Home. Hence, when the manufacturing labour force is small in an area, local wages are increasing in transport costs.

Consider now the effect on the wage of an increase in the size of the manufacturing workforce in Home. There are two offsetting effects. First, a larger workforce in manufacturing means a larger income in Home. Given that transport costs are positive, this extra demand will disproportionately benefit Home producers. This is the home market effect and it tends to increase wage. Note that the home market effect is weaker when transport costs are low. This is because lower transport costs make it easier to import foreign goods. Consequently, when Home income increases following an expansion of the manufacturing workforce, a greater share of that increase will be devoted to manufacturing goods produced in Foreign when transport costs are lower.

[16] The literature also uses the concept of trade costs rather than transport costs. We think of the former as a broader concept, which encompasses transport costs together with many other costs associated with trading goods remotely. For our purpose, we take these two concepts to be equivalent.

The second effect of a larger manufacturing work force is that there are more firms, which crowds the product market. When transport costs are high, the demand from Foreign will be low and the market faced by the Home firms will essentially boil down to the Home market. Then, a larger manufacturing workforce implies more Home firms but more Home firms simply lead to a more crowded local product market and this lowers prices and wages. This is the market crowding effect mentioned above. Note that the market crowding effect is also attenuated for lower transport costs. This is because lower transport costs make it easier to export manufacturing goods so that an expansion of the manufacturing workforce has a less detrimental effect on Home prices (and thus wages). It turns out that for high transport costs, crowding dominates and the wage curve is downward sloping (the solid line in figure 3a). In contrast, when transport costs are low, the home market effect can dominate leading to an upward sloping wage curve (the dashed line in figure 3a).

The reasons for this reversal are quite subtle. We have already shown above that when the manufacturing labour force is small in an area, local wages are increasing in transport costs. We now show that when the manufacturing labour force is large in an area, this relationship is reversed implying that the wage curve slopes up for low transport costs and down for high transport costs. To see this, it is easiest to take a specific example. Assume that Home employs 50 workers in agriculture and 100 workers in manufacturing, while foreign employs 50 workers in each sector. Consider first the market crowding effect, which depends on the number of manufacturing firms in Home relative to the size of the market for those firms. As transport costs decline, the manufacturing sector in Home gradually gains a better access to 100 consumers. Hence, as transport costs decline from infinity to zero, *the market for manufacturing firms* located in Home increases from 150 to 250, i.e., by 66%. Consider now the home market effect, which depends on the share of home manufacturing firms in the purchases of Home consumers. As transport costs decrease, workers in Home spend a higher fraction of their manufacturing expenditure on imported goods. More specifically, when transport costs are infinite, Home consumers buy their manufacturing varieties from the home manufacturing sectors, that is from 100 manufacturing workers. When transport costs are zero, they can buy their manufacturing goods from manufacturing in both areas, that is from 150 manufacturing workers. Hence, as transport costs decrease from infinity to zero, *the manufacturing market for consumers* in Home increases by 50%. Put differently, as transport costs decline, the market for Home producers increases faster than the market for Home consumers. Thus market crowding decreases faster than the home market effect. For sufficiently low transport costs the home market effect dominates the crowding effect and Home will pay higher wages. That is, for sufficiently low transport costs, local wages in large areas are decreasing in transport costs. Remembering that the relationship between local wages and transport costs is reversed for small areas, we see that for low transport costs the wage curve must be upward sloping while it must be downward sloping for high transport costs.

What about the cost of living? By assumption, this core NEG model neglects issues relating to land and housing and so the cost of living in an area will be determined only by the price index for the imperfectly traded (i.e. manufacturing) goods. When the manufacturing workforce in an area is small, the share of imported goods is large, which puts upward pressure on prices as these goods are subject to transport costs. As local market size increases, local prices are pushed downwards. Consequently, the cost of living declines with the size of the area.

Lower transport costs have two implications for the cost of living curve. First, it is obvious that, conditional on the size of the area, the price index is lower when transport costs are lower. Second, and slightly less obvious, the lower transport costs the less important is own area size in determining the price index and so the flatter is the cost of living curve. Thus in figure 3b the dashed line gives the cost of living with low transport costs, the solid line with high transport costs.

In figure 3c, where N_h^m and N_f^m are the size of manufacturing at Home and Foreign respectively, we report the resulting net wage curve, i.e., the difference between the wage and the cost of living, for both high and low transport costs. In each case, there is a symmetric curve for the other area since the sum of the population of both areas, N^m, is a constant. Thus we can draw the curves for both areas on the same figure (with a thick line for Home and a thin line for Foreign), the distance between the two vertical axes being equal to the total population. For high transport costs, the relevant net wage curves are represented by the solid lines which are decreasing: net wage declines with the size of the local manufacturing population since both the nominal wage and the cost of living increase. Both curves intersect for areas of equal size at A, which represent the symmetric equilibrium under perfect inter-area mobility of manufacturing workers. The equilibrium is stable in this case. To see this, imagine shifting a worker to Home from Foreign. The net wage falls in Home and rises in Foreign so that the worker wishes to move back to Foreign and so symmetry is restored. In contrast, for low transport costs, the relevant net wage curves are represented by the dashed lines, which are now increasing. The symmetric equilibrium at B is now unstable. Again, to see this, imagine moving one worker to Home from Foreign. In contrast to the situation with high transport cost, this perturbation raises the net wage in Home, thus attracting more workers and further raising the net wage. This suggests that with low transport costs, the only stable equilibria involve concentration of manufacturing in either Home or Foreign. Of course, given that the two areas start out symmetric, we cannot say which of these two equilibria will be reached.

One of the strengths of this class of model is that it allows investigation of integrative shocks. Indeed, in his original paper Krugman (1991) focuses on the question of what happens to the equilibrium as transport costs are reduced. The answer is given by the comparison between the two kinds of equilibria in figure 3c. If transport costs are high, then manufacturing is equally divided between the two areas, essentially because of the need to supply immobile consumers (workers in agriculture). As transport costs fall, it becomes easier to supply all consumers from a single area. At low transport costs the symmetric equilibrium is unstable, and the stable equilibria require all manufacturing activity be concentrated in one of the two areas (with the equilibrium being at either point C or C', depending on whether the agglomeration occurs in Home or Foreign, respectively).

The model demonstrates in a particularly sharp way how concentration can arise from market access effects, even without any externalities within the manufacturing sector. The work of Krugman (1991) has served as point of departure for a large literature (see Ottaviano and Thisse, 2004, for a survey). Before turning to these extensions, a comparison between the two frameworks analysed so far is warranted.

Comparing the two canonical models

As we discussed in the introduction, the approaches based on urban systems and NEG are often perceived as disjoint and commonly studied and further developed separately, often by different people. Although the approaches do differ, the previous two subsections make it clear that the differences between urban systems and NEG are about the key assumptions of the core models rather than the nature of the models themselves. In a nutshell, the core model of urban systems relies on local agglomeration effects and local congestion effects in the context of a large number of areas. NEG instead relies on firm level increasing returns and the existence of transport costs between a small, discrete number of areas.

These assumptions translate into differences for the shape of the three mains curves: the wage curve, the cost of living curve, and the labour supply curve. The wage curve is unequivocally increasing in the urban systems approach. Because of all sorts of local externalities (technological spillovers, thick labour market effects and input-output linkages) it is assumed that increasing returns occur in each area. In contrast in NEG the wage curve may be increasing or decreasing. This is because, as we have shown, the subtle trade-off between a crowding effect (a larger workforce leads to an increasingly crowded local market) and a home market effect (a larger workforce implies a larger local market which benefits all local firms) depends on the level of transport costs. Even when the wage curve is upward sloping, there are no pure local externalities in the NEG model, only pecuniary effects that occur because of transport costs and increasing returns to scale at the firm level.

Ultimately, the actual shape of the wage curve is an empirical question. However, it is important to note that a priori, the assumption made by the urban systems literature would appear to be relevant at smaller spatial scales where local externalities are expected to play a role. By contrast, the NEG assumptions are better suited to larger spatial units (regions, countries or even groups of countries) for which long distance market interactions are expected to play an important role while short distance effects become of secondary importance.

Turning to the cost of living curve, there is again a sharp contrast. The cost of living increases with population size for the urban system model while it decreases with size in NEG (at least in the core model of Krugman, 1991; we discuss other cases below). Again, as with the wage curve, the shape of the cost of living curve is an empirical issue. However, the spatial scale at which these theories are applied is likely to matter for this curve as well. With a more urban focus, the literature on systems of cities pays a lot of attention to rising land and commuting costs. Instead, NEG, with its more regional perspective, views commuting and housing costs as second order issues compared to the importance of market access and its impact on the price of consumption goods.

The final curve to consider is the labour supply curve. Both sets of models assume that manufacturing workers are perfectly mobile between areas and that they will move until net wages are equalised across areas. There is one difference in that assumptions on the spatial distribution of immobile workers fixes a-priori the number of spatial units in NEG approaches while the number of areas is determined endogenously in the urban systems approach. However, the role of the number of spatial units turns out to be relatively unimportant. As shown by Fujita et al. (1999), the results of the core NEG model generalise to a larger number of regions.[17] From our diagrammatic exposition above, it is clear that the urban systems approach can readily be generalised to a small, discrete number of cities (see Papageorgiou and Pines, 1999, for more on this).

Pulling all of this together, we would argue that there is no inherent contradiction between the urban system approach and NEG: the latter is trying to explain broad trends at large spatial scales while the former attempts to explain “spikes” of economic activity. Clearly, bringing these two approaches together in a unified framework is an important goal for future research.[18] This is important because distinguishing empirically between a “trend” and a “spike” is not easy, especially as a series of spikes can imply a trend when none may be present.[19]

4. Using our framework with existing extensions of the system of cities and NEG approaches

[17] However, for policy purpose the number of areas matters for the number of winners and losers and the magnitudes of the gains and losses as the spatial economy adjusts to shocks.

[18] See Fujita, Krugman, and Mori (1999) and Tabuchi (1998) for two very different early attempts.

[19] We note that exiting empirical work usually ignores this fundamental issue.

In this section, we use our framework to consider a number of extensions to the urban systems and NEG models that have been proposed in the literature. We start with two extensions to the NEG approach, before turning to extensions to the urban systems model.

The rise and fall of regional inequalities: adding housing

The core NEG framework can be enriched by either adding more features on the production side (changing the wage curve) or more features on the dispersion side (changing the cost of living curve).

Regarding the cost of living curve, one particularly natural extension is to impose some constraints on land supply by assuming that housing is imperfectly elastically supplied so that house prices are increasing with population.[20] This extension can easily be incorporated in to our diagrammatic framework by recognizing that the cost of living curve will now depend on both the cost of consumption goods and the cost of housing. As the size of the area rises, the price of consumption goods falls (as discussed above) but this fall is offset by rising house prices. At some point, we would expect rising house prices to dominate and this implies a bell-shaped cost of living curve. Perhaps the clearest way to picture this is to add together the cost of living curve in figure 1, which captures house prices and congestion, with the curve from figure 3, which captures the price index effect. This non-monotonic cost of living curve is represented in Figure 4b. As with the core NEG model, falling transport costs will have two effects on this cost of living curve. First, lower transport costs shift the curve up. Second, as transport costs decline, the first part of the bell shape for the cost of living becomes flatter as market access matters less and less. At very low transport costs, market access is essentially irrelevant and the cost of living curve is monotonically increasing in the size of the manufacturing work force.

For high transport costs, combining the wage curve from the core NEG model with a bell-shaped cost of living curve leaves the net wage curve essentially unchanged compared to figure 3, so symmetric areas, as represented by point A in figure 4c, is still the only stable equilibrium. For intermediate transport costs, a bell-shaped cost of living curve implies a bell-shaped net wage curve consistent with some, but not complete agglomeration of manufacturing workers in a single area. Note that this extension leads to a net wage curve that is similar to that in the urban system approach. This is particularly interesting given that the forces driving these two curves are very different. In this case, the symmetric situation is no longer a stable equilibrium. Instead the two stable equilibria at B and B' involve some agglomeration of manufacturing in one area or the other. Finally, when transport costs

[20] What we do here is to use our graphical framework to assess what happens if housing costs are added to the core Krugman model. This is in contrast with Helpman (1998) who presents a formal model incorporating housing but also changing other key assumptions of the core Krugman model (e.g., getting rid of the agricultural sector).

are very low, the net wage curve now intersects the net wage curve for the other area in its downward sloping region. The symmetric situation in C is again a stable equilibrium.[21]

Once again, we can work through the thought experiment of allowing transport costs to fall and think what this implies for equilibrium. When transport costs are very high, manufacturing workers are evenly distributed across areas because it would be too expensive to serve the agricultural demand of a 'peripheral' area (no manufacturing) from an agglomerated 'core'. Then, for intermediate transport costs, it pays for manufacturing workers to agglomerate and economise on transport costs while serving peripheral farmers from the core. Finally, for low transport costs, high housing costs in the core become dominant. Put differently, this extension of Krugman (1991) adds one dispersion force (housing costs), which is independent of transport costs. As we saw above, in Krugman (1991), all forces get weaker when transport costs decline. It is thus natural that this added dispersion force should dominate when transport costs are sufficiently low.

The effects of market integration on the location of activity are therefore ambiguous in this type of framework. At relatively high levels of transport costs a reduction in barriers promotes spatial disparities. At lower levels it permits the dispersion of activity and a narrowing of disparities. This is an example of the famous bell-shaped effect of market integration on regional inequality.

The rise and fall of regional inequalities: input-output linkages and labour mobility

A similar bell-shaped effect of market integration on regional inequality can be obtained from a very different extension of NEG. Following Puga (1999) and Krugman and Venables (1995), assume workers are mobile across sectors (agriculture and manufacturing) but not across areas. This last assumption allows us to simplify our diagram since labour market equilibrium only requires that agriculture and manufacturing workers get paid the same nominal wage rather than the same net wage, since all types of workers in an area will face the same cost of living.[22] The agricultural good is freely tradable across areas and is produced using labour and land. The wage in agriculture as a function of manufacturing employment is represented by the dashed upward-sloping curve in figures 5 a-c. This line slopes upwards because increased manufacturing employment in an area implies lower agricultural employment, higher land to labour ratios in agriculture and thus higher agricultural wages.

[21] Both the wage and the cost of living are increasing in city size. That the change in cost of living dominates depends on the so called 'no black hole' condition which impose the necessary constraints on the relevant parameters.

[22] Note however that welfare analysis requires looking at net wages. Workers in *different* regions may get paid different net wages in this context.

The manufacturing sector is as in Krugman (1991) with the added sophistication that existing varieties are also used as intermediate inputs by manufacturing producers. This implies that the shape of the wage curve for manufacturing will be the outcome of a trade-off between three different forces. First, and as in Krugman (1991), market crowding still occurs: as the size of the manufacturing sector increases locally, the price index declines. Under free-entry, a decline in the price index implies that the maximum wage that producers can pay also declines. Second, there is also a home-market effect as in Krugman (1991). Since workers are immobile geographically, the workings of this effect are slightly different from that in Krugman (1991). A large manufacturing sector locally implies a larger market for manufacturing goods because the manufacturing sector consumes its own varieties as input, whereas in Krugman (1991) the effect was driven by the increase in the number of workers. Third, there is an additional local increasing return working through cost linkages: A larger local manufacturing sector means a greater number of varieties and a lower price index for manufactured goods, which are used as inputs in to the manufacturing process as well as for final consumption. Coupled with free entry, this works to partially offset the effect of market crowding on wages. Just like in Krugman (1991), the crowding effect dominates when transport costs are high while the home market and cost linkage effects become more important for lower transport costs.

The plain curve in figure 5a represents the wage curve for manufacturing in the case of high transport costs. Consider an increase in the manufacturing workforce starting from zero employment in manufacturing. Cost linkages and the home market effect first dominate before crowding kicks in and lowers manufacturing wages as the manufacturing workforce expands. The equilibrium occurs at point A. The plain curve in figure 5b represents the manufacturing wage curve for intermediate transport costs. In this case, the home market effect reinforced by cost linkages dominates the crowding effect as manufacturing expands. There are now three equilibria at points B1, B2, and B3. The equilibrium in B2 is unstable so that area 1 is either at B1 while area 2 is at B3 or area 1 is at B3 while area 2 is at B1. Compared to A1, B3 implies a larger manufacturing workforce.[23] Finally the plain curve in figure 5c represents the manufacturing wage curve for low transport costs. In this case, cost linkages are much weaker so that the manufacturing wage curve is much flatter than in the previous case. There is a unique equilibrium at C. This equilibrium implies a more even distribution of manufacturing than for intermediate transport costs. The reason behind this is that concentration of manufacturing in one area bids up the wage in that area while it depresses the wage in the other area. When transport costs are low, manufacturing in the core area has to compete with manufacturing in the peripheral area. Since workers are the same, the only source of wage difference arises from the costs of importing intermediates and exporting output. As transport costs decline, the wage difference between the two

[23] Note that B3 could be to the right of the right-hand-side the vertical axis, which would imply complete specialisation in manufacturing.

areas has to decline, which implies manufacturing moves back to the periphery. Again, the bell-shaped effect of market integration is obtained.

Sorting

As shown by Combes, Duranton, and Gobillon (2005), workers tend to sort across cities according to observed and unobserved characteristics. More generally, skilled workers are over-represented in large cities while unskilled workers tend to live in smaller cities (Peri, 2002). Inspired by the analysis of Abdel-Rahman and Wang (1997), it is possible to show how such a pattern can emerge using our diagrammatic framework. The starting point of the analysis is to assume that workers with higher skills benefit more from agglomeration effects. If, for instance, agglomeration has a multiplicative effect on individual productivity (as routinely assumed by empirical studies), the benefits from agglomeration will increase with skills. Graphically this implies two wage curves, one for skilled workers and another for unskilled workers. This is represented in figure 6a. After taking into account the cost of living curve, the net wage curves for these two groups of workers are represented in figure 6c.[24] Unsurprisingly the net wage curve for skilled workers is above that for unskilled workers. With both groups of workers being perfectly mobile and assuming that skilled workers enjoy a higher reservation level, the two labour supply curves are also represented in figure 6c.

The equilibrium is represented by points A and B for skilled and unskilled workers. It can be verified that at this equilibrium, skilled workers have no incentive to move to a smaller unskilled city while unskilled workers would enjoy a lower utility by moving to a larger skilled city. Thus, sorting by skill is an equilibrium outcome.

Changes in specialisation and the structural transformation of cities

The extension of the urban system framework proposed by Duranton and Puga (2005) is useful for thinking about the transformation of the very largest city economies during the de-industrialisation period. This model was motivated initially by changes in the US urban system which saw cities shifting from being predominantly specialised by sector to being specialised by function (e.g., production, management, etc). To model this, Duranton and Puga (2005) use the Henderson (1974) framework but allow the spatial organisation of firms to be endogenous.

[24] We assumed that the cost of living was the same for both groups. This is unlikely to hold empirically since skilled workers tend to consume more land because of their higher wages. However, with land being a normal good, a higher cost of living for skilled workers is unlikely to offset completely the differences in wages.

Firms, in each sector, require both sector-specific inputs (e.g., car parts for car producers) as well as business services for their headquarters. There are agglomeration economies in all sectors including business services. Firms face a trade-off between spatial integration of both headquarters and production facilities and the spatial separation of these two functions. If firms decide to locate their production facilities and headquarters separately, then both parts of the operation can fully benefit from the relevant agglomeration externalities (sector specific inputs for the production facility, business services for the headquarters). In contrast, spatial integration allows firms to manage the interaction between production facilities and headquarters more efficiently because they save on communication costs between the two units, but this comes at the cost of more expensive inputs to both parts of the operation (because of the negative reciprocal crowding that both activities generate in the city).

When communication costs are high, it is very costly for firms to spatially separate their headquarters from their production plant. Consequently the demand for labour of spatially disintegrated firms is low. As a result, a functionally specialised city will pay low wages, captured by the bottom wage curve in figure 7a. In particular, this wage curve will be below that of a city specialised by sector (medium wage curve in figure 7a). However, when communication costs are low, it is beneficial for firms to spatially separate their facilities and, as a result, cities will be specialised by functions, production or business services. The advantage of functional specialisation is that headquarters (regardless of their sectors) will be able to exploit more fully the economies of scale in cities specialised in business services and management activities. For low communication costs, the wage curve of functionally specialised cities is represented by the top curve in figure 7a.

Using these wage curves, together with the cost of living curve represented in figure 7b, we draw, in figure 7c, the net wage curve for all three types of cities. It is clear that functionally specialised cities are not viable for high communication costs (the net wage curve does not intersect with the labour supply curve). For low communication costs, functionally specialised cities are viable and the equilibrium size of a city *functionally* specialised in management and business services when communication costs are low is larger than that of a city specialised by sector.

If we accept the idea that technological progress in the last 40 years has made it easier to manage production activities remotely (cheaper telecommunication, intranets, etc), we can shed some light on the structural transformation of the economy of cities such as London, New York or Paris. In the case of London, for instance, the structural transformation that took place between the 1960s and the 1990s can be understood as a successful transition from an integrated city with a production structure encompassing a range of manufacturing sectors, management, government, and the handling of goods to a functionally specialised city concentrated in business services, management, and high-end government activities. In figure 7c, this corresponds to a shift from A to B. This increased concentration of high value-added activities with large agglomeration economies in London may have taken place to some extent at the expenses of many cities from the rest of the country. As shown by Duranton and Puga (2005), if we assume (in line with empirical evidence) that business services benefit from stronger localisation economies than other industries, business centres will benefit from stronger agglomeration economies than cities combining production and management and in turn the latter will enjoy stronger economies of scale than purely production cities. Referring back to figure 2a, and holding communication costs constant, one can think of business services as industry 2, production activities as industry 1 with integrated cities as a hypothetical industry 3 with the net wage curve lying between industry 1 and 2. This implies that cities loosing their management functions also have a lower optimal size. In the London case, this implies that the counterpart of the growth of London was the population decline of cities in Northern England.

In a world of perfect mobility, this structural transformation would only imply higher wages in all cities because a more efficient urban organisation increases productive efficiency and with competitive land developers, the benefits of this increased efficiency will are passed on to workers in the form of higher wages. However, as we will see below, in a world of imperfect mobility and endogenous labour force participation, this population decline in Northern English cities may have had very negative implications.

5. Pushing the framework: imperfect labour mobility, planning regulations and labour force participation

Imperfect inter-area labour mobility and margins of adjustment

As mentioned above, the diagrammatic framework we introduce here allows us to be more general than previous literature. A key issue that we want to address are the implications of allowing for imperfect labour mobility between areas. In figures 2 to 7, we assumed flat labour supply curves, due to a perfect labour mobility assumption. We now move back to the case where mobility is imperfect, as drawn in figure 1. In this figure, the only stable equilibrium occurs in the region where net wage shows decreasing returns with respect to city size.[25] In contrast, we draw in figure 8c a steep upward sloping supply curve (very imperfect inter-area mobility) so that it intersects the net wage curve in the region of increasing returns (i.e., for a level of population below the optimum), giving a unique stable equilibrium at point A. This difference has some interesting implications. In Henderson (1974), the

[25] Note that this is necessarily the case under perfect mobility.

combination of perfect labour mobility between areas and competitive large agents (developers) implies all areas are at optimal population in long run equilibrium. In the absence of large agents, perfect mobility alone leads to overpopulated areas (see figure 2).

What we show here is that in the absence of large agents *and* with barriers to perfect mobility, areas can be too small in equilibrium. Put differently, barriers to mobility can reduce the inefficiencies associated with the absence of a market for cities. This echoes standard second-best results whereby one distortion can reduce the negative effects of another. However, as usual, correcting a distortion by introducing another distortion may come at a cost. With barriers to mobility, areas may not be able to exploit fully the extent of their increasing returns. If as suggested by the empirical literature, the costs of being too large are rather small compared to the opportunity costs of being undersized (Au and Henderson, 2004), limited labour mobility may have important hidden costs by preventing the exploitation of agglomeration economies. Limited labour mobility also has implications for the way in which the urban system adjusts to shocks, an issue to which we now turn.

Consider a positive productivity shock in the area. For any given level of city size, firms can now pay a higher wage and so the wage curve shifts upwards. As shown in figure 8, the new wage curve implies a new net wage curve (the two dashed lines) and a new equilibrium at point B. This new equilibrium implies higher wages, higher housing costs and a larger workforce. The magnitude of these effects depends on the slope of the different curves, i.e., on the three key elasticities of the model (that of wages to employment, of the cost of living to population, and of labour supply to net wages). What happens to the long run equilibrium now depends on our assumptions about the elasticity of labour supply. We can identify three different possibilities.

If labour is perfectly mobile between areas, labour supply is infinitely elastic and the higher net wage attracts an inflow of new workers up until the point where net wages are again equalised across areas. The area experiencing the positive productivity shock ends up with a larger population, but no change in net wages. Put differently, a positive productivity shock will be fully crowded out by an increase in the workforce.[26] The polar case of perfect immobility between areas (i.e., a vertical labour supply curve) implies the opposite: net wages increase by the full amount of the productivity shock, while the population is unchanged.[27]

When the elasticity of labour supply is neither zero nor infinite, the situation is more complicated and the effect of a positive productivity change will depend on the relative slope of the two curves: net

[26] This result depends on the assumption that there are a large number of cities. The case with a small number of cities is more complicated and is discussed in the Appendix.
[27] Clearly, the assumption that cost of living does not depend on wages in the city is crucial when reaching this conclusion. See our earlier discussion.

wage and labour supply. We can make three general points. First, the higher is the elasticity of labour supply, the greater is the quantity (i.e., population) response and the smaller the price (i.e., net wage) adjustment. Second, for equilibria in the region of increasing returns for the net wage, a positive productivity shock will be magnified and lead to an increase in net wage that is larger than the original productivity shock. In contrast, in the region of decreasing returns, the effect of the shock on net wage will be smaller than the initial shock. Third, the elasticity of net wages in turn depends on the elasticity of wages (positively) and the elasticity of the cost of living (negatively).

The analysis of other location-specific shocks follows directly from this. For instance a cheaper cost of living has the same effect on the net wage curve as a positive productivity shock. As shown in the Appendix, this analysis generalises to the case of a small discrete number of areas.

The analysis of common shocks is slightly more involved because a productivity shock in the economy affects not only the wage curve of an area but also its labour supply curve. Following a common positive productivity shock, the wage curve (which represents labour demand) shifts upwards as in the case examined above, but now the labour supply curve also shifts upwards. The final effect depends on how the two curves shift relative to each other. If the shifts of the two curves are perfectly symmetric (i.e., the shock has similar effects everywhere), the shift of the wage curve will be exactly offset by that of the labour supply curve so that the new equilibrium will occur for a higher net wage but with the same population. If the effects of the shock are not exactly the same everywhere, the area may gain or lose population. These gains and losses will be larger when labour supply is more elastic.

To see this more precisely, consider a common positive productivity shock leading to a proportional increase in productivity. This implies an anti-clockwise rotation of the wage curve around its origin in our figure. In turn, the net wage curve shifts upwards implying an increase in the optimal population size. If the labour supply curve also shifts upwards following this common productivity shock, we may obtain two (interior) equilibria rather than one (with the new equilibrium being unstable). Since not all areas can grow at the same time (we need to keep population constant), this implies that some areas will lose population. Assuming that the upward sloping labour supply curve reflects differences in the willingness to migrate then the least mobile workers will be stuck in undersized areas. Put differently, a positive common productivity shock may be bad news for some areas.[28] Such results echo that of Duranton and Puga (2005) outlined above which relies on a decline in communication costs between cities. The general conclusion we draw here is that technological progress, even though it may be "smooth" from a macroeconomic perspective, is likely to disrupt the urban system and imply a smaller optimal number of cities. Imperfect labour mobility clearly prevents this type of adjustment

[28] Of course the same type of result would be obtained if, instead of a common positive productivity shock, we would consider an improvement in commuting technology leading to a shift of the cost of living curve.

and can lead to large welfare losses. This type of problem may be mitigated by population growth (which implies a larger number of cities). However in countries with stagnant or decreasing populations, these issues are likely to become more acute in the future.[29]

We can finally turn to the effects of integrative shocks. For such shocks, the NEG framework is probably more appropriate for thinking through the consequences. These extensions have been proposed by Tabuchi and Thisse (2002) and Murata (2003). To model limited labour mobility, they use a standard discrete choice framework with heterogeneous workers: all workers are potentially mobile but the costs of moving are idiosyncratic. This case can be treated graphically in an identical manner to the analysis in the Appendix. The main result of Tabuchi and Thisse (2002) and Murata (2003) is that, with limited labour mobility, lower transport costs do not lead to unstable symmetric equilibria and catastrophic agglomeration, thus reversing the result of Krugman (1991) that we worked through above. The reason behind this result is that imperfect labour mobility provides an extra dispersion force: An upward sloping labour supply curve, which gets vertical as the whole population concentrates in one area, implies that the labour supply curve always cross the net wage curve from below regardless of the shape of the latter (provided degenerate cases such as a vertical net wage curve are avoided). Hence the equilibrium is always interior and stable.

Planning regulations

Another issue that we can address using our diagrammatic framework is that of the impact of planning regulations from both a positive and a normative perspective. To keep the analysis short and simple, let us examine the effects of a strict growth control policy, which forbids any further land development in a location after it hits some upper size bound. In effect, this implies an upper bound for size of the local population. Figure 9a presents a standard upward-sloping wage curve. The solid curve in figure 9b represents the cost of living under growth control while the dashed curve represents the cost of living in the absence of growth control. Using these two curves, in figure 9c, we obtain the net wage curve under growth control and the net wage curve without growth control (solid and dashed line respectively). The last curve represented in figure 9c is the labour supply curve. It is simple to read off that the equilibrium with growth control is at A while the equilibrium without growth control is at B.

The first important result regards the effects of growth control on whether cities are under or oversized. Just like barriers to mobility, growth control measures have an ambiguous efficiency effect in a second-best world. The reason is that in the absence of a mechanism, like large agents, to implement the efficient population size of an area, the area may be over-populated in equilibrium. In

[29] Again raising the issue of how to deal with the decline of some cities. See the end of section 4.

contrast, growth controls typically imply under-population. However, as highlighted before it is always dangerous to offset one distortion by introducing another distortion. Indeed, in figure 9c the net wage under growth control is actually smaller than the net wage if the city was oversized (compare the net wage at A to the net wage at B).

In case of a positive location specific shock when growth control is binding, note that the sole margin of adjustment is the cost of living. This is because a positive productivity shock raises wages. In turn this makes the area more attractive. Then, potential in-migrants bid up the price of housing. However, because of growth control, the population cannot increase. Hence the adjustment is such that the net wage must remain constant. This implies that the increase in housing costs completely offsets the increase in wages. Supportive evidence of this type of effect is given in Glaeser et al. (2005).

Labour force participation

We now enrich our framework to consider yet another margin of adjustment: participation in the labour market. This margin is usually neglected in the theoretical literature, although empirically it appears to play a major role. According to Decressin and Fatás (1995), in the European Union, when a region experiences a negative employment shock, 78% of the adjustment occurs through labour force participation in the first year and 50% through participation in the second year (the remaining adjustment occurring through unemployment and out-migration). Figure 10a shows how the standard wage curve moves when the location experiences a negative productivity shock (from the solid to the dashed line). Figure 10b is a standard cost of living curve. As usual, the difference between the wage and the cost of living leads to the net wage curve in figure 10c. The solid bell-shaped curve is the net wage before the shock and the dashed bell-shaped curve represents the net wage after the shock.

The main difference with the previous figures concerns the labour supply curve. The upward sloping straight line is a standard labour supply curve. With constant labour force participation, the effect of a negative productivity (or demand) shock is straightforward. Some workers leave the area after the shock and the local economy adjusts to a lower wage and a lower local population. On the figure, the equilibrium would shift from point A to B.

To model the participation decision, we assume that below a certain net wage threshold, workers drop out of the labour force and rely on the welfare system instead.[30] Consider again the same negative productivity shock, this time taking the participation constraint into account. After the shock the

[30] Whether the participation decision is driven by the level of nominal wage or instead by the net wage is an empirical question. In countries where unemployed workers are likely to receive housing subsidies as well as unemployment benefits in cash like the UK, it is natural to assume that participation is driven by net wages.

unconstrained equilibrium (point B) is not feasible because before reaching point B, at point D, the local economy hits the reservation net wage. Labour force participation then drops until labour supply crosses the net wage at point C. The gap between D and C on the X-axis is the extent of non-participation in the labour force. In this case, the adjustment to a negative shock implies not only lower wages and out-migration but also workers dropping out of the labour force and becoming unemployed or claiming alternatives forms of benefits such as disability benefits, etc.

Note that the consequences of a negative shock could be much worse than in the case just discussed. If for instance, the maximum of the dashed net wage curve (the net wage curve after the shock) were to go below the reservation level of workers, participation would then drop to zero.[31]

6. Conclusions

This paper develops a diagrammatic framework to analyse the spatial economy and how it adjusts to shocks. This framework allows a clear comparison between the core models of the urban systems approach and NEG. Existing extensions of these models can also be studied within the same framework. Beyond this our diagrammatic framework lends itself well to a number of policy-oriented extensions. In particular, we consider the implications of imperfect labour mobility, local (land) supply regulations, and labour force participation decisions, issues that are seldom studied in the existing literature. Further policy extensions to be dealt with in future work regard other spatial policies that we ignored here. The list of such policies is potentially very long and includes, among others, regional subsidies, the taxation of agglomeration rents, etc (see Baldwin et al., 2003, for an analytic treatment of these two types of policies in an NEG framework).

From a theoretical perspective, a weakness of our approach is that we still look at urban systems and NEG models separately whereas one of our conclusions is that these two approaches should be looked at together: NEG is preoccupied with broad spatial trends whereas the urban system literature is concerned with spikes in the economic landscape. The two are mutually dependent, but a full analysis of their joint implications must wait for another time.

Our results underscore the importance of three key relations. The importance of the relation between local employment and local wages has been acknowledged for a long time and much is known about it

[31] When the labour supply curve intersects the net wage curve in the region of increasing returns, some dramatic consequences are also possible. In particular, it may again be the case that, after a shock, the labour supply curve no longer has any intersection with the net wage curve. This would again lead the labour force participation to drop to zero. This case is particularly inefficient because a less steep labour supply line (i.e., more labour mobility) can imply another equilibrium with full labour force participation.

both theoretically and empirically. The issues surrounding the cost of living are well understood theoretically but empirically estimating the cost of living curve has proved elusive. Finally, we also showed that the local labour supply was a fundamental determinant of the way in which areas adjust to shocks. Much still needs to be learnt about local labour supply, both theoretically and empirically. We hope future work will address these issues.

Appendix: Interactions among a small number of areas under imperfect labour mobility

It is possible to generalise the analysis of shocks under imperfect mobility to account for the possible interactions among a small number of areas. This extension is particularly relevant when migration flows are mostly short distance. The concave upward sloping curve in figure A1 represents the net wage for area 1. It can be derived in exactly the same fashion as before from a wage curve and a cost of living curve, which we omit for simplicity. The second important curve in the figure is the net wage curve for area 2 (a thinner curve that mirrors the net wage curve for area 1). Assuming that the total population of the two areas is constant, this net wage curve shows directly the wage in area 2 for any level of population in area 1. Under perfect mobility, the equilibrium would imply the equalisation of net wages in the two areas at point A (which incidentally would be unstable since the intersection is in the region of increasing returns). With imperfect labour mobility, the supply of labour in area 1 is represented by the convex upward sloping curve in the figure. It intersects the net wage curve of area 1 at point A’. Note that this equilibrium implies a larger population in area 1. A possible reason behind this could be the following. In the region of increasing returns, there are forces pushing towards the concentration of the entire population in one area. However, workers may have personal preferences for one or the other area. These locational preferences imply a convex labour supply curve as in the figure and the workers staying in area 2 are those with strong preferences for this area.

Consider now a positive productivity shock in area 1. This productivity shock implies an upward shift of the net wage curve in the figure (the concave dashed curve). If the labour supply curve was determined only by the net wage in area 1, the new equilibrium would be found at point B. However, it is more reasonable to expect migration to be driven by the *differential in net wages* rather than the net wage at destination only. Consequently, we expect the productivity shock to shift the labour supply curve as well, although this shift will be downwards not upwards, which gives the dashed convex curve. This is because for any level of population in area 1, the difference in wages between the two areas is higher after the shock.

The shift of the net wage curve of area 1 between A' and C' now implies that the post-shock gap in net wages for the pre-shock population levels is represented by the segment C – C' (instead of C – A' before the shock). With C' – D' running parallel to C – D, the new gap in net wages can be represented by D – D'. Such a difference in net wages would have implied an equilibrium in D' before the shock. Put differently, aside from a shift in the net wage curve, the shock also implies a labour supply curve shift from A' to D'. The new equilibrium after the shock is then represented by point E. To summarise, a positive productivity shock in an area has two effects. There is a direct effect making the area more attractive in absolute terms. With a large number of areas (and no area being able to affect the national equilibrium), the adjustment would stop here. However, with a small number of areas, for any level of population an increase in the attractiveness of an area implies a larger difference in net wages between the two areas and thus a shift in the labour supply curve, which adds to the first effect.

References

Abdel-Rahman HM, Wang P (1997) Social welfare and income inequality in a system of cities. *Journal of Urban Economics* 41: 462-483

Baldwin RE, Forslid R, Martin P, Ottaviano GIP, Robert-Nicoud F (2003) *Economic geography and public policy*. Princeton University Press, Princeton, NJ

Combes PP, Duranton G, Gobillon L (2005) Spatial wage disparities: Sorting matters! Processed, London School of Economics

Decressin J, Fatás A (1995) Regional labour market dynamics in Europe. *European Economic Review* 39: 1627-1655

Dixit AK, Stiglitz JE (1977) Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297-308

Duranton G, Puga D (2000) Diversity and specialisation in cities: Why, where and when does it matter? *Urban Studies* 37: 533-555

Duranton G, Puga D (2004) Micro-foundations of urban agglomeration economies. In: Henderson V, Thisse JF (eds) *Handbook of Regional and Urban Economics*, volume 4. North-Holland, Amsterdam

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

Forlsid R, Ottaviano GIP (2003) An analytically solvable core-periphery model. *Journal of Economic Geography* 3: 229-240

Fujita M, Krugman PR, Mori T (1999) On the evolution of hierarchical urban systems. *European Economic Review* 43: 209-251

Fujita M, Krugman PR, Venables AJ (1999) *The spatial economy: Cities, regions, and international trade*. MIT Press, Cambridge, MA

Glaeser EL, Gyourko J, Saks R (2005) Urban growth and housing supply. *Journal of Economic Geography* (forthcoming)

Helpman E (1998) The size of regions. In: Pines D, Sadka E, Zilcha I (eds) *Topics in public economics. Theoretical and applied analysis.* Cambridge University Press, New York, NY

Helsley RW, Strange WC (1997) Limited developers. *Canadian Journal of Economics* 30(2): 329-348

Henderson JV (1974) The sizes and types of cities. *American Economic Review* 64: 640-656

Henderson JV (1988) *Urban development: Theory, facts and illusion*. Oxford University Press, Oxford

Henderson JV, Venables AJ (2004) The dynamics of city formation: Finance and governance. Processed, London School of Economics

Krugman PR (1991) Increasing returns and economic geography. *Journal of Political Economy* 99: 484-499

Krugman PR, Venables AJ (1995) Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857-880

Murata Y (2003) Product diversity, taste heterogeneity, and geographic distribution of economic activities: market vs. non-market interactions. *Journal of Urban Economics* 53: 126-144

Ottaviano GIP, Thisse JF (2004) Agglomeration and economic geography. In: Henderson V, Thisse JF (eds) *Handbook of regional and urban economics*, volume 4, North-Holland, Amsterdam

Ottaviano GIP, Tabuchi T, Thisse JF (2002) Agglomeration and trade revisited. *International Economic Review* 43: 409-436

Papageorgiou YY, Pines D (1999) *An essay on urban economic theory*. Kluwer Academic Publishers, Boston, MA

Peri G (2002) Young workers, learning, and agglomeration. *Journal of Urban Economics* 52: 582-607

Puga D (1999) The rise and fall of regional inequalities. *European Economic Review* 43: 303-334

Rosenthal SS, Strange WC (2004) Evidence on the nature and sources of agglomeration economies. In: Henderson V, Thisse JF (eds) *Handbook of Regional and Urban Economics*, volume 4. North-Holland, Amsterdam

Salop SC (1979) Monopolistic competition with outside goods. *Bell Journal of Economics* 10: 141-156

Tabuchi T (1998) Urban agglomeration and dispersion: A synthesis of Alonso and Krugman. *Journal of Urban Economics* 44: 333-351

Tabuchi T, Thisse JF (2002) Taste heterogeneity, labour mobility and economic geography. *Journal of Development Economics* 69: 155-177

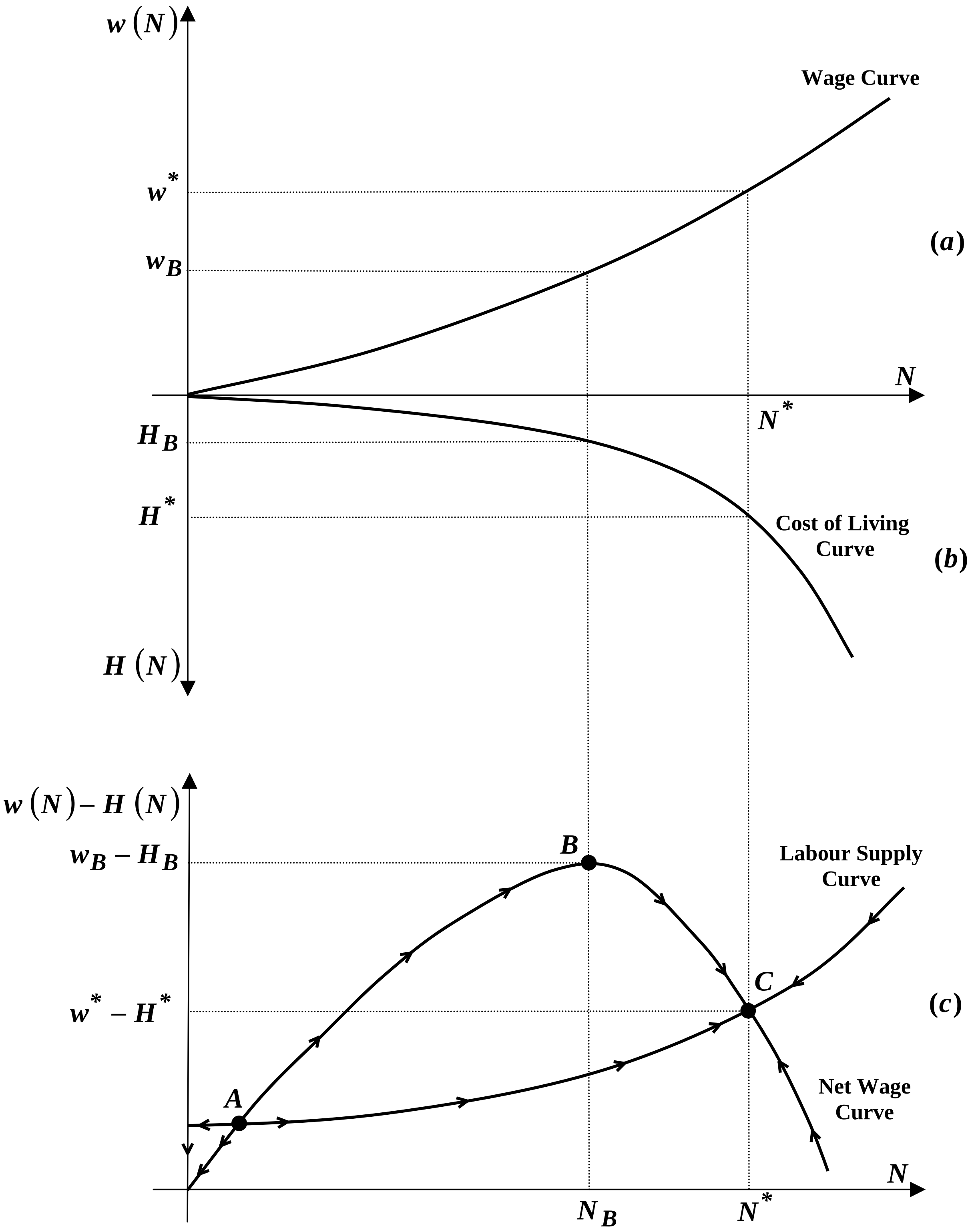

$w(N)$
Wage Curve
w^*
w_B
(a)
N
N^*
H_B
H^*
Cost of Living Curve
(b)
$H(N)$
$w(N) - H(N)$
$w_B - H_B$
B
Labour Supply Curve
C
$w^* - H^*$
(c)
A
Net Wage Curve
N
N_B
N^*

Figure 1

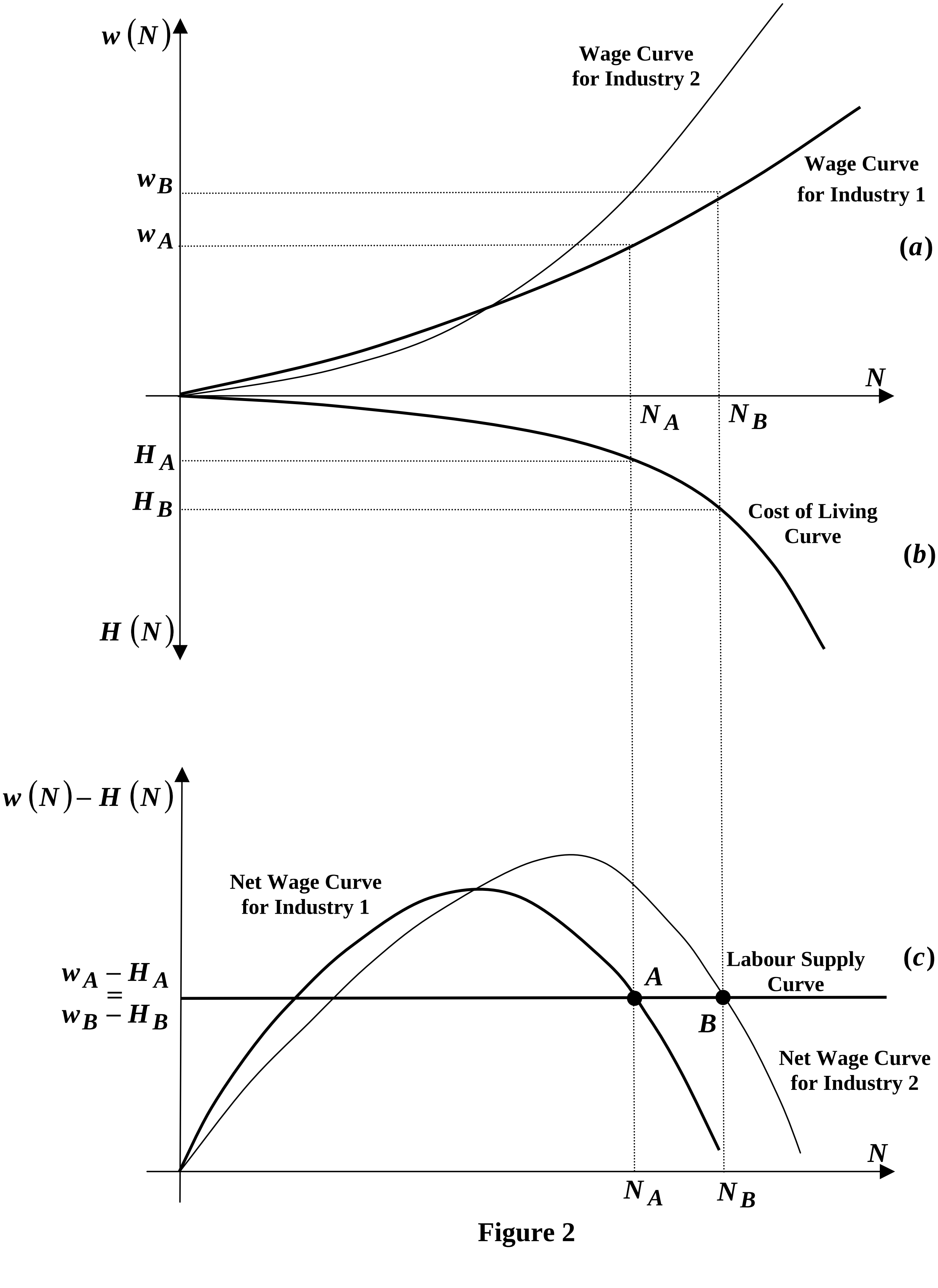

$w(N)$
Wage Curve
for Industry 2
w_B
w_A
Wage Curve
for Industry 1
(a)
N
N_A
N_B
H_A
H_B
Cost of Living
Curve
(b)
$H(N)$
$w(N) - H(N)$
Net Wage Curve
for Industry 1
$w_A - H_A$
$=$
$w_B - H_B$
A
B
Labour Supply
Curve
(c)
Net Wage Curve
for Industry 2
N
N_A
N_B

Figure 2

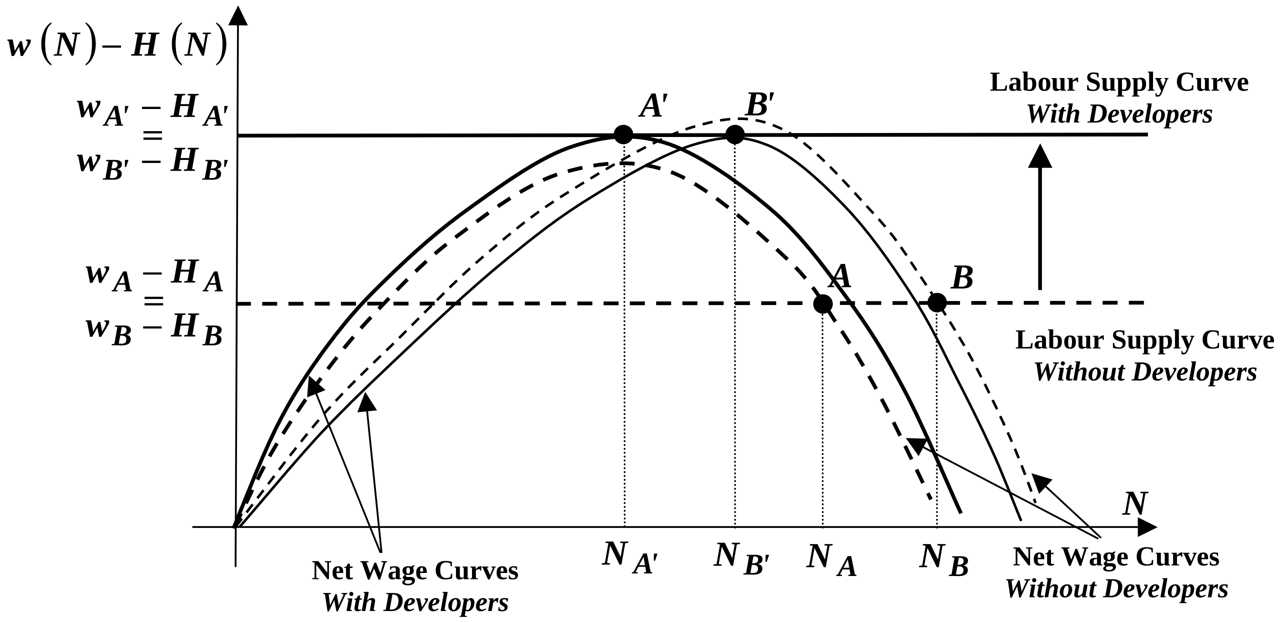

$w(N) - H(N)$
$w_{A'} - H_{A'}$
=
$w_{B'} - H_{B'}$
$w_A - H_A$
=
$w_B - H_B$
A'
B'
A
B
Labour Supply Curve
With Developers
Labour Supply Curve
Without Developers
N
$N_{A'}$
$N_{B'}$
N_A
N_B
Net Wage Curves
With Developers
Net Wage Curves
Without Developers

Figure 2d

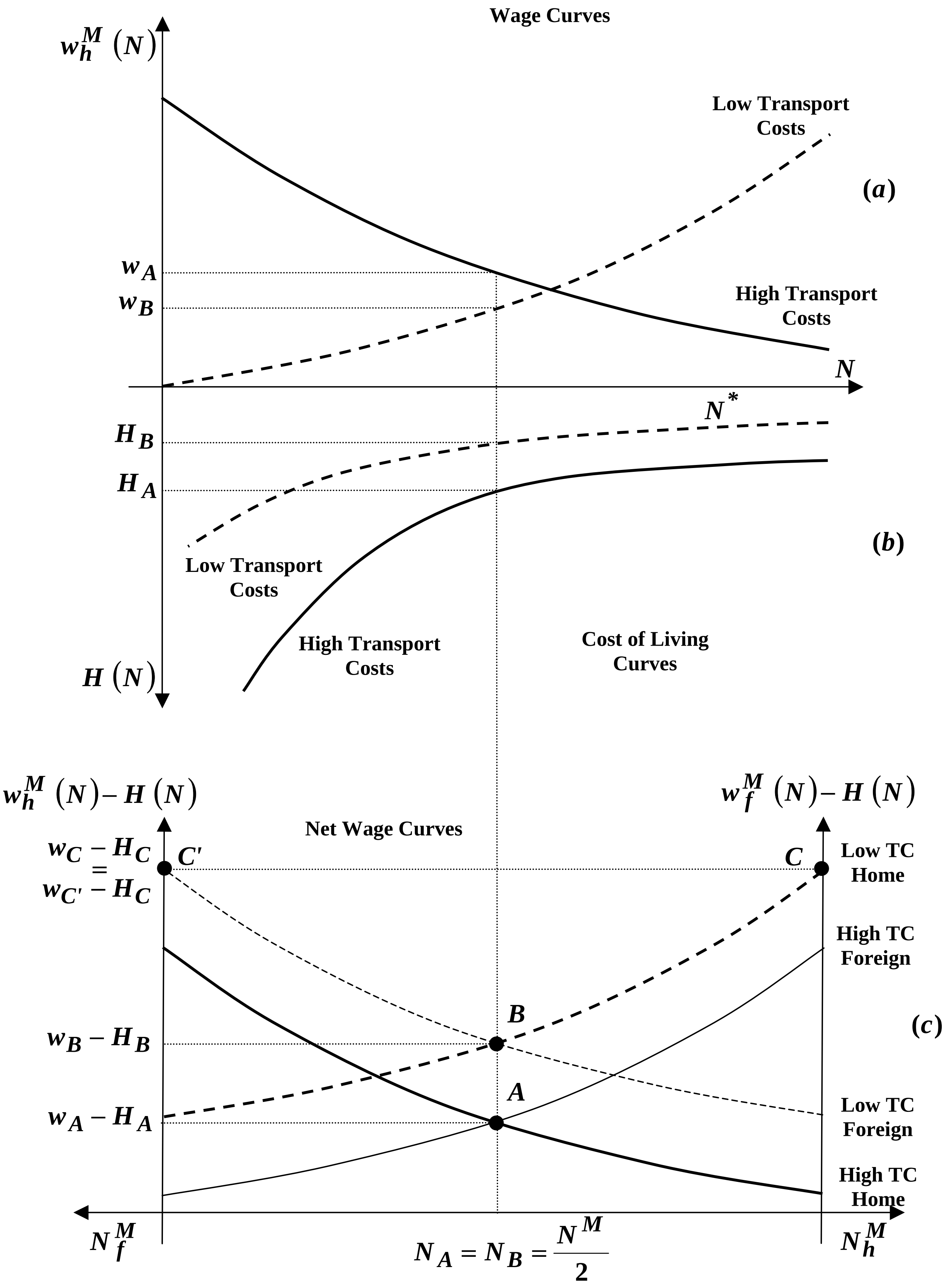

Wage Curves
$w_h^M(N)$
Low Transport Costs
(*a*)
w_A
w_B
High Transport Costs
N
N^*
H_B
H_A
(*b*)
Low Transport Costs
High Transport Costs
Cost of Living Curves
$H(N)$
$w_h^M(N) - H(N)$
$w_f^M(N) - H(N)$
Net Wage Curves
$w_C - H_C = w_{C'} - H_C$
C'
C
Low TC Home
High TC Foreign
B
(*c*)
$w_B - H_B$
A
$w_A - H_A$
Low TC Foreign
High TC Home
N_f^M
$N_A = N_B = \frac{N^M}{2}$
N_h^M

Figure 3

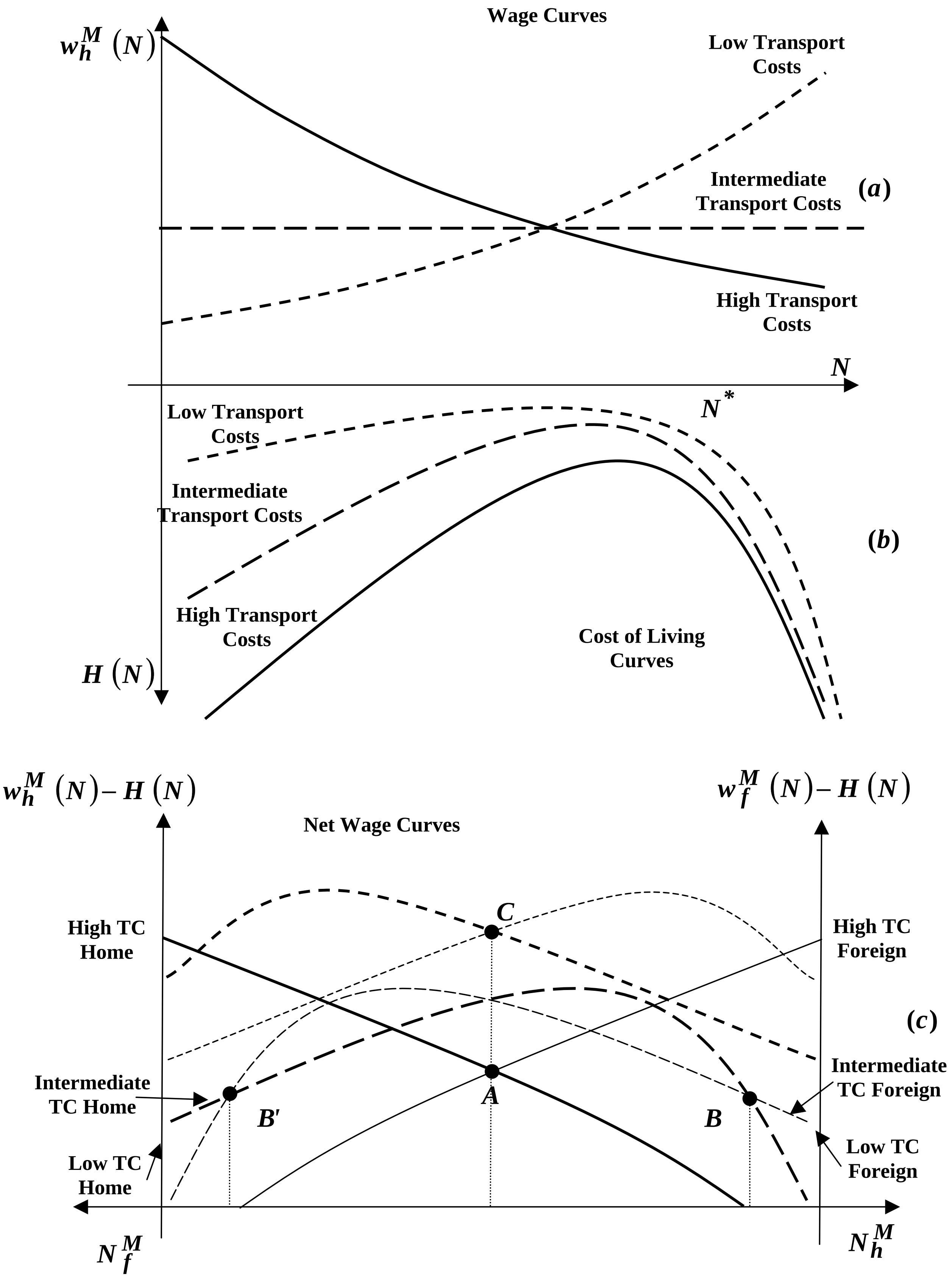

Wage Curves
$w_h^M(N)$
Low Transport Costs
Intermediate Transport Costs
(a)
High Transport Costs
N
N^*
Low Transport Costs
Intermediate Transport Costs
(b)
High Transport Costs
Cost of Living Curves
$H(N)$
$w_h^M(N) - H(N)$
$w_f^M(N) - H(N)$
Net Wage Curves
C
High TC Home
High TC Foreign
(c)
Intermediate TC Home
Intermediate TC Foreign
A
B′
B
Low TC Home
Low TC Foreign
N_f^M
N_h^M

Figure 4

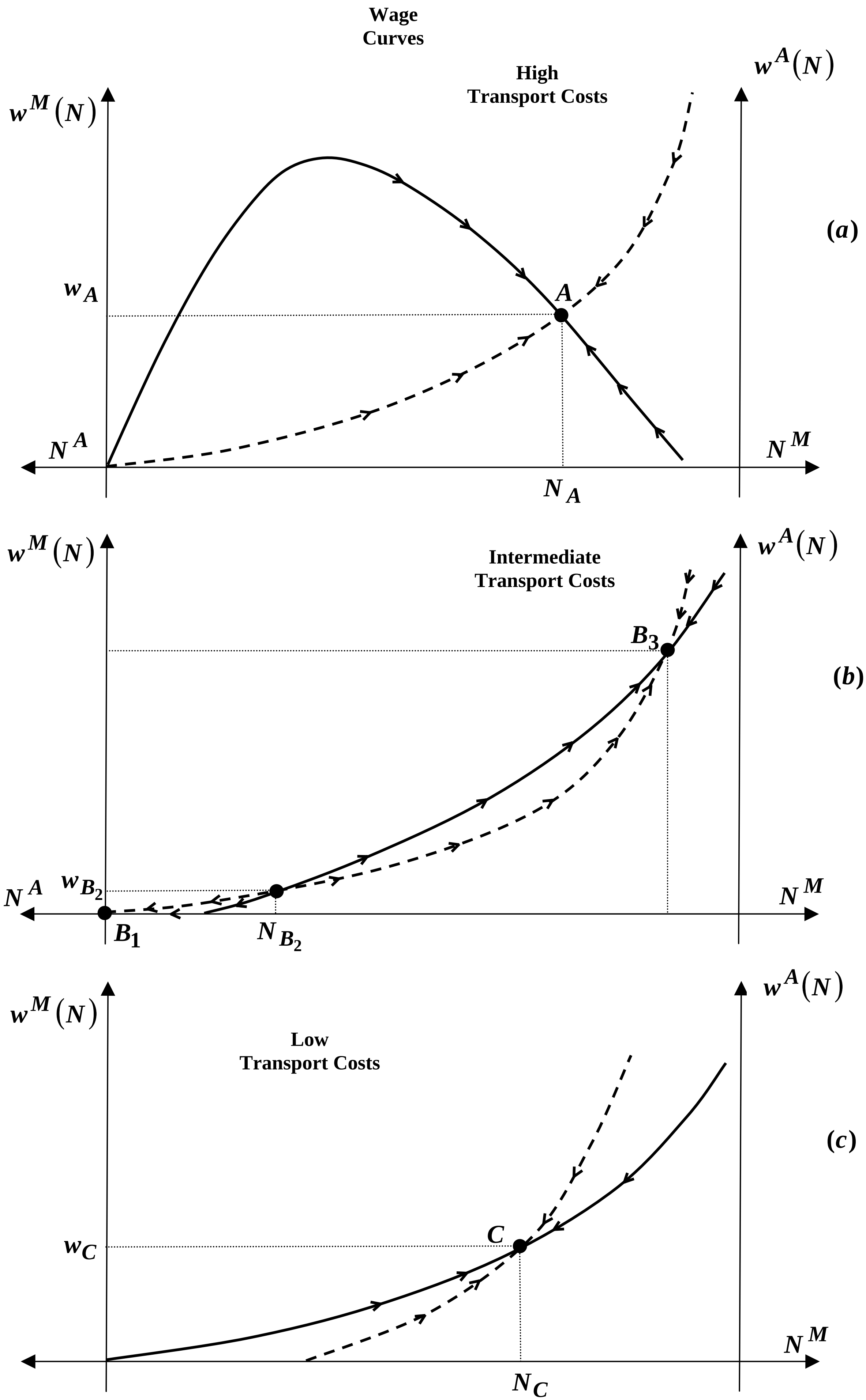

Wage Curves
$w^M(N)$
High Transport Costs
$w^A(N)$
(a)
w_A
A
N^A
N^M
N_A
$w^M(N)$
Intermediate Transport Costs
$w^A(N)$
B_3
(b)
w_{B_2}
N^A
N^M
B_1
N_{B_2}
$w^M(N)$
Low Transport Costs
$w^A(N)$
(c)
w_C
C
N^M
N_C

Figure 5

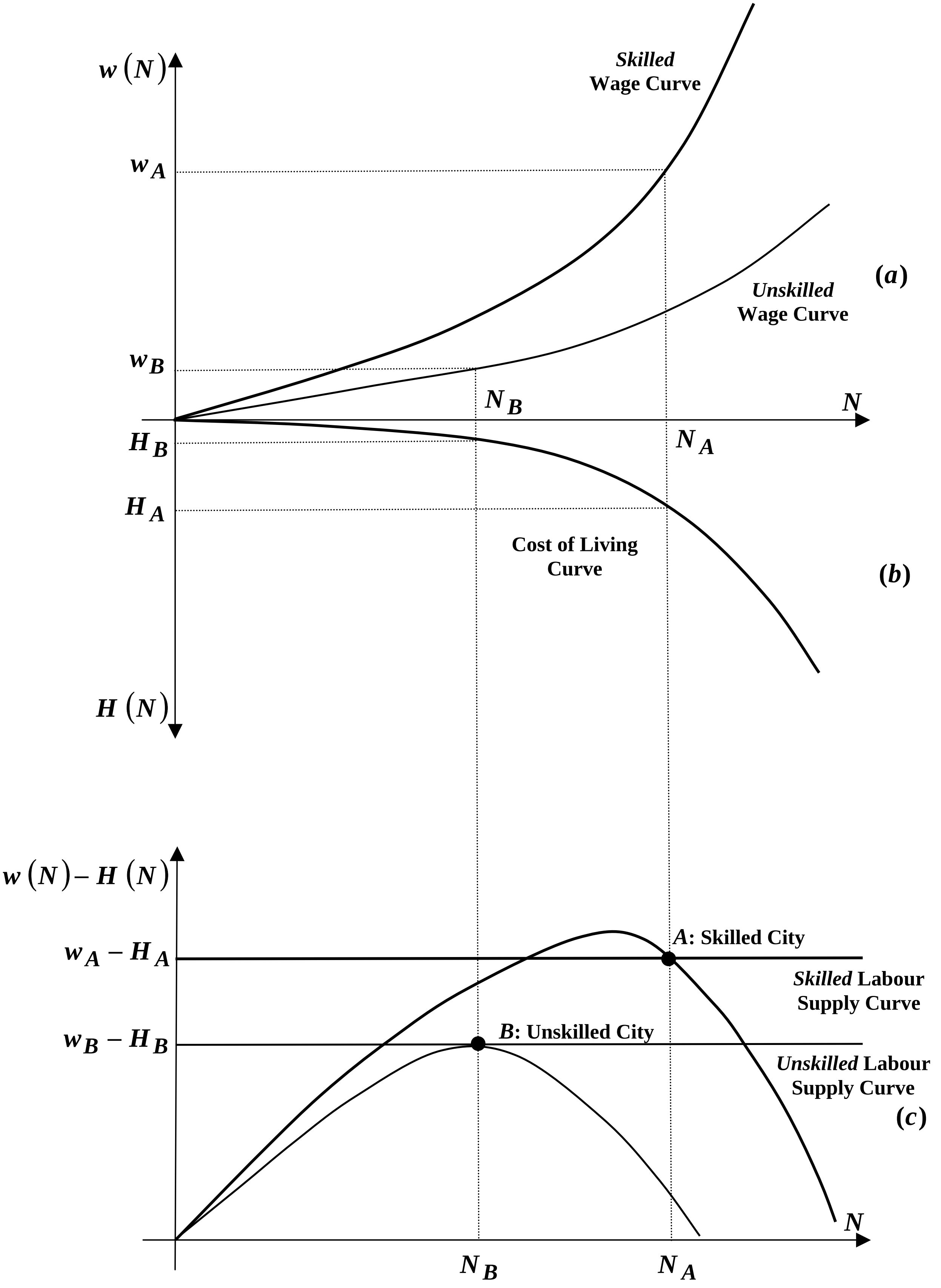

$w(N)$
w_A
w_B
Skilled
Wage Curve
Unskilled
Wage Curve
(*a*)
N_B
N
N_A
H_B
H_A
Cost of Living
Curve
(*b*)
$H(N)$
$w(N) - H(N)$
$w_A - H_A$
$w_B - H_B$
A: Skilled City
Skilled Labour
Supply Curve
B: Unskilled City
Unskilled Labour
Supply Curve
(*c*)
N
N_B
N_A

Figure 6

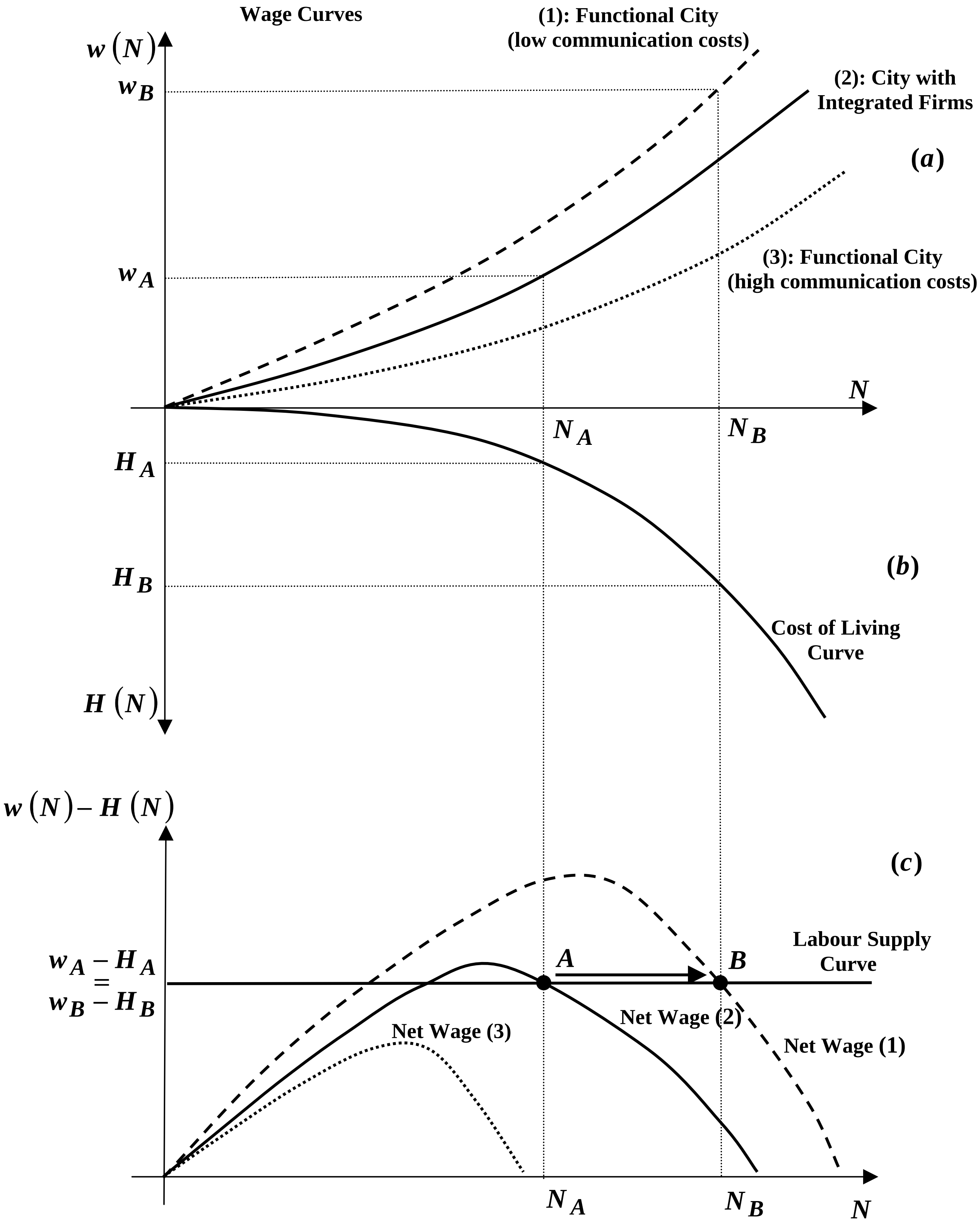

Wage Curves
(1): Functional City
(low communication costs)
(2): City with
Integrated Firms
(3): Functional City
(high communication costs)
(*a*)
$w(N)$
w_B
w_A
N
N_A
N_B
H_A
H_B
(*b*)
Cost of Living
Curve
$H(N)$
$w(N) - H(N)$
(*c*)
Labour Supply
Curve
$w_A - H_A = w_B - H_B$
A
B
Net Wage (3)
Net Wage (2)
Net Wage (1)
N_A
N_B
N

Figure 7

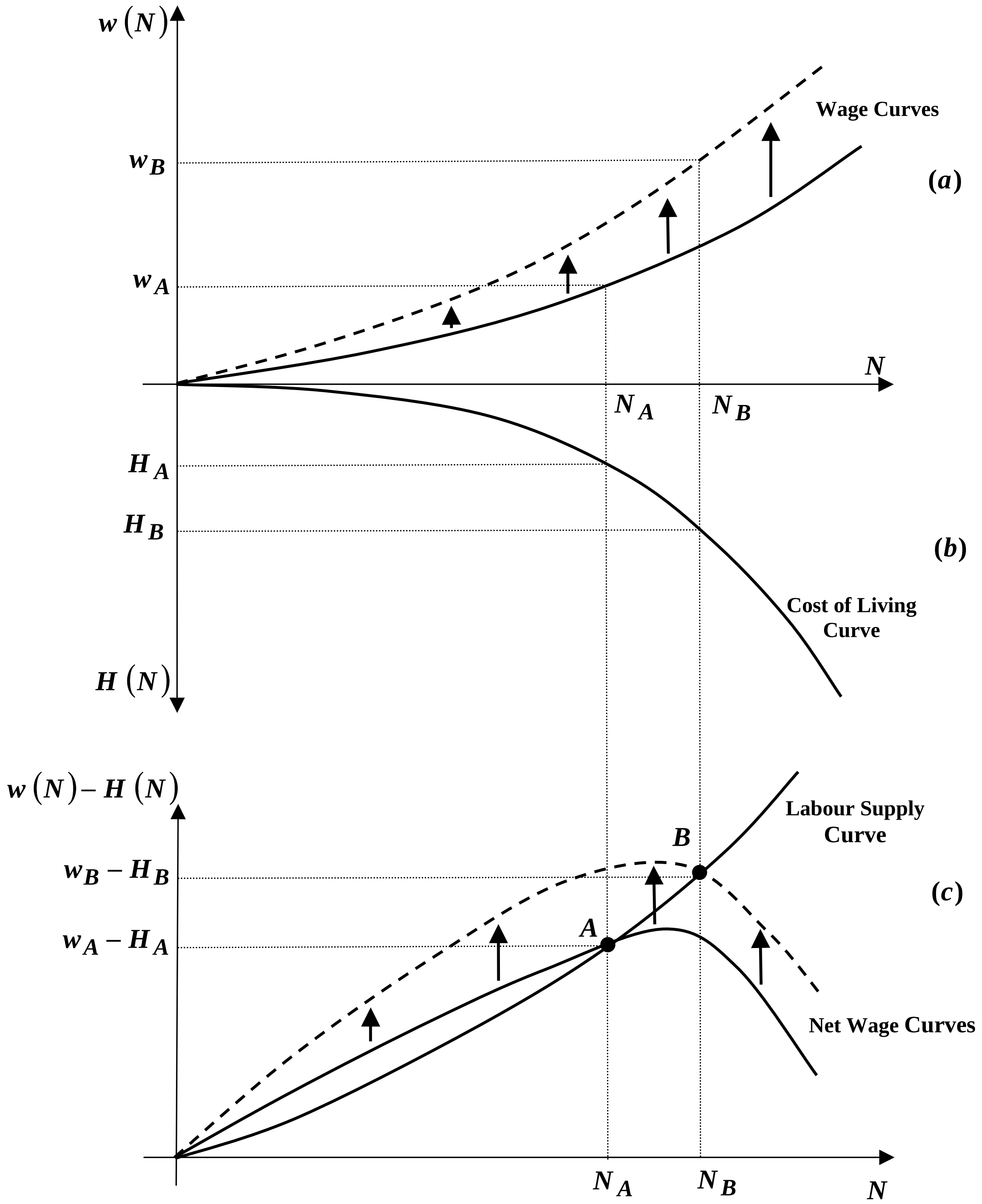
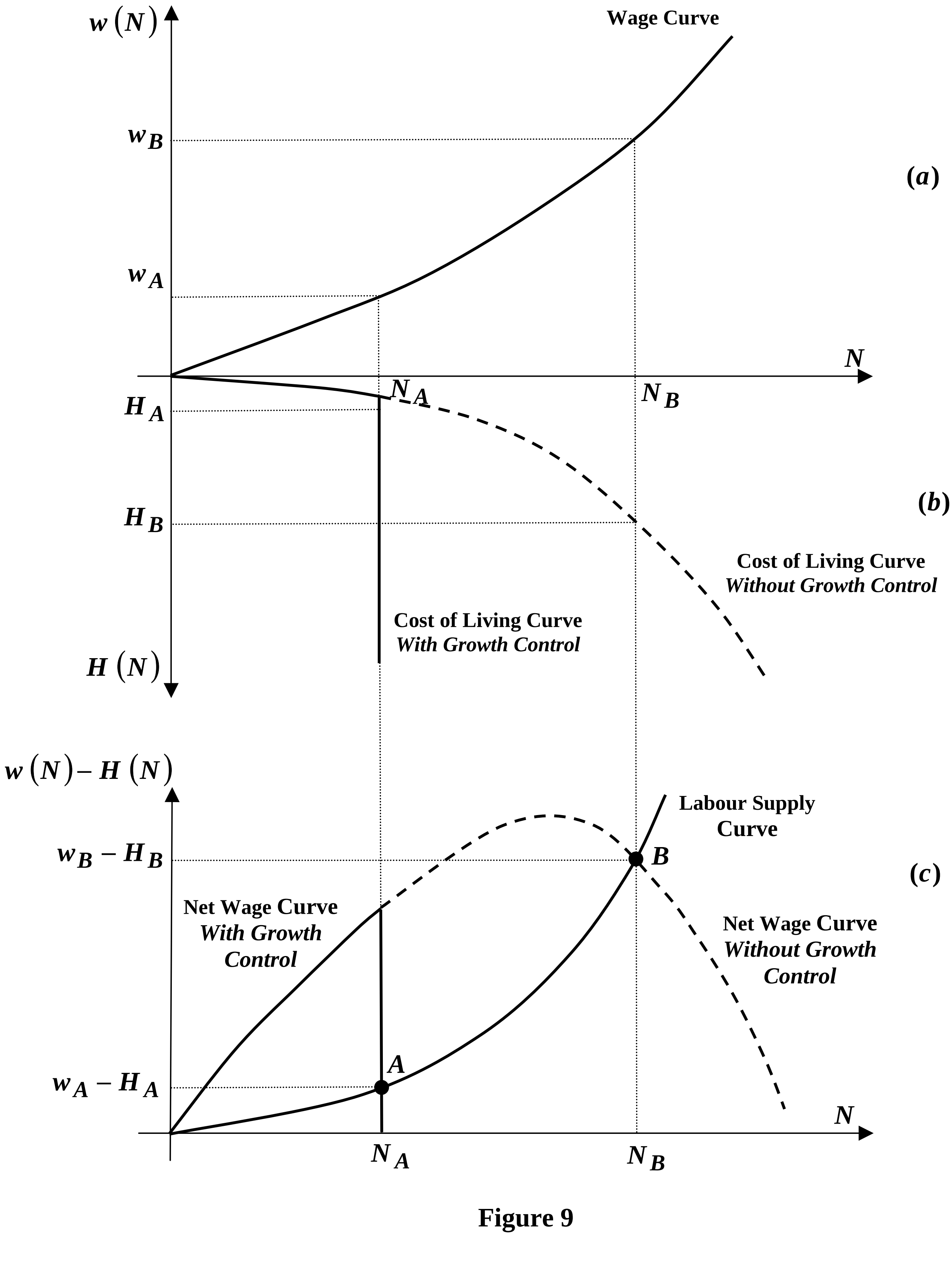

$w(N)$
Wage Curves
w_B
(a)
w_A
N
N_A
N_B
H_A
H_B
(b)
Cost of Living Curve
$H(N)$
$w(N) - H(N)$
Labour Supply Curve
B
$w_B - H_B$
(c)
A
$w_A - H_A$
Net Wage Curves
N_A
N_B
N

Figure 8



$w(N)$
Wage Curve
w_B
(a)
w_A
N
N_A
N_B
H_A
H_B
(b)
Cost of Living Curve
Without Growth Control
Cost of Living Curve
With Growth Control
$H(N)$
$w(N) - H(N)$
Labour Supply
Curve
$w_B - H_B$
B
(c)
Net Wage Curve
With Growth
Control
Net Wage Curve
Without Growth
Control
A
$w_A - H_A$
N
N_A
N_B

Figure 9

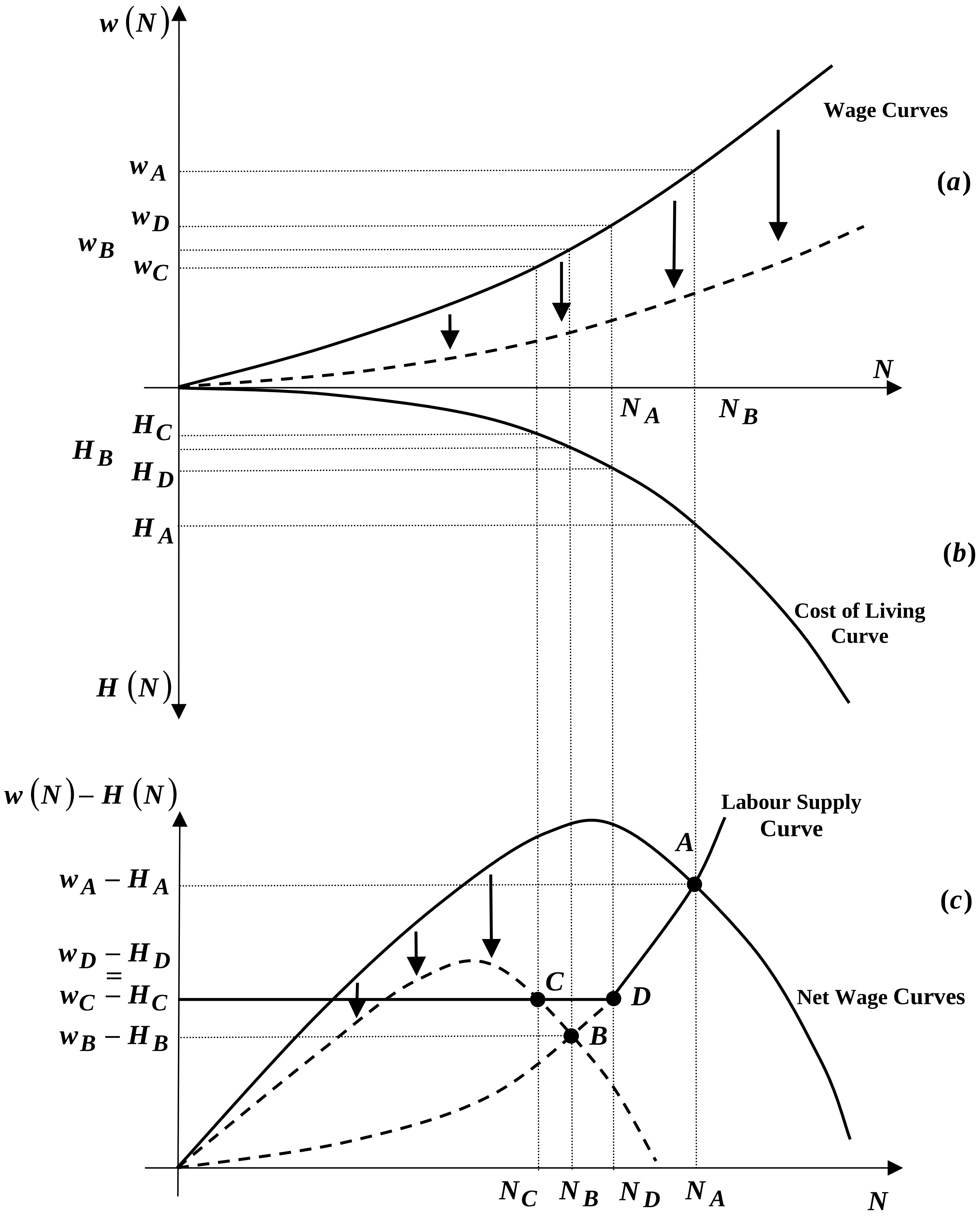

$w(N)$
Wage Curves
w_A
(a)
w_D
w_B
w_C
N
N_A
N_B
H_C
H_B
H_D
H_A
(b)
Cost of Living Curve
$H(N)$
$w(N) - H(N)$
Labour Supply Curve
A
$w_A - H_A$
(c)
$w_D - H_D$
$=$
$w_C - H_C$
C
D
Net Wage Curves
$w_B - H_B$
B
N_C
N_B
N_D
N_A
N

Figure 10

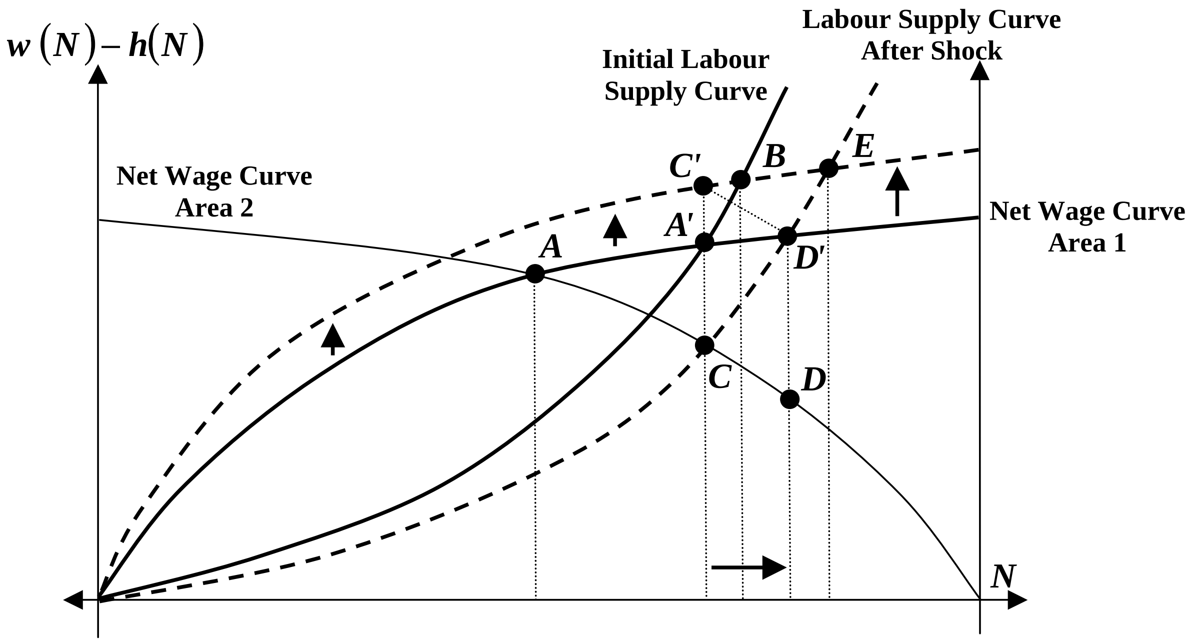

$w(N) - h(N)$
Labour Supply Curve After Shock
Initial Labour Supply Curve
Net Wage Curve Area 2
Net Wage Curve Area 1
C′
B
E
A′
D′
A
C
D
N

Figure A1