Chapter 52

THEORIES OF SYSTEMS OF CITIES

HESHAM M. ABDEL-RAHMAN

University of New Orleans, USA
e-mail: habdelra@uno.edu

ALEX ANAS

State University of New York at Buffalo, USA
e-mail: alexanas@buffalo.edu

Contents

Handbook of Regional and Urban Economics, Volume 4. Edited by J.V. Henderson and J.F. Thisse

DOI: 10.1016/S0169-7218(04)07052-2

Abstract

Economic theories of systems of cities explain why production and consumption activities are concentrated in a number of urban areas of different sizes and industrial composition rather than uniformly distributed in space. These theories have been successively influenced by four paradigms: (i) conventional urban economics emphasizing the tension between economies due to the spatial concentration of activity and diseconomies arising from that concentration; (ii) the theory of industrial organization as it relates to inter-industry linkages and to product differentiation; (iii) the New Economic Geography which ignores land markets but emphasizes trade among cities, fixed agricultural hinterlands and the endogenous emergence of geography; (iv) the theory of endogenous economic growth. Among the issues examined are specialization versus diversification of cities in systems of cities, how city systems contribute to increasing returns in national and the global economies, the factors that determine skill distribution and income disparity between cities, the impacts of income disparity on welfare, whether population growth should cause economic activity to become more or less concentrated in urban areas, and how resources should be allocated efficiently in a system of cities. Related to the last issue, we consider models where cities are organized by local planners or developers as well as cities that self-organize by atomistic actions. A conclusion of the theoretical study of city systems is that markets fail in efficiently allocating resources across cities when certain intercity interactions are present and that a role for central planning may be necessary.

Keywords

agglomeration, cities, growth, industry structure, trade, welfare

JEL classification: H23, H41, R11, R12, R13

1. Introduction

According to a United Nations Report (1996), the World's urban population increased from 30% in 1950 to 45% in 1995 and is expected to reach 50% in 2005. In industrial countries, more people live in urban areas but the increase in urbanization has been milder, rising from 61% in 1960 to 73% in 1993. The largest cities in the world have increased from only one city with a population of over 10 million in 1950 (New York) to fifteen such cities in 1995. According to a recent *National Geographic* article, the number of such *megacities* will reach 21 in 2015 "and the number of urban areas with populations between five and ten million will shoot from 7 [in 1950] to 37. This growth will occur mainly in developing countries, those least equipped to provide transportation, housing, water, sewers. Asia and Africa, now more than two-thirds rural, will be half urban by 2025" [National Geographic (2002, p. 77), brackets ours]. While urbanists from a variety of disciplines have reacted to these rapid urbanization trends with alarm, urban economists have been patiently developing theories of city systems since the mid-1970s. These theories emphasize the benefits as well as the costs of urbanization.

Most of the non-agricultural GDP in industrialized countries is produced in urban areas and cities are where new products are developed and where virtually all of technical innovation occurs. With trade liberalization and increased globalization as exhibited by NAFTA, GATT and the ongoing expansion of the European Union, the role of national governments in international trade is becoming less important. As a result, it is increasingly convenient to view the world as a network of cities of different sizes and comparative advantages, trading with one another. Such a view of economic activity increasingly competes with traditional regional and international economics. A comprehensive theory of a system of cities is an essential component of economists' efforts to understand and model economic growth and international trade.

Modern urban economics, traced to Alonso (1964), Mills (1967) and Muth (1969) spearheaded a theory of the internal structure of a city emphasizing land markets and land use. Dixit (1973) presented perhaps the most sophisticated discussion between positive and negative scale effects of city size. The beginnings of the systems-of-cities literature within modern urban economics can be traced primarily to Henderson (1974). He synthesized this Alonso–Mills–Muth theory of a city's internal structure with the concept of Marshallian externalities [Marshall (1890)], to explain the concentration of economic activity in cities. According to Lucas (1988), cities are the engines of long-run economic growth. Such an observation is not surprising to urban economists, since under the influence of Marshall, they have long maintained that the spatial proximity of market agents in cities fosters cost savings in consumption, production, search, matching and learning, knowledge spillovers and human capital accumulation.[2] External scale economies are now crucial in the long-run endogenous growth theory [e.g.,

[2] For a review of models emphasizing the micro-foundations of spatial agglomeration and city formation using search, matching and learning see the chapter by Duranton and Puga (2004) in this volume.

Romer (1986, 1987)] and also play a key role in the theory of spatial agglomeration that is now central to all systems of cities models. Theories of city systems have been successively influenced by four paradigms:

(i) conventional urban economics emphasizing the tension between economies due to the concentration of economic activity and diseconomies arising from that spatial concentration [Mills (1967), Dixit (1973), Henderson (1974)];
(ii) the theory of industrial organization as it relates to inter-industry linkages and to product differentiation [Dixit and Stiglitz (1977), Ethier (1982)];
(iii) the theory of endogenous economic growth [Romer (1986, 1987), Lucas (1988)];
(iv) the New Economic Geography [Krugman (1991)], which ignores land markets but emphasizes trade among cities, fixed agricultural hinterlands and the endogenous emergence of geography.

An ultimate unfulfilled challenge for theories of systems of cities is to explain the number and size of cities in a national economy with given population: "*How does a specific city size distribution emerge and how does it evolve in a free market*?" and "*What is the socially optimal city size distribution and how should it evolve*?" Regularities in the size distribution of cities in both developed and developing countries have been observed. Such distributions are characterized by a hierarchical structure in which there are a relatively small number of large cities and a larger number of smaller cities, commonly described by the rank–size rule. Despite several decades of research, systems of cities theories developed to date are not rich enough to explain the empirically observed city size distributions characterized by the rank–size rule or by its variants.[3]

A second basic challenge for systems-of-cities theory is explaining the variation in industrial composition across the spectrum of city sizes and the efficiency or inefficiency of such variation. Large cities at the top of the hierarchy, like New York, London, Paris and Tokyo are characterized by a diverse industrial structure while cities all over the world tend to be more specialized, the smaller they are. Explaining this would entail theoretical models of hierarchical city size distributions with a relatively small number of large cities producing most of the goods in the economy and a larger number of smaller and more specialized cities producing only a few goods.[4] As we shall see, the existing static theory of city systems has provided a variety of insights as to why specialization versus diversification occurs. The literature has distinguished between Marshall–Arrow–Romer (MAR or localization) externalities that arise from knowledge transfers within an industry, and Jacobs (1969) (or urbanization) externalities that arise from transfers between industries. Empirical work by Glaeser et al. (1992) has claimed that cities with a diversified industrial base, hence subject to Jacobs externalities, may

[3] The rank–size rule indicates that the population of each city in a system of cities, multiplied by its population rank equals the size of the largest city. For a survey of the rank–size rule and related work, see the chapter by Gabaix and Ioannides (2004) in this volume.

[4] Lösch (1954) built descriptive models of the hierarchical distribution of economic activity over space.

grow faster than specialized cities. Henderson, Kuncoro and Turner (1995) have redressed this claim showing that MAR externalities play a bigger role in traditional industries, while Jacobs externalities are more important in modern high-technology industries. Other empirical studies suggest that labor productivity and industrial growth are positively related to the local size of the industry as well as to the industrial composition of the city in which the industry is located.[5] The theoretical literature on city systems has not yet provided an explanation as to how the rate of growth of an industry or of a city as a whole is related to city specialization or diversification.

A third challenge for theory is to explain the observed skill distribution of the labor force within and between cities in the system. In particular, large cities tend to be populated with a labor force possessing a wide variety of skills while small cities tend to be populated by a labor force with relatively specific skills. As a result observed income disparity is relatively larger in larger cities. Furthermore, this issue of income disparity within a system of cities is becoming more important to understand in an urban context given that several national studies documented that income inequality has been rising over the past two decades.[6] There has been a dramatic widening of the real gap between the wages of highly skilled and poorly skilled laborers. Thus, it is important to examine, in a theoretical framework, whether the structure of a city system and the sorting of workers among cities by skill level contributes to the observed pattern of growing wage disparities.

This chapter will survey normative as well as positive issues studied by systems of cities models. We will review the static models of city systems as well as the more recent dynamic ones. We will examine the effects of exogenous population growth as well as endogenous economic growth in a system of cities. The following are some of the questions illuminated by the literature we are surveying: What are the centripetal forces favoring concentration of economic activity in large cities versus centrifugal forces favoring dispersion of economic activity to small settlements? What determines the number and the size of cities and how does this change as population grows? What are the institutional and atomistic mechanisms by which cities are created? When do cities specialize in production and when do they diversify? When do both specialized and diversified cities coexist in a national economy? How is specialization and diversification affected by costly trade among cities? What are the factors that determine the distribution of different types of labor force in a system of cities? What are the factors that determine income disparities within an urban system and what is the impact of such income disparity on overall welfare? What are the main engines of urban growth in a system of cities? What are the sources of market failure in the allocation of resources among cities and when is central planning necessary to achieve an efficient allocation?

The organization of the chapter is as follows: Section 2 presents a simplified equilibrium model of the internal structure of a single city emphasizing the organization of

[5] For survey of these empirical finding see the chapter by Rosenthal and Strange (2004) in this volume.

[6] See Juhn, Murphy and Pierce (1993) for these trends in the U.S. and Machin (1996) for the UK among others.

land use around a Central Business District. This is necessary so that the reader has a clear understanding of what goes on inside each city in a system of cities and so that the link with traditional urban economics is explicit. In Section 3 our attention turns to the agglomeration economies that help form cities and to the optimal city size implied when such agglomeration economies are offset by the expansion of cities. Our exposition focuses on three commonly used models of urban agglomeration that we present in simple stylized form. These are the local public good model of public economics, the product variety model of industrial organization and the labor productivity model of Marshallian (or MAR or localization) externalities. We show that the three models are in fact perfectly equivalent and can be made to produce identical cities by the choice of a single parameter in each model. Section 4 focuses on the institutional mechanisms (developers or local governments) that can be used to form and sustain self-financing and optimally sized cities in a system of cities, provided the cities do not interact with each other. The two mechanisms are shown to be equivalent. These institutional mechanisms are contrasted with city formation under self-organization and atomistic defection. In Section 5 we provide a broad summary of historical developments in the literature on systems of cities and we identify the key issues that arise in the design of such models. The rest of the survey is devoted to examining a number of specific models from the literature, emphasizing primarily equilibrium analysis. In Section 6 we survey the basic static models of specialized and diversified city systems characterized by homogeneous labor and we investigate trade, economies of scope and increasing returns in such city systems, while in Section 7 we present static models of a city system with labor heterogeneity and income inequality. In Section 8 we provide a survey of main ideas from the literature on the efficient allocation of economic activity into a system of cities under the proper form of fiscal decentralization. The most important conclusion of this theoretical study may be that markets fail in efficiently allocating resources across cities when intercity interactions are present. Section 9 is about growth, focusing on the effects of exogenous population growth as well as endogenous economic growth. A handful of models of dynamic city systems are surveyed there. Section 10 concludes by outlining some of the challenges ahead.

2. Internal structure of cities[7]

We assume monocentric and circular cities. All production occurs at a central point (the Central Business District, or CBD).[8] We assume that producers do not use land. The only input is the labor supplied by the consumers living in the city. Each consumer

[7] Virtually all authors modeling city systems used the model presented here.

[8] This is a major drawback because as cities grow they can spawn secondary centers. Such a mechanism of subcenter formation within existing cities competes with the emergence of new cities. The literature we are surveying has not studied both processes in the same model. Some thoughts on the interaction of the two processes are presented in Section 10.

uses one unit of land for a residence (fixed lot size) and is endowed by a unit amount of time that he allocates between labor and commuting to the CBD. The time-cost of commuting a unit distance in both directions (round trip) is t, an exogenous constant. If a consumer picks his residence location to be r miles from the CBD, then his labor supply is $H(r) = 1 - tr$. We will use N to denote the number of consumer/laborers residing in the city and $\bar{r}$ to denote the radius of the city. Then, since lot sizes are uniformly equal to one, $\bar{r} = \sqrt{N/\pi}$.[9] The aggregate labor, H, supplied to the CBD is obtained by integrating over the residential area:

$$H = \int_0^{\bar{r}} 2\pi r H(r)\,\mathrm{d}r = N\left(1 - kN^{1/2}\right), \tag{1}$$

where $k \equiv 2t/(3\sqrt{\pi})$. Since k is a normalization of t, we can use either k or t to measure the unit commuting cost.

The principle that determines city structure with zero relocation costs is that identical residents achieve the same level of utility no matter where within the city they locate. Indirect utility is of the form $V(\mathbf{p}, I)$, where $\mathbf{p}$ is the market price vector of the traded goods in the economy available for the residents of the city and I is the disposable income of any resident that is available for buying those goods. Since all residents face the same $\mathbf{p}$, for the value of utility to be invariant with location within the city, the disposable income must be the same for each resident. Disposable income is defined as income less location costs (commuting costs plus the rent of the unit-sized lot).[10] Hence, for equal utility to hold, location costs must be invariant with residence location. We assume that rent at the edge of the city is zero, because there is no non-urban use for land. We assume also that each unit time spent commuting is valued at the wage rate, w. Then, commuting cost (to the CBD and back) as a function of residential distance is $C(r) = wtr$. Hence, the rent on land at radius $0 \leqslant r \leqslant \bar{r}$ is $R(r) = t(\bar{r} - r)w$. The location cost of any one resident is then $R(r) + wtr = wt\bar{r} = wtN^{1/2}/\sqrt{\pi}$, and *aggregate location cost* (*ALC*) is

$$ALC(N) = \frac{wtN^{3/2}}{\sqrt{\pi}}. \tag{2}$$

This is independent of r. We assume that a local city manager (local government or private city developer) collects all rents and redistributes the average rent to each city resident. The *aggregate land rent* thus shared (*ALR*) is:

$$ALR(N) = \int_0^{\bar{r}} 2\pi r R(r)\,\mathrm{d}r = \frac{wtN^{3/2}}{3\sqrt{\pi}}. \tag{3}$$

[9] The maximum possible radius for a city is $1/t$, since a consumer residing beyond that radius would spend all of his time commuting and would have no time to work. The maximum population that can be accommodated by such a city is therefore $N_{\max} = \pi/t^2$.

[10] Where appropriate later, we will also deduct taxes in defining disposable income.

Note that both $ALR(N)$ and $ALC(N)$ are increasing functions of the city's population, N. However, $ALC(N)$ rises three times faster than $ALR(N)$. The disposable income of any consumer can now be calculated as

$$I(N) = w + \frac{ALR(N)}{N} - \frac{ALC(N)}{N} = \left(1 - kN^{1/2}\right)w. \tag{4}$$

The consumer has this income to purchase all the goods that will be offered in the city system. Note that as N grows the disposable income falls because the marginal (and average) resident spends more time commuting in a larger city. Therefore, adding residents to the city reduces utility. Under this *centrifugal* condition, cities would not exist. To provide an economic incentive for adding residents and creating a city, we must add *centripetal* forces so that the agglomeration of people and firms generates benefits. The following section provides three stylized economies of scale models for doing this.

3. Urban agglomeration and optimal city size

The first model we present relies on a public good. This is funded collectively by the city residents and is the driver for the urban agglomeration. As more residents join the city, the average cost of the public good declines. The second model is based on consumer demand for a variety of products or producer demand for a variety of intermediate inputs. Such consumers or producers concentrate in a city in order to make a large local market that supports greater product or input variety. The third model is the case of a productivity increase from the concentration of labor in the same industry, also known in the literature as the black-box model of Marshallian externalities. One way to interpret this black-box model is that the productivity of each worker is enhanced by the innovative ideas freely contributed by the labor force working in close proximity.

In each of the models, the agglomeration results in an optimal city size: utility increases as residents are added until it peaks at some optimal city size when the centrifugal force of commuting cost and the centripetal agglomeration force are balanced at the margin. Thereafter, utility is reduced as more residents are added. We will show that the three models are equivalent in reduced form and imply the same indirect utility as a function of the city's population. More precisely, we will show that product variety in the second model is, in fact, a local public good as is also the aggregate labor supply in the third model. Each model also illustrates a different form of market failure and we discuss how it is corrected or rendered inconsequential. In Section 4, we will see how the Henry George Theorem is the appropriate mechanism for achieving the optimal city size.

In each case, we will write the indirect utility of the monocentric city of Section 2 as $V(\mathbf{p}, Q, I(N))$ where $\mathbf{p}$ is the vector of the market prices of the consumption goods, Q is the quantity of a local public good if one exists, and $I(N) = (1 - kN^{1/2})w$ is the disposable income of Section 2, with w the urban wage. To get closed-form solutions, we will specialize to the functionally separable form: $V(\mathbf{p}, Q, I(N)) = v(\mathbf{p})f(Q)I(N)$.

3.1. Local public good

A commonly cited reason for urban agglomeration is that consumers locate in close proximity in order to have access to a local public good that they finance collectively [Flatters, Henderson and Mieszkowski (1974), Stiglitz (1977), Arnott (1979), Arnott and Stiglitz (1979)]. We will assume that the city produces a private good, say good x, traded costlessly in the larger economy with price p_x and that this good is produced competitively under constant returns, with labor the only input in production. Hence, $w = p_x$. Suppose that the direct utility function is of the form $U = x^{\alpha} y^{\beta} f(Q)$, $\alpha + \beta = 1$, where x is the quantity consumed of the locally produced good and y is the consumption of another good imported from other cities, while Q is the aggregate expenditure on the local public good. Assume that $f(Q) = Q^{\mu}$ where $\mu > 0$. Clearly, the source of market failure in this model is the presence of the public good. This market failure is corrected by determining Q so that consumer utility is maximized. A resident pays a lump sum tax, T, to finance the public good. Hence, $Q = TN$. Indirect utility is

$$V(\mathbf{p}, Q, I(N) - T) = \alpha^{\alpha}\beta^{\beta} p_x^{-\alpha} p_y^{-\beta} \left[w\left(1 - kN^{1/2}\right) - T\right] Q^{\mu}, \quad \text{where } T = \frac{Q}{N}.$$

The lump sum tax that maximizes utility, given N, is $T^* = (\mu/(1+\mu))I(N)$.[11] The after-tax utility is

$$\begin{aligned}\widetilde{V}(\mathbf{p}, N, I(N)) &\equiv V(\mathbf{p}, NT^*, I(N) - T^*) \\ &= \left(\frac{\mu}{1+\mu}\right)^{\mu} \left(\frac{1}{1+\mu}\right) p_x^{1+\mu-\alpha} p_y^{-\beta} N^{\mu} \left(1 - kN^{1/2}\right)^{\mu+1}.\end{aligned}$$

The first-best optimal city size is found by maximizing $\widetilde{V}(\mathbf{p}, N, I(N))$ with respect to N. In the functionally separable case, this optimal size is given as a fraction of the maximum possible city size, $N_{\max}$:

$$N^* = \frac{9\mu^2}{9\mu^2 + 6\mu + 1} N_{\max}. \tag{5}$$

This shows that the optimal city size increases as the unit commuting cost (k) falls (recalling from footnote 7 that $N_{\max} = 4/(9k^2)$) and increases as μ, the elasticity of utility with respect to public expenditure, increases. When $\mu = 0$, $N^* = 0$ and cities cannot exist (there is no centripetal force). And as $\mu \to \infty$, $N^* \to N_{\max}$, the centripetal force dominates causing maximally sized cities.[12]

[11] We will see in Section 4 how this optimally determined tax confiscates all of the land rent when the city population, N, is endogenous.

[12] In a variation of the public good model, Abdel-Rahman (2000) assumed that the public good is a form of infrastructure such as water, electrical or sewer system. Investment in this public good is assumed to reduce the fixed costs of setting up firms. The larger the investment in infrastructure is, the larger the number of firms that will enter the city.

3.2. Product variety and the home market effect

Agglomeration can also be caused by consumers locating in the same city to create a large home market. If the consumers value product variety, they benefit from a large home market because more unique products will emerge and be viable in a larger market. This, in turn, will drive up consumer utility providing the basis for an even larger market. Thus, suppose that each of the N consumers has a utility function that is Dixit–Stiglitz CES [Dixit and Stiglitz (1977)] given by

$$U = \left(x_0, \sum_{i=1}^{m} x_i^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}$$

where $\sigma > 1$, x_0 is a numeraire good and m is the number of varieties manufactured in the city (x_i being the consumption of the ith such variety) and assume for now that these are not traded with the other cities.[13] In Dixit–Stiglitz, part of the disposable income of the consumer is used to buy x_0. Again following Dixit and Stiglitz, we assume that each of the m products is manufactured by a different firm, and by an increasing returns technology requiring a fixed labor input, f, plus a marginal labor input, c, per unit produced. The ith firm's total cost of producing z_i units of output is then $w(f + cz_i)$. All firms pay the same wage, w, and are perfectly competitive in the labor market. In this model, the wage is taken as the numeraire price. Hence, $w = 1$. In the output market, the firms are monopolistically competitive and achieve a Chamberlin equilibrium [Chamberlin (1933)]. This imperfect competition in the output market is the basis of the market failure. Because the consumer's demands for the products are symmetrical and there is also symmetry among the many firms, the Cournot–Nash markup condition is approximately $p_i(1 - 1/\sigma) = c$, where σ is the price elasticity of the consumer's demand for each product. Hence, the equilibrium product prices are $p = (\sigma/(\sigma - 1))c$. This monopolistic markup in the differentiated goods industry is the cause of the market failure which comes about because the allocation of income between the m varieties and the numeraire good is distorted. To simplify things, we adopt a version of the Dixit–Stiglitz utility without the numeraire good. In this version, income can only be spent on the varieties and, as a result, there is no distortion and the market failure is inconsequential. The indirect utility is

$$V(\mathbf{p}, I(N)) = \left(\sum_{i=1}^{m} p_i^{1-\sigma} \right)^{1/(\sigma-1)} I(N).$$

[13] Alternatively, firms utilizing differentiated services as inputs are motivated to locate in the same city in order to create a large home market and thus increase the variety of services available to them. Krugman (1980, 1991) and Abdel-Rahman (1988) have emphasized models in which the consumer cares about consumption variety, while Abdel-Rahman and Fujita (1993) and Anas and Xiong (2003) have emphasized models in which firms producing a homogeneous product demand a variety of intermediate inputs. The former models use the Dixit and Stiglitz (1977) utility function, while the latter use the Ethier (1982) production function.

From the zero profit condition of each firm, $z = f(\sigma - 1)/c$. The aggregate labor demand is $m(f + cz) = mf\sigma$. Since the labor supply in the city is given by (1), the number of products at equilibrium is

$$m(N) = \frac{N(1 - kN^{1/2})}{f\sigma}.$$

Using this to evaluate the utility,

$$V(\mathbf{p}, m(N), I(N)) = \left(m(N)p^{1-\sigma}\right)^{1/(\sigma-1)} I(N) = p^{-1}m(N)^{1/(\sigma-1)}\left(1 - kN^{1/2}\right)$$
$$= \left(\frac{\sigma - 1}{\sigma c}\right)(f\sigma)^{1/(1-\sigma)} N^{1/(\sigma-1)}\left(1 - kN^{1/2}\right)^{\sigma/(\sigma-1)}.$$

And

$$\widetilde{V}(\mathbf{p}, N, I(N)) \equiv \ln V(\mathbf{p}, m(N), I(N)) = -\ln p + \frac{1}{\sigma - 1}\ln m(N) + \ln I(N)$$

under the assumed symmetry of the varieties. This last expression, borrowed from Anas (2004) allows us to see that the consumer cares about disposable income, which decreases with city size, and the number of products that increase with city size. Only as $\sigma \to \infty$, products are viewed as perfect substitutes and variety becomes unimportant. Now define $\omega \equiv 1/(\sigma - 1)$ and note that $\omega + 1 = \sigma/(\sigma - 1)$. The optimal city size is again given by (5) where μ is now replaced by ω. As $\sigma \to \infty$ and $\omega \to 0$, all varieties are perceived as perfect substitutes. Then, N^* will be the size of a company town producing a single variety, since there is no benefit to be derived from a large home market. But as $\sigma \to 1$ and $\omega \to \infty$, $N^* \to N_{\max}$.

The consumer benefits from a larger number of goods (firms) but the fixed cost limits the number of firms at equilibrium, causing the firms to price their products above marginal cost. Hence, in Dixit and Stiglitz (1977), there is a collective incentive to tax the consumers and use the proceeds to subsidize the fixed costs of firms. This is formally equivalent to taxing consumers to pay for the public good in Section 3.1. To see this equivalence, recall that the indirect utility in the public goods model was

$$V(\mathbf{p}, Q, I(N) - T) = v(\mathbf{p})Q^{\mu}\left[\left(1 - kN^{1/2}\right) - T\right],$$

where T was the head tax so that $NT = Q$. In the current model, the number of firms (goods variety) may be viewed as a public good with after-tax utility

$$V(\mathbf{p}, m, I(N) - \Theta) = v(\mathbf{p})m^{\omega}\left[\left(1 - kN^{1/2}\right) - \Theta\right], \quad \text{where } m = N\Theta/f.$$

Maximizing this with respect to Θ, $\Theta^* = (\omega/(1 + \omega))/(1 - kN^{1/2})$, where $\omega = 1/(\sigma - 1)$. This is identical to

$$T^* = \frac{\mu}{1 + \mu}\left(1 - kN^{1/2}\right)$$

in the public good model. To complete the argument, we just need to show that the labor market clears under Θ^*. This requires

$$m = \frac{H(N)}{f\sigma} = \frac{N(1 - kN^{1/2})}{f\sigma},$$

which was shown earlier to follow from labor market clearing in the absence of any tax. That the tax is not needed to achieve the efficient allocation is a reflection of the fact that the market failure is rendered inconsequential by the absence of the numeraire good of Dixit–Stiglitz. The market allocation is efficient for a given N by laissez-faire (i.e., the efficient m is provided by the market), despite the monopolistic markup. This contrasts with the public good model of Section 3.1 where, as is well known, there is no market mechanism that can guarantee the efficient provision of the public good according to the Samuelson condition.

3.3. Labor productivity as an external economy

Suppose that aggregate labor, H, enters the CBD production function via two channels. First, labor is the only input purchased by each firm to produce a numeraire good under a constant-returns technology.[14] Second, it is the source of a Marshallian external effect on all firms, so that the more labor is employed in the CBD, the more productive the labor employed at each firm becomes. This is a classical Chipman (1970) type externality compatible with competitive equilibrium. The production function of firm i is $y_i = A(H)h_i$ where h_i is the labor employed by that firm, and H is the aggregate labor employed by all firms in the CBD. Because of constant returns, the aggregate CBD production function can be written as $Y = A(H)H$, where the external scale effect, $A(H)$ with $A'(H) > 0$, is the marginal and average product of labor and aggregate labor supply, $H = H(N)$, is given by (1). Let us assume that $A(H) = H^a$, $a > 0$, namely that the positive external effect increases with the total amount of labor's time devoted to work. Let us assume that the CBD labor market is competitive. Then, labor is paid its private marginal product and $w = A(H) = H^a = N^a(1 - kN^{1/2})^a$. Then, the indirect utility in the functionally separable case is

$$\widetilde{V}\big(\mathbf{p}, N, I(N)\big) \equiv V\big(\mathbf{p}, I(N)\big) = v(\mathbf{p})N^a\big(1 - kN^{1/2}\big)^{a+1}.$$

The city size, measured in population N at which this utility is maximized, is again given by (5) but now μ is replaced by a. When $a = 0$, there are no external returns to scale and $N^* = 0$. But as $a \to \infty$, $N^* \to N_{\max}$.

Market failure in this model arises from the fact that individual firms do not have an incentive to reward labor for the positive externality it confers on production. More precisely, the social marginal product (*SMP*) of labor is $A'(H)H + A(H) > A(H)$,

[14] In this model there is no explicit public good, so we will assume that $f(Q) = 1$ or equivalently that $V_Q = 0$.

where $A(H)$ is private marginal or average product (*PMP*). $SMP - PMP = A'(H)H$ and if producers were to cover the gap, they would incur losses. The income needed to cover the gap must be derived from another source and, as we shall see in Section 4, this source in the first-best case is the aggregate land rent. Assume for now that labor is paid its *SMP* from any source, $w = A'(H)H + A(H) = (1+a)H^a$. The functionally separable indirect utility derived above is now just multiplied by $1+a$ and, hence, the optimal city size, N^*, remains unaffected by the market failure. This model is also reduced-form-equivalent to the previous two. In this case, the city's labor force itself is equivalent to a public good. Write the after-tax indirect utility as $V = H^a(1 - kN^{1/2} - G)$ where G is the head tax and impose the constraint $GN - sH = 0$, where s is the subsidy per unit of labor. Optimizing with respect to G, given N and s, we get

$$G^* = \frac{a}{1+a}\left(1 - kN^{1/2}\right).$$

3.4. Discussion and notes on the literature

We saw that, in each of the models discussed, the reduced form indirect utility is of the form $\widetilde{V}(\mathbf{p}, N, I(N))$ with $\partial\widetilde{V}/\partial N > 0$ given $I(N)$, $\partial\widetilde{V}/\partial I(N) > 0$ and $\partial I(N)/\partial N < 0$. At the optimal size, N^*, the marginal agglomeration benefit from adding one more resident to the city just equals the marginal disutility from the increase in aggregate location cost:

$$\left.\frac{\partial\widetilde{V}(\cdot)}{\partial N}\right|_{N=N^*} = \left(\frac{\partial\widetilde{V}(\cdot)}{\partial I(N)}\right)\left(-\frac{\partial I(N)}{\partial N}\right)\Bigg|_{N=N^*}.$$

There is an important difference between the public good model of Section 3.1 on the one hand, and the product variety and labor productivity models of Sections 3.2 and 3.3, on the other. As is well known, given population N, there is no market structure that guarantees the optimal provision of the public good, Q, in Section 3.1. In Section 3.2, as we saw, laissez-faire provides the optimal diversity, m, given N, despite the markup imperfection, provided that there is no numeraire good as in Dixit and Stiglitz (1977). Likewise, in Section 3.3, the laissez-faire allocation is also efficient given N, provided that there is full employment.

The above three models of city formation have been used widely, but they are not the only ones. Helsley and Strange (1990) developed an alternative model with micro-economic underpinnings different from those discussed above. In their model, both the firms producing in the CBD as well as the workers/consumers hired by the firms are horizontally differentiated on the same unit circle. For a firm, its position on the unit circle identifies the firm's skill requirements in the labor market and for a worker, position on the unit circle identifies the skills that worker possesses. Workers are more productive when they work for firms with the skill requirements that best match theirs. Firms know only the expected skills of the workers they hire. As more firms enter the CBD, firm density on the unit circle increases, the labor market becomes thicker and workers are better matched to firms. Hence, productivity increases. Firms have fixed costs, so that

each firm must command a finite range of laborer skills and the unit circle does not fill up with firms. Wages and profits are determined by a simple bargaining mechanism.

4. City formation mechanisms

Each of the models discussed in Section 3 calls for collective action to setup a city. In the first model, a tax was needed to finance the public good, while in the second (with the numeraire good present) the tax would be needed to subsidize firm entry costs in order to increase the city's product variety. In the third model, the gap between the social and private marginal products of labor would be closed by a Pigouvian subsidy to labor. In this section we will examine the relationship between the public expenditure, the tax structure and the population size at the optimum.

We will consider two planned city formation mechanisms. In the first, the city is set up and managed by a utility-maximizing local government representing the community of the city's residents. The government also decides the population of the city. In the second mechanism, there is a city developer who maximizes profit from city development, but must compete for residents in a national market. To attract residents to his city, the developer must be sure to setup and finance the city in such a way that residents moving into the city (recall that moving costs are zero) cannot do worse than achieving a reservation utility level. These two mechanisms will be shown to be equivalent under the assumption that the city-development market is contestable and that there is no limit to the number of cities.[15]

4.1. Community planning: welfare maximization

It is most natural to consider the problem of city formation in the context of the public good model presented in Section 3.1.[16] Suppose that a local planner or government sets a head tax, T, to be paid by each resident joining the city and uses the aggregate tax revenue to pay for the public good. Thus, $TN - Q = 0$. The government maximizes the utility of the representative city resident. The government will determine Q, N and T while land rents, $\mathbf{p}$ and w are determined by the markets. The problem of the local government is to solve:

$$\text{Maximize}_{Q,N,T}\, U = V(\mathbf{p}, Q, I(N)) \quad \text{subject to:}$$

$$NT - Q = 0,$$

[15] The assumption that there is no limit to the number of cities is not realistic because space on the earth or in any given national territory is limited. However, all systems-of-cities models except those in the tradition of the New Economic Geography ignore this feature.

[16] Not only is the exposition of the public good model easier, but indirectly the discussion also applies to the models of Sections 3.2 and 3.3 since we showed that the number of firms producing varieties or the labor force are implicitly public goods.

$$I(N) - w + T + \frac{ALC(N)}{N} - \frac{ALR(N)}{N} = 0.$$

We may optimize after substituting $I(N)$ into the utility and then forming the Lagrangian, $\Im$, with Lagrangian multiplier λ assigned to the tax budget constraint. The first-order conditions with respect to Q, T and N are respectively as follows:

$$\Im_Q: \quad \frac{\partial V}{\partial Q} - \lambda = 0;$$

$$\Im_T: \quad -\frac{\partial V}{\partial I} + \lambda N = 0;$$

$$\Im_N: \quad \frac{\partial V}{\partial I}\frac{\partial I}{\partial N} + \lambda T = 0.$$

Solving the first for λ and plugging that into the second, we get

$$N\frac{\partial V/\partial Q}{\partial V/\partial I} = 1. \tag{6}$$

Given N and T, this condition – known as the Samuelson condition – can be solved for the optimal level of the public expenditure, $Q^*(T, N)$. The right side is the marginal cost of the public good measured as a dollar of public expenditure, while the left side is the aggregate willingness to pay for a dollar of public expenditure, measured as the marginal rate of substitution between the public good and after-tax disposable income, summed over all the N residents. Next, we can substitute for λ and for $T = Q/N$ in the third constraint. The resulting condition is

$$Q = \left[\frac{-\partial I/\partial N}{\frac{\partial V/\partial Q}{\partial V/\partial I}}\right] N. \tag{7}$$

Given, Q and T, this equation can be solved for the optimal population $N^*(Q, T)$.

Two important facts are revealed by (7). First, note that the bracket evaluated at the optimal solution gives the optimal tax, T^*. Note that the denominator inside the bracket is the marginal rate of substitution and, from the Samuelson condition, (6), it equals $1/N$. Then the bracket can also be written as $[\cdot] = (-\partial I/\partial N)N$. The parenthesis is the marginal cost imposed on each city resident from the addition of a new resident. This marginal cost is expressed as a decrease in after-tax disposable income due to increased rent plus commuting cost net of redistributed (average) land rent. Multiplying by N gives the after-tax disposable income reduction over *all* residents caused by the addition of the marginal resident to the city. The first important fact, therefore, is that each resident is levied a Pigouvian tax equal to the social marginal cost the resident imposes on the city.

The second important fact is that the tax imposed on each resident is equal to the average land rent in the city. This can be seen directly by evaluating the derivatives and

thus showing that

$$T^* = [\cdot] = \frac{ALR(N^*)}{N^*} = \frac{k}{2}(N^*)^{1/2}.$$

Hence, the optimal Pigouvian tax summed over all residents confiscates the entire land value of the city. This result, in turn, is known as the Henry George Theorem: optimizing welfare with respect to the number of residents in the city, N^*, requires levying a Pigouvian tax, T^*, on each resident the proceeds of which over all residents confiscates the city's aggregate land value, $ALR(N^*)$ – hence, causing no distortion – and using these proceeds solely to finance the optimal quantity of the public good, Q^*, implied by the Samuleson condition.

4.2. *City development*

In the foregoing, the community planner sets the population of the city without considering the alternatives which residents can obtain by going elsewhere. The planner must make sure that the maximum utility which residents can obtain by joining his city is higher than what they can get elsewhere and that there are enough potential residents available to draw from. This model obviously ignores the possibility of competition from other community planners setting up similar cities. It has been argued that with no limit on the number of cities that can emerge, competition among cities – under the assumption of costless consumer mobility among cities – would imply that the residents of all cities must be at the same utility level. Such an observation has led to a stylized model of city development in which each city is set up and managed by a profit-maximizing developer who is a utility taker. It is assumed that the developer owns all the land the city will need (having bought such a land from alternative users that do not suspect that a city will replace them) and then proceeding to sell each parcel to the highest bidder. Because this is the behavior of a perfectly price discriminating monopolist, it is immediately obvious that it is efficient and, hence, it is equivalent to assuming that the developer lets the land market for residences operate competitively. The developer pays for the public good out of the land rental proceeds. The developer chooses the level of public expenditure and the number of residents in such a way that each resident will not do worse than the national reservation utility level. If cities are set up and city developers make a profit, then new developers will enter the city-development market and will set up competing new cities. Such competition will cause all rents on land being invested in the local public good and developers making zero profits.

Suppose that $\overline{U}$ is the exogenous national reservation utility level. The developer of a representative city in the public good case must solve the following problem:

$$\text{Maximize}_{Q,T,N}\Pi = TN - Q \quad \text{subject to}$$

$$V\left(Q, w - T - \frac{ALC(N)}{N} + \frac{ALR(N)}{N}\right) - \overline{U} = 0.$$

The first-order condition with respect to Q, T and N are as follows, where μ is the Lagrangian multiplier of the utility constraint:

$$\Im_Q: \quad \frac{\partial V}{\partial Q} - \frac{1}{\mu} = 0;$$

$$\Im_T: \quad -\frac{\partial V}{\partial I} + \frac{1}{\mu} N = 0;$$

$$\Im_N: \quad \frac{\partial V}{\partial I}\frac{\partial I}{\partial N} + \frac{1}{\mu} T = 0.$$

These are identical with the corresponding first-order conditions of the community planning (Section 4.1) problem. The two problems yield an identical optimizing solution Q^*, T^*, N^* with $1/\mu^* = \lambda^*$ if $\overline{U}$ in the city-development problem equals U^* in the community welfare maximization problem.

We now turn to the labor productivity problem of Section 3.3. In this case the source of the agglomeration economy is the external returns to scale in the CBD production. A city developer solves:

$$\text{Maximize}_{N,S} \Pi = ALR(N) - SN \quad \text{subject to}$$

$$A\big(H(N)\big) + S - \frac{ALC(N)}{N} = \hat{I},$$

where $H(N)$ is the aggregate labor supply given by (1). The developer rents out all the land in the land market, organized competitively, collecting aggregate rent $ALR(N)$ and paying out a subsidy, S, to each resident. Because workers are freely mobile between cities, the developer must set the subsidy so that workers get the national disposable income, $\hat{I}$, that prevails in other cities set up by competing developers. $A(H(N))$ is the average product and the private income of the worker and $ALC(N)/N$ is the location cost of a worker in a city with N workers. Solving for S from the constraint and plugging the result into the objective and simplifying using (2) and (3), it becomes

$$\Pi = A\big(H(N)\big)N - kN^{3/2} - \hat{I}N,$$

where the second term is the aggregate commuting cost. This suggests that we can alternatively view the developer as owning all production, collecting all rents, selling the output in the national market, paying all commuting costs (the second term) and paying each worker $\hat{I}$. With many competing developers setting identical cities, the developer makes zero profit. The first-order condition together with zero profit gives

$$S^* = \frac{1}{2} k N^{*1/2} = N^* A'\big(H\big(N^*\big)\big)H'\big(N^*\big).$$

This says that the subsidy per person equals the per capita rent which, in turn, equals the difference between the social and average product of the worker. Hence, once again, the Pigouvian subsidy confiscates the total land rent, causing no distortion.

The efficiency results provided above are a direct consequence of contestability in the city formation market. Hence, the absence of large sunk costs is crucial. But if sunk costs required for city formation are large and developers are not utility takers, Helsley and Strange (1994) showed that in a game-theoretic framework with a fixed finite number of cities, the system of cities will be inefficient. This inefficiency is due to the strategic interaction between developers [Scotchmer (1986)]). In another strategic community development model, Henderson and Thisse (2001) examined the equilibrium formation of communities in which the number of communities is endogenous and the communities are differentiated by the income of their residents. Profit-maximizing developers that interact strategically choose the level of provision of a public good, an entry fee and a unit price for housing to attract households with different incomes to their community. The internal spatial structure of the communities is ignored.

4.3. Self-organization by atomistic defection: are developers needed?

Collective mechanisms of city formation relying on developers or local governments are not entirely realistic. An alternative view is that cities are formed, develop, evolve not by centralized action but by the atomistic decisions of consumers and firms. Under this view, consumers are assumed to be myopic and to act independently. They move to an alternative city when they can get a higher utility there. This alternative paradigm, known also as *self-organization*, is more realistic and gives rise to a variety of positive and normative questions about city formation. In the literature, Henderson has repeatedly emphasized the role of developers setting up cities, especially in the United States. On the other hand, Krugman's New Economic Geography has emphasized the formation of cities by a cumulative process caused by the defection of economic agents from other places (including from cities established earlier). The view that developers set up cities is an idealization. In the U.S. developers sometimes play an important role in setting up suburban subdivisions and in some cases whole towns or edge cities. But once these are set up, the developers sell and additional growth happens by atomistic defections. Local governments take over from initial developers but do not behave like the community welfare maximizers of Section 4.1.

It is important to identify the conditions under which developers are needed for cities to emerge, to emerge at the optimal time and to guarantee that cities have the optimal industrial composition. Autarkic optimal size is given by Equation (5) or one of its variants discussed in Section 3. Efficiency of the city system occurs when all cities are so sized. Then, shifting a resident from one city to another reduces utility in both cities. Hence, such an efficient equilibrium is also stable. There is strong agreement in the literature that this efficient and stable outcome cannot be guaranteed by atomistic defection. Henderson (1974), Anas (1992), Pines (2000) and Henderson and Becker (2000) have all presented analyses in which, under self-organization, existing cities get grossly overpopulated before new cities emerge.

To see the reason why, in the case of symmetric cities, cities emerge too late, suppose that a total population P must be divided between two city populations[17] N_1, N_2 such that $N_1 + N_2 = P$. The one-city possibilities of $N_1 = P$, $N_2 = 0$ and $N_1 = 0$, $N_2 = P$ are allowed. Assume that the two cities do not interact with each other and that each consumer is also a firm producing and consuming his own output. Cities can exist because the productivity of each consumer increases with the population of the city (the model of Section 3.3 applies). Pretend that the only action of the city government is to collect and redistribute land rent to its residents. Equilibrium requires consumers deciding in which city to locate and moving freely from the lower utility city to the higher utility one, and thus equalizing utilities. Clearly, the symmetric allocation, $N_1 = N_2 = P/2$, is always an equilibrium but it is not always stable or always optimal. The one-city allocations are also equilibria.[18] There are also asymmetric equilibria with both cities populated. Suppose that N^* is the population level where utility peaks in a city. Then there are two such asymmetric equilibria where utilities are equal: $V(N_1) = V(P - N_1)$. In one of these $N_1 < N^*$ and in the other $N_1 > N^*$. The conclusions depend strictly on the assumed shape of the locus of the asymmetric equilibria in (N_1, N_2) space. In this space, the vertical and horizontal axes correspond to the one-city equilibria and the 45 degree line corresponds to the symmetric equilibria. The slope of the asymmetric equilibria locus is negative everywhere. Anas (1992) assumes that the asymmetric locus is convex to the origin and does not cut the axes. This corresponds to the situation where agglomeration economies build up quickly with city size but decline gently once autarkically optimal city size is reached. There are two regimes separated by a bifurcation point.[19] For $P < 2N^*$, there are three equilibria. These are the two one-city equilibria that are locally stable and the symmetric equilibrium that is locally unstable. For $P > 2N^*$, there are five equilibria. These are the two one-city equilibria, the symmetric equilibrium and the two asymmetric equilibria. In this regime, the one-city equilibria continue to be locally stable, the symmetric equilibrium becomes locally stable, but the asymmetric equilibria are unstable. Now suppose that we start from a very low P, letting it grow exogenously. Since, for low P, the symmetric equilibrium is unstable, we will start with one city. At what point will a second city emerge under atomistic defection of a small population increment from the first? The answer is that in all likelihood this will never happen. For this to happen at some P, the asymmetric equilibrium locus would have to become asymptotic to the two axes. Only in that case, the migration of a single agent from the established city to the virgin site of the second city could establish a higher utility causing a catastrophic cumulative migration process to ensue, ending on a symmetric equilibrium. Otherwise, population will continue to be stuck in the first city, regardless of how overpopulated the first city becomes. Clearly, under these circumstances, atomistic migration fails and a developer or government is needed to start the

[17] This presentation borrows heavily from Anas (1992).

[18] But this requires that $U(P) > U(0)$. If not, then a one-city equilibrium is unstable even under the slightest population perturbation.

[19] The bifurcation pattern is identical to the *pitchfork bifurcation* of Krugman.

second city. That is because a developer or government can orchestrate the simultaneous migration of a critical mass of agents from the first city to establish the second.

It is also shown in Anas (1992) that the socially optimal time when the second city should be set up is t^* such that $V(P(t^*)) = V(P(t^*)/2)$. For most reasonable utility functions, this point in time is much earlier than the point in time when atomistic defection might cause the second city to self-organize. From this we can conclude that with a small number of cities the market process causes new cities to be set up too late if at all. Exceptions to the case examined by Anas do exist. Pines (2000) examined a case in which the autarkic utility curve rises to the optimal city size, N^*, thereafter falling steeply. Although it is hard to produce a realistic example of such a case, his results are interesting only because he assumed that utility falls to zero with sufficiently high population. That causes the asymmetric locus to cut both axes. In this case, atomistic defection can start a new city, but only if that atomistic defection can happen at a precise point in time. If that happens, then the cumulative process evolves smoothly with the original city declining and the new one rising in population until the symmetric equilibrium is reached. Otherwise, if that exact point in time is missed, atomistic defection would still work but only by causing a catastrophic migration process.

In all of the above-mentioned papers, the authors assumed situations in which new cities are intended to be identical to existing ones. Anas and Xiong (2004) examine a setup in which there are two industries. A homogeneous and competitive final good (manufactures) industry uses differentiated, monopolistically competitive services as intermediate inputs. If these inputs are not locally available, the final goods producers will import them from other cities. Optimal city size exists because the differentiated inputs impart an external scale economy on production (see Section 3.2). Suppose that initially there is only one city producing its own services for the use of the final good industry. It is shown that either services or manufactures can move out and set up a second smaller city at an available site by purely atomistic defection with such a new city growing smoothly from infinitesimal size.[20] This means that developers are not needed to set up the new city. They also show that the new city can emerge before the existing one is overpopulated and emerge at the optimal time.

5. Key issues and a summary of historical developments in the literature

We now wish to provide a characterization of the issues that arise in the specification of models of city systems and to follow that with a brief and broad overview of the historical developments that have occurred in the literature.

5.1. Key issues

To our knowledge, all models of systems-of-cities in existence assume that a system of cities is characterized by the costless migration of economic agents (consumers and

[20] Hadar and Pines (2003) have extended the analysis.

firms) among the cities. This assumption, though highly objectionable, forms the basis of an equilibrium principle, namely that cities are organized in such a way that, in equilibrium, agents of the same type are indifferent about the city in which they locate.[21] For example, cities with better local public goods or higher natural amenities should offer workers lower wages in the labor market and higher land rents in the land market, a form of compensating differentials. At equilibrium, these compensating differentials result in equal utility in all cities for agents of the same type.

Beyond this principle of costless migration, city-system models face a large number of basic questions that must be resolved in one way or another in specific models. Forming a clear list of such questions is essential for understanding and classifying the various contributions. Our list – which we believe is complete – consists of fourteen categories of questions.

1. *Are cities in the system identical in size and in industrial composition or are they different*? Cities may be identical in the composition of industries or they may differ. If a city contains only one industry, it is referred to as a *specialized city*. If it contains all of the modeled industries (or at least more than one) it is called a *diversified city*. All models of city systems have either specialized or diversified cities. See our discussion in Section 6.2.
2. *Is the number of commodities produced in the city system predetermined or endogenously determined*? Henderson (1974) and Wilson (1987) have emphasized alternative models with a predetermined number of commodities, whereas the New Economic Geography of Krugman (1991) emphasized models with an endogenously determined number of goods.
3. *Are city industries perfectly or imperfectly competitive or a mix of the two*? Again, Henderson (1974) and Wilson (1987) model industries as competitive, whereas Henderson and Abdel-Rahman (1991) and the New Economic Geography model them as imperfectly competitive. Abdel-Rahman (1990b, 1994) and Anas and Xiong (2003) examined models with both perfect and imperfectly competitive industries.
4. *Are cities in the system composed of homogeneous or heterogeneous worker/laborers*? *Is this heterogeneity exogenously imposed or endogenously determined*? Both types of models exist in the literature. See our discussion on models with labor heterogeneity in Section 7.
5. *Are cities set up by community planners (local governments) or developers, or do they become self-organized by atomistic defection of economic agents from existing cities*? Henderson and Becker (2000) presented a comparison of these city formation mechanisms. See also our discussion in Sections 4.1–4.3.
6. *How are cities linked with each other*? There is a variety of potential linkages that must be addressed. The most basic, as explained above, is costless migration

[21] There are, however, models of costly migration among countries by Hercowitz and Pines (1991) that could also be applied to intercity migration.

among cities. Are commodities produced in different cities traded among cities? Do agents located in one city own endowments (e.g., land shares) in other cities? Are public goods provided in one city excludable from residents of other cities? Namely, are they local or semi-local public goods? In city-system models with Dixit–Stiglitz varieties that are traded [e.g., Hadar and Pines (2004) or Anas and Xiong (2003, 2004)], the number of varieties produced in a city is equivalent to a semi-public good.

7. *What is the nature of transportation costs for goods traded between cities*? In the literature, the assumption of zero transport costs (in Henderson and Wilson type models) and that of positive iceberg transport costs (in Krugman type models) have played important roles.
8. *What is the nature of externalities operating within and between cities? Are these black-box technological externalities or pecuniary externalities arising from market transactions*? Our discussion in Section 3 illustrated the public good and variety models in which externalities are pecuniary and the labor productivity model in which the externality is technological.
9. *Is the number of cities in the city system predetermined and fixed or is the number and type of cities endogenously determined*? Both cases have been examined rather extensively. See, for example, Hadar and Pines (2004) or Tabuchi (1998) for the former and Henderson and Abdel-Rahman (1991), Anas (2004), Anas and Xiong (2003) for the latter.
10. *Does the city system have an explicit spatial dimension (e.g., cities located on a circle or a linear space) or is the geography abstract or highly simplified*? Treatment of cities on an explicit spatial geography has appeared only in the New Economic Geography [Krugman (1991)]. All other models of systems of cities are formulated on an abstract geography with zero transportation cost or positive transport costs but symmetrically located cities.
11. *Is the internal structure of cities (land market) treated explicitly or is it suppressed*? A simplified model of internal city structure, such as that of Section 2 is standard in the literature, except for the basic models in the New Economic Geography that ignore internal structure when dealing with city systems.
12. *Is the city system linked with a rural (non-urban) sector in the economy*? Kanemoto (1980) and Anas and Xiong (1999) present a system of cities with a rural sector and costless rural to urban migration, whereas in the standard approach of the New Economic Geography [e.g., Fujita, Krugman and Venables (1999)], rural workers are distributed over space and are treated as immobile.
13. *Does the city-system model generate unique or multiple equilibria under the same parameter values? Are there symmetric equilibria (where all cities are identical) as well as asymmetric equilibria (where cities differ in sizes or types)*? The presence of externalities in city systems has resulted in models where multiple equilibria are known to arise and are prevalent. But virtually all authors focus on the symmetric equilibria.

14. *What are the sources of market failure in the system of cities and what are the differences between market equilibrium and first-best and second-best allocations*? Our discussion of Sections 3 and 4 highlighted several types of market failure at the city level and Section 8 extends this efficiency discussion to the entire system of cities with a potentially unlimited number of sites where cities can emerge.

5.2. *Historical development of the field*

Historically, the development of models of systems of cities has proceeded as follows. In the earliest models by Henderson (1974), industries are perfectly competitive producing homogeneous goods and the number of such goods produced in the city system is predetermined. Increasing returns at the city level arise from Marshallian localization (industry-specific) externalities and not from pecuniary market transactions. There may be goods that are locally consumed and other goods that are traded between cities. However, trading costs are always assumed zero. Hence, as we shall see later, different traded goods are not produced in the same city because doing so would crowd the land market, raising rents and commuting costs without any offsetting benefit. The number of cities is endogenous since city developers are free to enter and form new cities according to the Henry George Theorem. Hence, city sizes are efficient. Geography is suppressed and there is no rural sector. Generally, there is a unique equilibrium (or if multiple equilibria exist, these are not specified or not examined).

A second line of modeling is based on the literature of local public goods as it applies to city formation [Stiglitz (1977)]. Wilson (1987) proposed a model in which there are two tradable private goods produced under constant returns with land and labor and without any localization economies. A third, non-traded good is a local public good. The cost of trading the two private goods among cities is again zero and again there is no rural sector. In such a model, the autarkic optimal city size is determined as an inverse U-shaped function of the city's labor force. This shape reflects the fact that the per-capita cost of the public good decreases as population is added to the city (see Section 3.1) while the marginal product and, hence, the wage received by each person also decreases. Thus, an optimal city size exists in such a model even without any commuting cost. The key question in such a model addressed by Wilson is whether the two private goods industries will locate in the same city or in separate specialized cities and how to locate them optimally. In contrast to the Henderson-type model, as we shall see, the Wilson type model has been shown to result in specialization even in the absence of land-market crowding effects (e.g., commuting costs).

In the late eighties, Hobson (1987), Abdel-Rahman (1988) and Rivera-Batiz (1988) and Abdel-Rahman and Fujita (1990) adapted the Dixit and Stiglitz (1977) model of product differentiation and monopolistic competition to an urban setting, starting models that differed considerably from those of Henderson and Wilson. In these models, firms are treated as imperfectly competitive according to Section 3.2 and subject to internal economies of scale, each firm producing a unique differentiated product. These

are either final products sold to consumers [as in Dixit and Stiglitz (1977)] or intermediate products used as inputs in the production of homogeneous final products [Ethier (1982)]. By the free entry of firms, the number of goods is made endogenous and – as explained in Section 3.2 – the variety of these goods is a pecuniary external scale economy at the city level. The differentiated goods are assumed to be city-specific and are not traded between cities. As in the Henderson models, geography is again unspecified and there is no rural sector. Welfare analysis as well as market equilibrium are of interest. Because different industries can locate in the same city to share differentiated intermediate inputs, a variation of these product differentiation models have also been used by Abdel-Rahman (1990b) to explain urban diversity as we will see in more detail in Section 6.3.

In all of the above models, the system consists of specialized or diversified cities. However, in developed and developing countries specialized as well as diversified cities coexist. In an attempt to generate more realistic city system, Abdel-Rahman and Fujita (1990) and Abdel-Rahman (1996) adopted the idea of economies of scope in production to an urban setting which leads to the coexistence of specialized and diversified cities in equilibrium. But specialization versus diversification can also be explained without assuming economies of scope. In Anas (2004) and Anas and Xiong (2003), as we shall see, urban diversity versus specialization is explained as a result of the interplay of trading costs, on the one hand, and urban location costs on the other. In Duranton and Puga (2001), specialized and diversified cities coexist as a result of a dynamic model in which the city choice decisions of the firm are tied to the product cycle in a dynamic model.

Another development is based on the ideas in Krugman (1991) and is now called the New Economic Geography (NEG). In this type of model, the industry structure is again of the Dixit–Stiglitz–Ethier type. But there are several important additional features all introduced by Krugman. First, unlike in the other two model types, trading cost is positive and of the iceberg type. Second, self-organization under atomistic defection (not developers or community planners) is the city-formation mechanism. There is either a predetermined number of cities or, at least, not an unlimited number. In the basic versions of these models, cities do not have land markets and are treated as points. Multiple equilibria, not welfare analysis is the focus of the analysis. Most importantly, a rural sector with immobile farmers is usually present and plays a key role in the formation and location of cities. A specific but simple geography (e.g., line or circle) is usually treated. City systems are derived by complex simulation exercises [see Fujita and Mori (1997) and Fujita, Krugman and Mori (1999)].

Finally, most models of city systems assumed that cities are populated with identical households/workers. Helsley and Strange (1990), Kim (1991), Henderson and Becker (2000), and Abdel-Rahman and Wang (1995, 1997) introduced models with multiple but exogenously defined worker types. Abdel-Rahman (2002) developed a model in which the skill distribution of workers within the system of cities is determined endogenously through self-selection.

6. Homogeneous labor

We now turn to the most basic models of city systems under the assumption of homogeneous labor, and discuss the most important issues that have been studied using such models. These are the presence of constant returns in the extensive margin as new cities are formed when cities are not interacting, the industrial specialization versus diversification of cities, the presence of economies of scope among industries and how this affects the formation of diversified cities, the importance of trade among cities producing manufactures and/or differentiated products and how such trade causes the emergence of increasing returns with respect to the aggregate population of the city system.

6.1. The simplest case: identical isolated cities

The simplest city-system model consists of n identical cities that have no interaction except that the consumers are free to migrate costlessly among cities establishing an equal-utility equilibrium. Each city is set up at the autarkic efficient size of N^* consumers so that $nN^* = P$, the national or regional population. Although the number of cities must be an integer, it is commonly assumed that P is large enough relative to N^* so that n is treated as a continuous variable to give a good approximation. City economies may be according to one of the three models discussed in Section 3 and may be setup by a developer, a local government or self-organize under atomistic defection. Assuming large P, $n = P/N^*$ is a stable equilibrium since with this number of cities, consumers in each city will achieve the highest level of utility and thus atomistic defection to other cities could not result in higher utility. Similarly, if cities were set up by developers and $n > P/N^*$, each city is smaller than optimal size and city developers make losses. Some of them must exit the market, $N \to N^*$ and $n \to P/N^*$. Conversely, if $n < P/N^*$, each city is larger than optimal size and developers make positive profits. New developers should emerge with n and N tending to equilibrium again.

What are possible reasons why the cities do not have any interaction? According to the local public good model of Section 3.1, each city produces the same local public good and the same consumption good. Hence, nothing is gained by trading these goods. According to the model of Section 3.3, all cities would produce the same consumption good, again there being no reason to trade. City output would be consumed locally and any excess could be exported to the rest of the world from the CBD. In the case of the product variety model of Section 3.2, however, things are a bit different. In this case, because consumers have an extreme taste for variety (or producers an extreme bias for input variety) they would have a strong desire to not only consume their local varieties but to also import all other varieties produced in the other cities. For such cities not to trade differentiated goods, intercity trading costs have to be infinitely high, an extreme case.

The most important property of a city system with isolated non-interacting cities is a result established in the early papers of Henderson (1974), Upton (1981) and Henderson

and Ioannides (1981). According to this result, even though each city is subject to internal economies of scale with respect to city population, due to centripetal forces such as those of Section 3, the city system on aggregate exhibits constant returns to scale at the extensive margin. As the aggregate population, P, grows, this constant returns is maintained by spawning new identical cities of size N^*, with all aggregates such as gross product growing at the rate of population growth. This property illustrates the importance of an economy being able to spawn new cities. When this is not possible, additional population growth must be accommodated in existing cities and these cities become larger than their autarkic optimal size and causing welfare losses.

6.2. *Specialization versus diversification*

A central issue is whether cities in the system will be specialized or diversified in production. Specialization occurs when production in the city consists of only one industry. Diversification is normally defined as two or more industries co-locating in the same city. The simplest way to study the problem is to assume that there are just two industries and then explore the conditions under which these two industries will locate in the same city.

6.2.1. *Specialization*

Suppose that the two industries have no direct connection with one another. Each produces a homogeneous good that can be traded among the consumers of the city system. Then, there are two opposing effects that must be considered. Although the literature has not been explicit about these effects, we will try to describe them carefully here.[22] The first may be called the *trading-economy effect*. This means that if each city contains both industries producing enough output to meet the local demand for both goods, then the city need not import either good. Such a diversified city saves importation cost for its residents and raises their utility. The second effect is the *crowding-out effect*. This means that if a city contains two industries it will be larger in labor force and in population. Average commuting costs will be larger in such a city lowering utility. Thus, unrelated constant returns to scale industries crowd each other out in land markets favoring the specialization of cities.

This crowding-out effect does not exist in Krugman's New Economic Geography because in those models cities are typically modeled as not having any land markets. On the other hand, in the Henderson type models with traded goods, the cost of trading goods among cities is always assumed to be zero. Hence, the trading-economy effect is non-existent. Then, the crowding-out effect remains unopposed and cities are always specialized. We can easily generalize this intuition to $k = 1, \ldots, K$ industries as long as there is no limit on the number of cities that can be formed. Suppose that all goods

[22] We follow Anas and Xiong (2003) who discuss these two effects explicitly.

can be traded within and among cities at zero cost. Then, there will be n_k specialized cities of type-k containing industry k and N_k residents. Given a total population P, $\sum_{i=1}^{K} n_k N_k = P$. Each of these cities will be set up at the efficient size implied by the economies of scale of their respective industry. If the efficient size of a city specializing in good $k+1$ is larger than the efficient size of a city specializing in good k, rents in a type-k city will be higher. To compensate, so that utility is invariant among cities, wages in the type $k+1$ city should also be higher than those in the type k. Meanwhile, the number of cities of each type will adjust so that the total quantity of each good k demanded in the city system is supplied.

The above specialization result may be contrasted with that which occurs in the model by Wilson (1987) described earlier. Recall that in that model there were homogeneous consumer-workers, two industries without localization economies, zero trading cost and a city-specific public good. It was assumed that developers set up cities. Hence, a city would be specialized if the developer found it beneficial to allow production of only one good in his city. This is an argument for specialization and trade based on the theory of clubs rather than on the theory of localization externalities. Papageorgiou and Pines (1999) provide an analysis of this model, arguing that whether cities will be specialized or diversified at the optimum of the city system depends on the asymmetry between the technologies for producing the two private goods and on the complementarities between private and public good production. Typically, in the optimum of a system of specialized cities, one will be larger than its autarkic size and the other smaller, and trade of the specialized goods will take place. At such an optimum, the city with the larger public good provision is overpopulated, pays a lower wage and has higher rents (the theory of compensating differentials) and the two types of cities differ in consumption mix as well, because the city-specific wage (determined by the city-specific marginal product of labor) means that the relative prices of the two goods are different depending on the type of city one lives in.

6.2.2. Economies of scope

One way to depart from the completely specialized city system is to assume linkages between the various industries. Perhaps the most obvious way of doing so is to assert economies of scope when the two goods are produced together in the same city. See Panzar and Willig (1981) for the non-spatial theory of economies of scope and for its relation to agglomeration economies, see Goldstein and Gronberg (1984).[23]

Abdel-Rahman (1990a) developed the first spatial model of asymmetric economies of scope. There are two homogeneous goods where the production functions are $x_1 = f(H_1, H_2)h_1$ with $f_1, f_2 > 0$ and $x_2 = g(H_2)h_2$ with $g_1 > 0$, where x_1, x_2 are

[23] See Helsley and Strange (1993) for a non-spatial micro foundation model of urban agglomeration in which matching in the used-assets capital market enhances the salvage value of the assets from failed projects. For a review of models based on search, learning and matching, see Duranton and Puga (2004) in this volume.

the outputs of firms in industry 1 and 2, respectively, h_1, h_2 are labor inputs in the firms and H_1, H_2 are aggregate labor inputs at the level of the industries. Thus, industry 1 has *urbanization economies* [Jacobs (1969)] because it reflects external economies of scale that operate across industries in the same city. Industry 2 has *localization economies* [Marshall (1890)]. The author examined a system consisting of a diversified city and a city specializing in industry 2. The main result is that the diversified city can be larger in equilibrium than the specialized city if at least one of the industries exhibits decreasing returns to scale at some point.

In Abdel-Rahman and Fujita (1993) economies of scope are modeled on the cost side. It is assumed that developers can set up cities, minimizing location plus production costs. There are again two final goods ($i = 1, 2$) with cost functions $C_i = F_i + c_i X_i$. F_i is a fixed cost that must be incurred for production to start and c_i is a constant marginal cost. X_i is industry output. Since trading costs are zero, pure specialization would occur with each city producing only one of the goods. To offset this, it is assumed that when the two goods are produced in the same city, then the fixed cost of production is $F_d < F_1 + F_2$. The cost of producing both goods in such a city is then $C_d = F_d + c_1 X_1 + c_2 X_2$. Clearly, if such fixed cost savings are not sufficiently high, cities will be specialized because the crowding-out effect is not completely offset by the cost savings of joint production. But if these savings are sufficiently high all of the cities in the system will be diversified producing both goods. There are also mixed equilibria with some cities diversified while others are specialized in one good. The mixed equilibria, in which the specialized city produces good 1, would result if $F_d > F_1$ but $F_d < F_2$. The authors showed that the diversified cities would be larger than the specialized ones.

A related paper is by Abdel-Rahman (1994). In this model there are two constant returns final goods industries producing goods traded at zero cost. Each produces using homogeneous labor plus a service specialized to that industry. Services are assumed not to be tradable among cities. The production of the two services is subject to economies of scope. In this case the saving from joint production can occur either in the fixed or the variable cost of service production. Again, when the saving from joint production of services is not high, pure specialization occurs with each city producing one final and its related intermediate good. In this case, the existence of cities is dependent on the presence of a scale economy (fixed cost) in the production of the service. Conversely, if the savings from the joint production of the two services are sufficiently high, all cities are purely diversified in equilibrium, each producing both final goods and both services. If the economy of scope lowers variable costs rather than fixed costs then a mixed equilibrium exists in which the specialized and diversified cities co-exist and, in such an equilibrium, the specialized cities can be larger than the diversified ones.

The above models treat the economy of scope as a black box because they do not show explicitly the source from which the productivity gains or the cost savings arise. A model that does so explicitly is the sharable inputs model of Abdel-Rahman (1990b). The city produces a traded good and a local public good. Both production processes are constant returns and use a homogeneous labor input plus the entire variety of dif-

ferentiated, non-traded services also produced in the city. These shared services are monopolistically competitive as in Ethier (1982).

6.2.3. Diversification without economies of scope

While the above models achieve diversification by imposing economies of scope in the production process, it is more challenging to do this without imposing such a relationship. In fact, when economy of scope is imposed, the result of diversification in cities follows almost directly from the premise of the model. The obvious alternative is to assume that the two final goods industries are unrelated in production and that the two goods are costly to trade. If these goods are sufficiently costly to trade, then there will be a significant trading-economy effect to offset the crowding-out effect. Abdel-Rahman (1996) presents such a model. In this model, there are two final goods that can be traded among cities. One of these is traded for consumption while the other is used to produce commuting which is assumed to require monetary as opposed to time expenditure. Each manufacturing industry uses labor and differentiated services specialized to that industry. The specialized services are monopolistically competitive, as in the previous models. Because the industries do not use the same services, there are no economies of scope. Like in all Abdel-Rahman models, the services cannot be traded among cities. Two equilibria are analyzed. The first occurs when the cost of trading manufactures among cities is sufficiently low. Then all cities are specialized in one of the goods and its associated services at equilibrium. Clearly, this is because the crowding-out effect dominates the trading-economy effect. In the second type of equilibrium, all cities are identical and purely diversified. Each city produces both manufactures and the associated services for each and there is no trade. This equilibrium occurs when trading costs are so high that the trading-economy effect dominates the crowding-out effect.

6.2.4. Intercity trade of services

In all of the preceding models with services, the fact that these services cannot be traded provides the source of the industry-level scale economy. But is the assumption that services cannot be traded realistic? On the one hand, Abdel-Rahman and Fujita (1990) have described these producer services as "repair and maintenance services, engineering and legal support, transportation and communication services, and financial and advertising services." While some of these services are difficult to export to other cities, it is increasingly true that communication, financial (banking, insurance, investment) and advertising services are traded among cities. The recent adoption of the Internet as a communication device has increased trading of such services. Hence, the assumption that services can be traded among cities at some cost while final goods are also tradable at some other cost appears realistic. This has been examined in Anas and Xiong (2003). The setup of their model is similar to Abdel-Rahman (1996) but differs in several minor and one major respect. The minor differences are that commuting within cities costs time not money and that both final goods are consumed (rather than one being used

to pay for commuting). The major difference is that both services and final goods are tradable among the cities but at different unit cost rates. The final goods (manufactures) are produced using labor and the largest possible variety of services special to that industry. Hence, each manufacturer will import all services of his industry from all other cities that produce them as well as use his locally available services. By assuming that the two industries are symmetric in technology and that the demand for their products is also symmetric, the analysis is greatly simplified. Two equilibria are analyzed using these assumptions of symmetry. In one all cities are diversified (containing both manufacturing industries and their associated services). In the other equilibrium, half of the cities are specialized in each manufacture and its associated services. The number of cities and the size of a city are the same in the two equilibria. Hence, the size of an industry located in a diversified city is half as big as it would be if the same industry were located in a specialized city. This means that a firm located in a diversified city will have exactly half the number of local services available to it than a firm located in a specialized city. This is exactly how the crowding-out effect operates in the model. On the other hand, the trading-economy effect means that there will be no trading of manufactures if all cities are diversified. It is shown that either equilibrium can yield higher utility, depending on various parameters of the model. For example, increasing the share of services in production favors specialized cities since, in such cities, more services are locally available and fewer have to be imported. For the same reason, if services are more expensive to trade, this favors specialization. Conversely, if manufactures are expensive to trade, this increases the savings from the trading-economy effect and favors the equilibrium of diversified cities. If commuting cost increases, cities become smaller and there are fewer service varieties available locally, which increases the utility of a specialized city relative to a diversified city because in the specialized city there are more service varieties locally.

6.2.5. *Product cycles*

Most recently, specialization and diversification has been analyzed in a dynamic model that resulted in system of cities in which diversified and specialized cities co-exist. Duranton and Puga (2001) developed the first model of product cycles in the systems of cities literature.[24] In this model, a metropolitan area plays the role of a nursery for new products. They employed a model of product development where firms experiment with prototypes in a diversified city until they find the ideal production process. After the firm identifies the ideal production process, it moves to a specialized city to start mass production. The main result of the paper is to identify the conditions that result in a unique steady state in which specialized and diversified cities coexist. However, the size of the diversified city is the same as that of the specialized one, which is not consistent with

[24] Henderson, Kuncoro and Turner (1995) had provided empirical support for this product cycle model in a system of cities.

empirical observation nor with other theoretical models that result in the co-existence of diversified and specialized cities.[25]

6.3. Increasing returns with traded varieties

In some of the models discussed above, the presence of costly trade among cities was the basis for the incentive to save trading costs and, hence, played a major role in whether two industries would locate in the same city or not. A second important consequence of trade among cities is the possible emergence of increasing returns in aggregate national population. Recall that in the models of city systems with isolated cities (see Section 6.1), the city system exhibited constant returns in the extensive margin as cities were created in response to population growth.

To see how trade among cities can induce increasing returns in aggregate population, we may construct a simple extension of the model of product variety that we examined in Section 3.2. In that model, we looked at only one city producing m manufactures in equilibrium, in symmetric industries. Consumers viewed these products as imperfect substitutes and had a strong enough taste for variety to want to consume all of these products at any price. What would happen if we had n such identical cities each producing m varieties of manufactures distinct from those produced in any other city? Henderson and Abdel-Rahman (1991) provided an answer, but they assume zero trading costs. Following Anas (2004), we assume an abstract geometry in which the n cities are located symmetrically with respect to one another. We assume, as he does, iceberg transport costs so that $1/\tau$ is the multiple of the demanded quantity of any manufacture that must be shipped from one city to any other. When $\tau = 1$, the cost of transportation is zero and when $\tau = 0$, the cost of transportation is infinitely high. Given that P is the national population in the entire city system and given that each city is set up at its autarkic efficient size, N, the number of symmetric cities in this economy will be $n = P/N$.

We may easily examine two aggregate quantities in this economy. One is the level of utility and the other is the *gross domestic product* (GDP) of the economy. Starting with the latter and using previously derived quantities (see Section 3.2),

$$\text{GDP} = nmz = \frac{P}{N}\frac{N(1-kN^{1/2})}{f\sigma}\frac{f(\sigma-1)}{c} = \frac{\sigma-1}{\sigma c}\left(1-kN^{1/2}\right)P.$$

GDP grows linearly with national population (or per-capita GDP is constant). Hence, there are no increasing returns. This is not at all surprising since the benefits of variety occur on the demand side not on the supply side. Turning to utility, we write the direct utility first by recognizing the symmetry we have imposed. It is

$$U = \left[mx_i^{(\sigma-1)/\sigma} + (n-1)mx_{-i}^{(\sigma-1)/\sigma}\right]^{\frac{\sigma}{\sigma-1}}$$

[25] For models in which diversified and specialized cities co-exist, see Abdel-Rahman and Fujita (1990) and Abdel-Rahman (1994) discussed earlier.

where x_i is the quantity of each variety the consumer buys that is produced in his own city and x_{-i} is the quantity of each variety the consumer buys that is produced in each of the other cities. At this point, it is convenient to switch to the indirect utility which takes the form,

$$V = \left[mp^{1-\sigma} + (n-1)m\left(\frac{p}{\tau}\right)^{1-\sigma} \right]^{1/(\sigma-1)} I(N).$$

Using the previously derived quantities to make substitutions, we get

$$V = p^{-1}\left[N + (P-N)\tau^{\sigma-1}\right]^{1/(\sigma-1)}\left(1 - kN^{1/2}\right)^{\sigma/(\sigma-1)}.$$

Taking the first derivative with respect to P,

$$\frac{\partial V}{\partial P} = \frac{\tau^{\sigma-1}}{p(\sigma-1)}\left(1 - kN^{1/2}\right)^{\sigma/(\sigma-1)}\left[N + (P-N)\tau^{\sigma-1}\right]^{(2-\sigma)/(\sigma-1)}.$$

Thus we see that utility increases with aggregate population except in the case of $\tau = 0$, when cities do not trade because trading cost is prohibitively high. Furthermore, when the taste for variety is sufficiently strong ($\sigma < 2$) then utility increases with P at an increasing rate. In a system of cities model in which producers used all the input varieties while consumers purchased the homogeneous output of these producers, per capita GDP would increase with population, but utility would not.

How would increasing returns with respect to national population become exhausted? There are no models in the literature that directly analyze this issue. One obvious answer is that the land limitation of countries [present in Anas and Xiong (1999)] and the limited number of sites with certain amenities such finite seashore [present in Helpman and Pines (1980)] would at some point restrict returns from additional population growth as land on which new cities can be set up diminishes.

7. Heterogeneous labor

A common feature of all the models discussed in Section 6 is that they ignore the heterogeneity of consumers and workers. Models with such heterogeneity fall into two groups:

(1) those that exogenously introduce different types of workers, and
(2) those that generate the types of workers endogenously.

Helsley and Strange (1990) and Kim (1991), permit horizontal differentiation among workers in a model in which productivity gains are driven by better matching between workers and firms. However, in their model, all cities are identical and all workers achieve the same equilibrium expected utility. Henderson and Becker (2000), Abdel-Rahman (1998) and Abdel-Rahman and Wang (1995, 1997) have different types of exogenously specified workers achieving different utility levels in equilibrium, while Abdel-Rahman (2002) endogenously generates household types.

In Abdel-Rahman (1998), there are skilled and unskilled workers and with leisure in the utility function. Cities in this model are formed due to investment in public infrastructure. The model generates a system of two types of cities in which one type produces a food product with unskilled workers, and the other type produces a high-tech manufacturing product with the use of skilled workers. Workers sort themselves into these two types of cities. The model identifies the determinants of income inequality, which includes productivity and infrastructure effects. Then, the paper analyzes the impact of income inequality on social welfare.

In Henderson and Becker (2000), there are households who are either entrepreneurs or workers, but in this case both types locate in the same city and the model results in a system of identical cities. Cities are formed due to an externality resulting from intra-industry specialization as in Becker and Murphy (1992). The paper examines city formation by large land developers as well as self-organization and a combination of both. The main focus is efficiency of equilibrium and the requirement to achieve efficiency under different city formation mechanisms.

Abdel-Rahman and Wang (1995, 1997) examine a situation in which the national population of workers is exogenously divided into unskilled and skilled laborers. The unskilled workers are homogeneous while the skilled workers are distributed uniformly on the unit circle. The economy has two goods: food requiring unskilled labor input and a high-tech good requiring input from skilled workers. The food industry is subject to a localization economy, a decrease in the average cost of providing a form of city-specific infrastructure that facilitates food production. The two goods can be traded among the cities at zero cost. Clearly, the setup just described favors specialization since there are no direct connections between the industries. Thus, there is neither a trading-economy effect, nor economies of scope, nor input sharing to oppose the crowding-out effect that would occur if both goods were to be produced in the same city. Because of this, the equilibrium would have a core–periphery structure. There would be peripheral cities accommodating unskilled workers only, specializing in the production of food and core cities accommodating the skilled workers specializing in the production of the high-tech good.

An important question is whether at equilibrium all core cities would be identical accommodating workers of all skills or whether they would be different, accommodating workers of a subset of skill ranges. The authors assume that firms cannot setup new cities taking with them workers whose skills are best matched with the firms, i.e., coalitions of workers and firms are not allowed. This condition is insured by assuming that firms do not know a priori the skills of their workers nor workers the skill requirements of firms and that firms and workers cannot sign pre-migration contracts after matching up in one city and prior to relocating to another. Thus, a worker migrating to a new city expects to incur search costs after arriving there. Although this assumption is perhaps somewhat artificial, it ensures that high-tech cities will all be symmetric in equilibrium. As in Helsley and Strange (1990) such cities will accommodate more than one firm since the density of firms on the unit circle confers a productivity economy. The authors examine a special equilibrium outcome in which the core consists of a sin-

gle large metropolis while the periphery consists of many identical and smaller cities specializing in food production and accommodating only unskilled workers. In the first paper, a symmetric Nash wage bargaining rule, as in Diamond (1982), determines a uniform wage for all skilled workers regardless of how well they are matched to their firms. However, the income of the skilled core workers is higher than that of the peripheral unskilled workers. In the second model, the Nash wage bargain is asymmetric so that firms can differentially reward workers with whom they are better matched. It is assumed that the better-matched workers have higher bargaining power and extract a higher wage. As a result, the second model produces income inequality within the core as well as between the core and the periphery. Unearned income does not play a role in the income distribution because the authors assume that all land is publicly owned and that local planners or city developers use the aggregate land rent to finance the source of the city-level scale economy. In the peripheral cities, the aggregate rents are used to pay for the local infrastructure investment, while in the core the aggregate rent is used to subsidize the fixed costs of the firms. The main claim of these two papers is that changes that improve efficiency can worsen income inequality. For example, the authors show that lower search costs and better matching in the high-tech industry cause the income distribution to become more unequal.

An extension of the above work is by Abdel-Rahman (2002). In this paper, all workers are homogeneous in skills a priori but are vertically differentiated in innate ability defined as the uniform distribution on a unit interval. However, educational investment is endogenous in the model in the sense that each worker may decide whether to acquire specialized training that causes his productivity to rise to a higher level and enables him to obtain a higher wage. If a worker decides not to acquire specialized training, he gets only a basic level of education that is provided as a public good. Only one good is produced, but there are two technologies for producing it. The good may be produced by a basic technology that utilizes workers who get only basic education or it can be produced by a specialized technology that utilizes workers who choose to acquire skills. At equilibrium, workers with higher a priori abilities can acquire skills more cheaply and all workers above a certain reservation ability level choose to acquire skills. The skills acquired are proportional to ability. Because the good is homogeneous, there is again no trading between cities and each city is self-sufficient at equilibrium. There are, however, several types of equilibria. If all workers acquire specialized training, all cities are identical high-technology cities. If no one acquires specialized training, all cities are identical low-technology cities. If, as is most realistic, if only some of the workers acquire specialized training then there are two types of specialized cities co-existing. One type is a high-technology city populated by skilled workers only and the other, a low-technology city populated by unskilled workers. Under realistic parameter values, there would be one or only a few high-technology cities and many low-technology cities. The lowest ability consumer living in a high-technology city would be indifferent between this situation and that of living in a low-technology city.

8. Efficiency and the role of central planning in city systems[26]

In Section 4 we examined optimal resource allocation within individual cities which do not interact. In Sections 6 and 7 we reviewed a variety of equilibrium models of systems of cities, making only slight reference to several issues of optimal resource allocation in such a system. We now pass to a more comprehensive treatment of issues of efficient resource allocation in a system of cities, where such cities may or may not interact with each other.

Henderson (1977), Helpman and Pines (1980) and Kanemoto (1980) were the first to pose problems of optimal resource allocation in a system of cities while some general issues have been laid out by Hochman (1981, 1997).[27] In Helpman and Pines, there is a large number of sites where cities can be set up. Each site is endowed with a different level of a natural amenity. Their model is a variant of the public goods model of Section 3.1. The quality of the public good created in a city is higher, the higher is the level of the natural amenity and the higher is the public expenditure. In addition, consumers derive utility from a composite good and from private lot size. In contrast to the model in Section 2, used in virtually all system of cities models, their lot size increases with distance from the CBD [also true in Henderson (1977) and in Kanemoto (1980)]. The main result is that it is first-best efficient to allocate more population to the sites with the highest amenity levels. However, the presence of commuting cost limits city sizes and, hence, causes sites with lower amenity levels to also be developed. If the importance of the public good in the utility function is high relative to that of residential lot size, then the optimal allocation implies large cities at high population densities that are set up at sites of high amenity. If, on the other hand, land is more important in tastes than is the public good, then there is a larger number of smaller cities with lower densities developed at the lower amenity sites. To decentralize the first-best optimum, each city can be set up by an independent developer, acting as in Section 3.1, setting up the city at its autarkic-efficient size. The developer collects the local aggregate land rent and invests it in the public good, while letting the land market allocate lot sizes. The role that falls on the central planner is to simply specify which sites should be developed.

First-best central planning can also be examined in the context of the labor productivity model of Section 3.3. Suppose that, as in, e.g., Black and Henderson (1999), there are two industries with aggregate production functions $A_i(H_i)H_i$, where $i = 1, 2$ is the

[26] We will focus on systems of cities models with the number of cities endogenous and many. Anas (1992), Tabuchi (1998), Pines (2000), Papageorgiou and Pines (1999, 2000), Hadar and Pines (2003, 2004) have addressed various issues of optimal resource allocation when there are just two cities. Henderson (1977) treated a large number of cities, but not large enough to assume away lumpiness problems. In this case, the efficient allocation of population between cities requires the equalization of the gap between social marginal product and social marginal cost, whereas an equilibrium allocation would equalize private marginal returns to labor.

[27] Kanemoto (1980) dealt with formulations corresponding to the labor productivity case as well as the public goods case.

industry, and H_i is the labor supplied to industry i. The two goods can be traded at zero cost leading to complete specialization of cities by industry. The utility function is $U = x_i^{\alpha} y_i^{\beta}$, $\alpha + \beta = 1$, where $i = 1, 2$ denotes the city type, with cities of type 1 producing good x and cities of type 2 producing good y. The central planner's first-best resource allocation problem can be stated as:

$$\text{Max}_{x_1, x_2, y_1, y_2, n_1, n_2, N_1, N_2, U} U \quad \text{subject to:}$$

$$x_1^{\alpha} y_1^{\beta} - U = 0, \qquad x_2^{\alpha} y_2^{\beta} - U = 0,$$

$$n_1 N_1 x_1 + n_2 N_2 x_2 - n_1 A_1(H_1) H_1 = 0,$$

$$n_1 N_1 y_1 + n_2 N_2 y_2 - n_2 A_2(H_2) H_2 = 0,$$

$$n_1 N_1 + n_2 N_2 - P = 0.$$

This says that the central planner maximizes the equal utility level by deciding per capita consumptions in each city, the city sizes of each type and the number of cities of each type given aggregate population and a monocentric organization of each city as in Section 2. Hence, from Section 2, $H_i = N_i(1 - kN_i^{1/2})$ is the labor supply in a type i city. The third and fourth constraints conserve the system-wide output of each good, while the last constraint conserves population. The solution is one in which the result of Section 3.3 holds exactly: in each city, the local aggregate rents should be used to pay the subsidies that cover the gap between social marginal and average products. So, in the decentralized optimum, a different developer may set up each city, while the central planner indicates the type of each city and determines the number of each type (n_1, n_2) by issuing a license to each city developer.

In the above examples we see that the Henry George Theorem holds in the decentralized first-best optimum. The reason is that the agglomeration economy (public good in Helpman and Pines (1980) or labor productivity in the above formulation) exists only at the city level. The theorem fails to hold, when – with the number of cities endogenous – adding population to one city confers positive or negative external effects on other cities. A demonstration of this can be constructed using the model of product variety discussed in Section 3.2. Consider the extension of that model to a system of cities as formulated by Anas (2004). He assumes that all cities are symmetrically located with respect to one another and that all varieties are traded nationally with positive iceberg trading cost. In such a setting, he shows that the first-best optimal resource allocation problem of the central planner is:

$$\text{Max}_{x_i, x_{-i}, m, z, N} U(x_i, x_{-i}, m, z) = \left[m x_i^{(\sigma-1)/\sigma} + \left(\frac{P}{N} - 1 \right) m x_{-i}^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)}$$

given P and subject to:

$$N x_i + \left(\frac{P}{N} - 1 \right) N x_{-i} \left(\frac{1}{\tau} \right) - z = 0,$$

$$m(f + cz) - N\left(1 - kN^{1/2}\right) = 0,$$

$$N_{\min} \leqslant N \leqslant N_{\max}.$$

Here, m is the number of firms (varieties) allocated to each city, N is the population allocated to each city with P/N the number of cities implied (where P is the national population), z is the output produced by each firm, x_i is the quantity consumed of each local variety, and x_{-i} the quantity consumed of each imported variety. The first constraint conserves the output of each firm, while the second conserves the allocation of labor in each city. In the third constraint, $N_{\min}$ is the population necessary to produce a single variety, indicating minimum possible city size.

Two diametrically opposed special cases of the above formulation appeared in the earlier literature. One is that of Abdel-Rahman and Fujita (1990). They implicitly assumed that the cost of trading varieties between cities is infinite ($\tau = 0$). Since, in their model, varieties cannot be traded, the economy should produce each variety in every city and there will be no externalities conferred on one city when adding population to another. In this extreme case, the Henry George rule holds again and developers acting as in Section 4.2 decentralize the first-best optimal allocation with the number of local varieties behaving like a public good and land rents subsidizing fixed costs of setting up firms. The opposite case is one where varieties can be imported from other cities at zero cost ($\tau = 1$). This was assumed in Henderson and Abdel-Rahman (1991). Under their assumption, complete specialization is optimal and each variety is produced in a separate town of minimal size.

Anas (2004) showed that this complete specialization equilibrium occurs under any positive transport cost ($0 < \tau < 1$) as long as national population, P, is fairly large. In the complete specialization equilibrium, each city produces a homogeneous composite good and a single variety that is exported to all other cities. Note that the externality is at the level of aggregate population. Adding more people to the system allows more traded varieties to be produced in new specialized cities causing welfare to increase in every city. City size is determined by the labor needed to produce a single variety plus the locally needed composite good (if one exists). It is not surprising that complete specialization should be optimal when varieties can be traded at zero cost [as Henderson and Abdel-Rahman (1991)] had assumed, because there is no trading cost to save from the two products being produced in the same city. But the result in Anas (2004) that complete specialization is optimal even with positive trading cost is surprising. This occurs because as national population increases the centrifugal forces arising from trading are strengthened.

Hochman (1997) has examined a generalization of functional forms for the complete specialization case. His key point is that each specialized town is a natural monopoly (the marginal cost curve is everywhere below the average cost curve). He notes the presence of two distortionary effects:

(1) the inefficiency of monopolistic competition causes more varieties (cities) at the optimum;

(2) each firm (city) ignores the advantages of an increase in the number of varieties and this causes fewer and larger cities at the optimum as compared to the laissez-faire equilibrium.

The optimum allocation requires marginal cost pricing. This results in optimal cities being larger and each city producing more of its variety. Because the marginal cost curve is below the average cost curve, marginal cost pricing requires a subsidy being given to the firm as in the classic problem of natural monopoly regulation. However, the Henry George rule fails since the local aggregate land rent falls short of the aggregate subsidy required and the local government has neither the means nor the incentive to pay out such a subsidy. This means that, to achieve the first-best optimum, the central planner should subsidize each variety via a head tax levied on each consumer. As shown in Hochman (1997), this subsidy can be per unit of the variety produced or, as shown by Henderson and Abdel-Rahman (1991), it can be lump sum and equal to the fixed cost of the firm. Furthermore, when each completely specialized city produces a local variety and a homogeneous good as well, as was true in the non-spatial model of Dixit and Stiglitz (1977), then Henderson and Abdel-Rahman (1991) and Hochman (1997) conclude that neither the Henry George Theorem holds, nor can the planner's first-best optimum be decentralized.

There are other contexts in which central planners are shown to be essential. Anas and Xiong (2003) present a model in which two homogeneous goods are traded at cost among cities and so are the varieties that are used as intermediate inputs in the production of these goods. Under the same parameter values (including the unit iceberg trading costs) they examine two possible equilibria. In the first type of equilibrium, each city specializes in only one homogeneous good and produces some input varieties locally. In the second type of equilibrium, each city is diversified producing both homogeneous goods and varieties. Depending on parameter values, either one of the two equilibria can yield higher utility (and be optimal). Without a central planner, such an optimum may not be achievable. A planner is needed to specify the industry mix of the cities according to the optimum allocation. Without such planning, the city system can get stuck in the inferior equilibrium.

Yet another context in which central planning is necessary to achieve the optimum is illustrated by Anas and Xiong (1999). In this model, when cities are formed, there is a benefit to the rural sector in that rural population densities decrease and agricultural productivity increases. Hence, the optimum involves a cross-subsidy from rural to urban workers.

9. Growth

One basic challenge is to explain how the city size distribution evolves over time as population grows exogenously or as endogenous economic growth occurs. A hypothesis that has found some support in the literature is the *parallel growth hypothesis*. According to this hypothesis, population growth causes cities of different sizes to grow

at the same rate so that at any moment in time the relative sizes of alternative cities are unchanged. Eaton and Eckstein (1997) provide empirical support for the parallel growth hypothesis. They observed that the relative populations of the top 40 urban areas in France (1876–1990) and Japan (1925–1985) remained essentially unchanged. Black and Henderson (1999, p. 254) provided further support for this hypothesis: "Despite entry of new metropolitan areas, the relative size distribution of cities is astonishingly stable over time, exhibiting no tendency to collapse ("converge" to a common city size), spread, go bimodal, and so forth, with the actual distribution fluctuating little between decades." The parallel growth hypothesis may be questioned by observing the emergence of megacities all over the world, an observation we noted in the Introduction. If the trend were to continue, sometime in the future an increasingly large proportion of the world's population could come to reside in such megacities.

9.1. Exogenous population growth

The evolution of a city size distribution over time may be viewed as a tug-of-war between centripetal and centrifugal forces. These are not only operative at the level of an individual city, determining the size of that city as we saw in Sections 2 and 3, but they are also operative at the system of cities level. Centripetal forces favor concentration of economic activity in a small number of large cities, while centrifugal forces favor concentration of economic activity in a large number of small cities. The key question is whether population growth modifies the relative strength of these two opposing forces. If, as population grows, the centripetal forces become increasingly strong relative to the centrifugal force then the number of cities could decline with their sizes increasing. Conversely, if the centrifugal forces become increasingly strong with population growth, then the number of cities should increase with city sizes decreasing.

For a system of identically sized cities (e.g., a flat city-size distribution), this question has been posed in Anas (2004). Suppose that the exogenous system-wide population is expressed as $P(t) = n(t)N(t)$, where t is time, $n(t)$ is the number of cities as a function of time and $N(t)$ is the size of a single city. In the simplest Henderson-type setup that was examined in Section 6.1, cities are isolated (do not interact with each other); hence as population grows, city size remains unchanged while more cities are spawned. However, as we saw in Section 7, when the cities trade with each other, then the creation and size of one city confers an externality on all the others. Then, depending on the nature of trade, cities grow or shrink in size on the growth path. Anas (2004) identifies:

(a) concentration ($\dot{P}(t) > 0 \Rightarrow \dot{n}(t) < 0,\ \dot{N}(t) > 0$),
(b) balanced growth ($\dot{P}(t) > 0 \Rightarrow \dot{n}(t) \geqslant 0,\ \dot{N}(t) \geqslant 0$),
(c) de-agglomeration ($\dot{P}(t) > 0 \Rightarrow \dot{n}(t) > 0,\ \dot{N}(t) < 0$).[28]

[28] Kanemoto (1980) and Henderson and Ioannides (1981) have examined exogenous growth without trade.

Clearly, which pattern obtains will depend on how the theoretical model is specified. In Anas (2004) the model subjected to this test is that inspired by the assumptions of the New Economic Geography (NEG) adapted by him to a system-of-cities context. There is no agricultural sector as in the NEG,[29] but instead cities are organized as in Section 2 and the number of cities is endogenous. This urban economy of n cities produces m varieties and all nm varieties in the city system are consumed by each of the consumers, hence traded among the cities subject to iceberg transportation cost as in the NEG. The focus is the optimal behavior of the city system under a hypothetical central planner. Anas proves that such a city system *always* produces the third pattern of de-agglomeration in its most extreme form. As population increases exogenously, new cities are spawned on the welfare maximizing path but each city continues to get smaller and smaller until, eventually, all cities suddenly become completely specialized mini-factory towns producing a single variety that is traded to all the other cities. This result is extremely robust. Provided sufficient population growth occurs, this optimal de-agglomeration result is inevitable for all finite values of the Dixit–Stiglitz elasticity of substitution and the unit iceberg trading cost. What causes this robust de-agglomeration is the consumer's extreme taste for more varieties stemming from the Dixit–Stiglitz utility that is so central to the New Economic Geography as well as to the monopolistic-competition-based city size models of Abdel-Rahman and Fujita (1990). For a given aggregate population, more national varieties can be produced by splitting the population into small specialized towns because smaller towns require less commuting, afford more labor hours and in this way allow more varieties to be produced.

There are important implications of this result. *First*, it means that under NEG type of assumptions, observed city size distributions with cities getting larger as the world population grows cannot be efficient unless the population growth is accompanied by steep declines in the unit cost of commuting or by steep increases in the cost of trading goods and services between cities. Historically, it is true that commuting costs have fallen, but they have stabilized in recent times. But the cost of trading between cities has also fallen, not risen. This means that city-system theories that rely on Dixit–Stiglitz tastes (or on Ethier production functions) with intercity trading of differentiated goods and services may not be useful for modeling actual urbanization trends. *Second*, the de-agglomeration result suggests how better theory may be developed by modeling trade among cities in a more realistic manner.

9.2. Endogenous economic growth

Palivos and Wang (1996) constructed a one-sector, one factor dynamic equilibrium model of a single city in which city growth is endogenous. In this model, the engine of urban growth is the spillover of knowledge among individuals as in Romer (1986). The black-box production function of the only consumption good is $y = Ak^{\alpha}K^{1-\alpha}$

[29] For a criticism of the treatment of agriculture in the NEG, see Pines (2001).

where k and K denote per capita and aggregate human capital in the city. Thus, higher population in the city implies higher aggregate human capital stock and thus higher productivity at the individual level (due to interpersonal spillovers). As in all urban models, the growth of the city is bounded due to the rise in the average transportation cost as a result of the physical increase in the size of the city (see Section 2). It is assumed that the consumer has a constant elasticity of intertemporal substitution and a constant rate of time preference. The paper characterized the steady-state equilibrium growth path in a decentralized economy as well as the socially optimal growth path. In the decentralized economy, each agent determines the paths of consumption and investment in human capital and a developer determines the city's population. In the social optimum, the developer chooses the optimal growth path of consumption and investment as well as the population. The main result of the paper is that the optimal city size and growth rate are lager on the optimal path than on the equilibrium path, because central planning by the developer internalizes the spillover externality. The model is identical in form to the labor productivity model of city size described in Section 3.3 and, hence, the aggregate human capital level is equivalent to a public good. Unfortunately, the authors do not examine the implications of their model in a system of cities framework.

Ioannides (1994) synthesized the Dixit–Stiglitz model of product diversity and monopolistic competition (see Section 3.2) with Romer's (1987) model of increasing returns due to specialization. In this model, which employs an overlapping generations formulation, consumers invest their life savings in urban overhead capital, enabling producers to constantly create new product varieties. Because the cost of trading varieties between cities is zero, there is no home market effect and, as in Henderson and Abdel-Rahman (1991), each variety is produced in a separate and completely specialized city organized by a developer. Thus, growth in the economy is tied to the growth in the number of varieties, same thing as the growth in the number of cities. In the steady state of such a model, with all cities identical, a balanced agglomeration path is exhibited with the number of cities growing at the same rate as aggregate population and the size of each city remaining constant.

Black and Henderson (1999) developed a model of urban growth of a system of two types of cities. The economy produces an intermediate and a final good. Production of the final good requires the intermediate good as an input, but because the cost of shipping the intermediate good is assumed to be zero, locating the two goods in the same city causes only crowding effects with no offsetting benefit. Hence, at equilibrium there are two types of specialized cities, one city type producing the intermediate good while the other type specializing in the final good and importing the intermediate good. Both goods are produced using workers. Each worker is a firm. The production function of a worker uses that worker's human capital as an input, but a worker's productivity depends on two external effects as well. One of these is the total number of workers in the same city. This measures the Marshallian flow of creative ideas or information that can be obtained freely from other workers. The other external effect is the average level of human capital in the city measuring the richness of the information flowing from the labor force. This formulation is equivalent to the labor productivity model of

our Section 3.3. At the level of the worker production is constant returns, but the social marginal product exceeds the private marginal product because of the external effects. Under the assumption that all workers are identical, each worker's human capital level equals the average level of human capital in the city of that worker. Each city is set up by a developer (see Section 4.2) and is self-financing under the Henry George rule. Equilibrium as well as socially optimal growth paths are analyzed.

In models with several types of cities and human capital accumulation at the level of a worker, a problem arises from the interaction of the migration decision with the human capital accumulation decision. That is because workers would choose to switch cities over their lives and they would accumulate human capital according to the sequence of city types they resided in.[30] To avoid this complexity, Black and Henderson assume that workers belong to an infinitely lived dynasty that is centrally administered. The size of the dynasty grows exogenously at a constant rate. The dynasty maximizes the lifetime discounted utility of its representative member (a worker). Each worker is allocated the same amount of consumption regardless of the city type in which that worker lives. The workers' savings (income net of commuting, rent and consumption) becomes new human capital. The dynasty allocates the human capital contributed by its workers in a city to its newborn members in that city. The dynasty also decides what fraction of its members should live in type 1 and type 2 cities. The model provides a framework for studying the effects of urbanization and city size distribution on economic growth as well as the effects of economic growth on urbanization, since the number of each type of city is endogenously determined. At steady state, the model can replicate each of the growth patterns identified above (concentration in fewer cities, balanced growth and de-agglomeration). More precisely, the number of cities increases if human capital accumulation is not strong enough. Otherwise, if human capital accumulation is strong enough, then the number of cities decreases. City sizes for each city type grow at approximately twice the rate of human capital accumulation. Thus, the model replicates the parallel growth hypothesis.[31] The social optimum involves higher consumption and human capital accumulation relative to the equilibrium since, in the equilibrium, the positive externalities from the localization of production are not internalized.

Black and Henderson mention that between 1900 and 1950, the average metropolitan population in the U.S. tripled, while the number of metro areas doubled under a national population growth rate of 1.4% per year. In the same period, the percentage

[30] Eaton and Eckstein (1997) treat migration more realistically in that individual consumers migrate from city to city over their lives but become assimilated to the average human capital level of the city in which they reside. However, in their model, the number of cities is constant and not endogenous As explained in the introduction, Black and Henderson and Eaton and Eckstein claim empirical support for the parallel growth hypothesis.

[31] The fact that all three growth patterns are possible under human capital accumulation is an important result because in models of city systems based on the New Economic Geography under costly trade between cities but no human capital, city sizes continually decrease under exogenous population growth [see Anas (2004) for the proof].

of the 17-year-old population that had completed high school rose nine-fold from 6.3 to 57.4 percent nationally (p. 254). These numbers correspond to a compound national population growth rate ($\dot{P}/P$) of 1.4% per year, a city size growth rate ($\dot{N}/N$) of 2.2% per year, a growth rate in the number of metro areas ($\dot{n}/n$) of 1.396% per year, and a human capital growth rate ($\dot{h}/h$) of 4.5% per year. From $P = Nn$, it follows that

$$\frac{\dot{n}}{n} + \frac{\dot{N}}{N} = \frac{\dot{P}}{P}.$$

In the model, city growth is proportional to human capital growth by the equation

$$\frac{\dot{N}}{N} = 2\varepsilon\frac{\dot{h}}{h}$$

where ε is the elasticity of income with respect to human capital. From the annualized growth rates given above, it is implied that $\varepsilon = 0.244$ (a 10% increase in human capital should generate only a 2.44% increase in income). Meanwhile, from

$$\frac{\dot{n}}{n} = \frac{\dot{P}}{P} - \frac{\dot{N}}{N} = 0.014 - 0.022 = -0.008,$$

implying a slight annual rate of decrease in the number of cities, not the doubling observed by Black and Henderson. The empirical part of the paper reports regressions that establish a strong association between metropolitan-area human capital levels and metropolitan populations. However, the fact that the theoretical model does not appear consistent with the data (as just explained) may be suggestive that human capital growth rates may be only part of the explanation for predicting city growth rates.

10. Challenges ahead

Systems of cities models are challenged by several empirical observations that have gained relevance and validity in recent decades. The first of these is the apparently growing trend toward larger and larger cities and the concentration of the world's population in such cities. This is sometimes referred to as the primacy phenomenon and is recognized as a persistent violation of the rank–size rule. As we saw earlier, existing models of human capital accumulation in cities do not capture this phenomenon because they have so far focused exclusively on the parallel growth hypothesis. Growth models are needed to explain how some cities grow relatively larger than others and how they become diversified as they grow. Yet, there is no dynamic model at the present time that captures the relationship between growth and diversification. The intuitive notion that diversification speeds up the growth process has found some empirical support [Glaeser et al. (1992)] but a solid theory is sorely lacking.

A second area of concern is the growth of polycentric cities all over the world. While this trend has been widely recognized [see Anas, Arnott and Small (1998)], all systems of cities models to date have relied on the monocentric model of an urban area (see

Section 2). The problem with this approach is that it suppresses an important margin across which dispersion takes place. As population grows, new cities are set up. But instead of accommodating additional population in new cities, existing cities can spawn new subcenters. Subcenter formation is a major way in which existing cities get bigger, accommodating a larger economic base. Some urban economists have developed models of subcenter formation and urban job dispersion [see Anas and Kim (1996), Anas and Xu (1999) for completely closed general equilibrium formulations based on pecuniary externalities or Lucas and Rossi-Hansberg (2002) for models based on Marshallian spillovers], but these have not been integrated into systems of cities models. These thoughts are not unrelated to those in the previous paragraph, because the ability to form subcenters encourages bigger and bigger cities and the spawning of urban subcenters in turn may be a major determinant of faster economic growth.

Finally, the idea of iceberg trading cost introduced into the New Economic Geography by Krugman (1991), hides strong dispersive tendencies. As shown in Anas (2004), cities vanish under population growth except in the case where trading costs are infinite. This suggests that the setup of the NEG can benefit greatly by a synthesis of the NEG with the models of human capital accumulation surveyed in Section 9, in order to more meaningfully explore the role of trade on city formation in a growing economy.

Acknowledgements

An earlier version of this chapter was presented at the 49th North American Meetings of Regional Science Association International, held in San Jose, Puerto Rico in November 2002. We are grateful to Robert Helsleg, our discussant, to David Pines for his careful reading of the paper and for his detailed comments and to the editors of this volume for several suggestions about the exposition.

References

Abdel-Rahman, H.M. (1988). "Product differentiation, monopolistic competition and city size". Regional Science and Urban Economics 18, 69–86.

Abdel-Rahman, H.M. (1990a). "Agglomeration economies, types, and sizes of cities". Journal of Urban Economics 27, 25–45.

Abdel-Rahman, H.M. (1990b). "Shareable inputs, product variety, and city sizes". Journal of Regional Science 30, 359–374.

Abdel-Rahman, H.M. (1994). "Economies of scope in intermediate goods and a system of cities". Regional Science and Urban Economics 24, 497–524.

Abdel-Rahman, H.M. (1996). "When do cities specialize in production?" Regional Science and Urban Economics 26, 1–22.

Abdel-Rahman, H.M. (1998). "Income disparity, time allocation and social welfare in a system of cities". Journal of Regional Science 38, 137–154.

Abdel-Rahman, H.M. (2000). "Multi-firm city versus company town: a microfoundation model of localization economy". Journal of Regional Science 40, 755–769.

Abdel-Rahman, H.M. (2002). "Does the structure of an urban system affect income disparities?" Journal of Regional Science 42, 389–410.

Abdel-Rahman, H.M., Fujita, M. (1990). "Product variety, Marshallian externalities and city sizes". Journal of Regional Science 30, 165–183.

Abdel-Rahman, H.M., Fujita, M. (1993). "Specialization and diversification in a system of cities". Journal of Urban Economics 33, 189–222.

Abdel-Rahman, H.M., Wang, P. (1995). "Toward a general-equilibrium theory of a core–periphery system of cities". Regional Science and Urban Economics 25, 529–546.

Abdel-Rahman, H.M., Wang, P. (1997). "Social welfare and income inequality in a system of cities". Journal of Urban Economics 41, 462–483.

Alonso, W. (1964). Location and Land Use. Harvard University Press, Cambridge, MA.

Anas, A. (1992). "On the birth and growth of cities: laissez-faire and planning compared". Regional Science and Urban Economics 22, 243–258.

Anas, A. (2004). "Vanishing cities: what does the New Economic Geography imply about the efficiency of urbanization?" Journal of Economic Geography, 181–199.

Anas, A., Arnott, R.J., Small, K. (1998). "Urban spatial structure". Journal of Economic Literature 36, 1426–1464.

Anas, A., Kim, I. (1996). "General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration". Journal of Urban Economics 40, 217–232.

Anas, A., Xiong, K. (1999). "Public investment in a core–periphery model". Working Paper. State University of New York at Buffalo.

Anas, A., Xiong, K. (2003). "Intercity trade and the industrial diversification of cities". Journal of Urban Economics 54, 258–276.

Anas, A., Xiong, K. (2004). "The formation and growth of specialized cities: efficiency without developers or Malthusian traps". Regional Science and Urban Economics. In press.

Anas, A., Xu, R. (1999). "Congestion, land use and job dispersion: a general equilibrium analysis". Journal of Urban Economics 45 (3), 451–473.

Arnott, R.J. (1979). "Optimal city size in a spatial economy". Journal of Urban Economics 6, 65–89.

Arnott, R.J., Stiglitz, J.E. (1979). "Aggregate land rents, expenditure on local public goods and optimal city size". Quarterly Journal of Economics 93, 471–500.

Becker, G., Murphy, K. (1992). " The division of labor, coordination costs, and knowledge". Quarterly Journal of Economics 107, 1137–1160.

Black, D., Henderson, J.V. (1999). "A theory of urban growth". Journal of Political Economy 107 (2), 252–284.

Chamberlin, E.H. (1933). The Theory of Monopolistic Competition. Harvard University Press, Cambridge.

Chipman, J.S. (1970). "External economies of scale and competitive equilibrium". Quarterly Journal of Economics 84, 347–385.

Diamond, P. (1982). "Aggregate demand management in search equilibrium". Review of Economic Studies 49, 217–227.

Dixit, A. (1973). "The optimum factory town". Bell Journal of Economics and Management Science 4, 637–654.

Dixit, A., Stiglitz, J.E. (1977). "Monopolistic competition and optimal product diversity". American Economic Review 67, 297–308.

Duranton, G., Puga, D. (2001). "Nursery cities: urban diversity, process innovation, and the life cycle of products". American Economic Review 91, 1454–1477.

Duranton, G., Puga, D. (2004). "Micro-foundations of urban agglomeration economies". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam. This volume.

Eaton, J., Eckstein, Z. (1997). "Cities and growth: theory and evidence from France and Japan". Regional Science and Urban Economics 27, 443–474.

Ethier, W. (1982). "National and international returns to scale in the modern theory of international trade". American Economic Review 72, 389–405.

Flatters, F., Henderson, J.V., Mieszkowski, P. (1974). "Public goods, efficiency and regional fiscal equalization". Journal of Public Economics 3, 99–112.

Fujita, M., Krugman, P., Mori, T. (1999). "On the evolution of hierarchical urban systems". European Economic Review 43, 201–259.

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

Fujita, M., Mori, T. (1997). "Structural stability and evolution of urban systems". Regional Science and Urban Economics 27, 399–442.

Gabaix, X., Ioannides, Y.M. (2004). "The evolution of city size distributions". "Theories of systems of cities". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam. This volume.

Glaeser, E., Kallal, H.D., Scheinkman, J.A., Shleifer, A. (1992). "Growth in cities". Journal of Political Economy 100, 1126–1152.

Goldstein, G.S., Gronberg, T.J. (1984). "Economies of scope and economies of agglomeration". Journal of Urban Economics 16, 63–84.

Hadar, Y., Pines, D. (2003). "On the market failure in a Dixit–Stiglitz setup with two trading cities". Journal of Public Economic Theory 5 (4), 549–570.

Hadar, Y., Pines, D. (2004). "Population growth and its distribution between cities: positive and normative aspects". Regional Science and Urban Economics 34 (2), 125–154.

Helpman, E., Pines, D. (1980). "Optimal public investment and dispersion policy in a system of open cities". American Economic Review 70, 507–514.

Helsley, R.W., Strange, W.C. (1990). "Matching and agglomeration economies in a system of cities". Regional Science and Urban Economic 20, 189–212.

Helsley, R.W., Strange, W.C. (1993). "Agglomeration economies and urban capital markets". Journal of Urban Economics 29, 96–112.

Helsley, R.W., Strange, W.C. (1994). "City formation with commitment". Regional Science and Urban Economics 24, 373–390.

Henderson, J.V. (1974). "The sizes and types of cities". American Economic Review 64, 640–656.

Henderson, J.V. (1977). Economic Theory and the Cities. Academic Press, New York.

Henderson, J.V., Abdel-Rahman, H.M. (1991). "Urban diversity and fiscal decentralization". Regional Science and Urban Economics 21, 491–510.

Henderson, J.V., Becker, R. (2000). "Political economy of city sizes and formation". Journal of Urban Economics 48, 453–484.

Henderson, J.V., Ioannides, Y.M. (1981). "Aspects of growth in a system of cities". Journal of Urban Economics 10, 117–139.

Henderson, J.V., Kuncoro, A., Turner, M. (1995). "Industrial development in cities". Journal of Political Economy 103 (5), 1067–1090.

Henderson, J.V., Thisse, J.-F. (2001). "On strategic community development". Journal of Political Economy 109, 546–569.

Hercowitz, Z., Pines, D. (1991). "Migration with fiscal externalities". Journal of Public Economics 46 (2), 163–180.

Hobson, P. (1987). "Optimal product variety in urban areas". Journal of Urban Economics 22, 190–197.

Hochman, O. (1981). "Land rents, optimal taxation and local fiscal independence in an economy with local public goods". Journal of Public Economics 15, 290–310.

Hochman, O. (1997). "More on scale economies and cities". Regional Science and Urban Economics 27, 373–397.

Ioannides, Y. (1994). "Product differentiation and economic growth in a system of cities". Regional Science and Urban Economics 24, 461–484.

Jacobs, J. (1969). The Economy of Cities. Vintage, New York.

Juhn, C., Murphy, K., Pierce, B. (1993). "Wage inequality and rise in returns to skill". Journal of Political Economy 101, 410–442.
Kanemoto, Y. (1980). Theories of Urban Externalities. North-Holland, Amsterdam.
Kim, S. (1991). "Heterogeneity of labor markets and city size in an open spatial economy". Regional Science and Urban Economics 21, 109–126.
Krugman, P. (1980). "Scale economies, product differentiation and the pattern of trade". American Economic Review 70, 950–959.
Krugman, P. (1991). "Increasing returns and economic geography". Journal of Political Economy 99, 483–499.
Lösch, A. (1954). Die Räumliche Ordnung der Wirtschaft. Yale University Press, New Haven. Translated by Woglom, W.H. and Stopler, W.F. as *The Economics of Location*.
Lucas, R. (1988). "On the mechanics of economic development". Journal of Monetary Economics 22, 3–42.
Lucas, R., Rossi-Hansberg, E. (2002). "On the internal structure of cities". Econometrica 70 (4), 1445–1476.
Machin, S. (1996). "Wage inequality in the UK". Oxford Review of Economic Policy 12, 47–64.
Marshall, A. (1890). Principles of Economics. MacMillan, London.
Mills, E.S. (1967). "An aggregative model of resource allocation in a metropolitan area". American Economic Review 61, 197–210.
Muth, R. (1969). Cities and Housing. University of Chicago Press, Chicago, IL.
National Geographic (2002). "Megacities: the coming urban world". National Geographic (December), 70–99.
Palivos, T., Wang, P. (1996). "Spatial agglomeration and endogenous growth". Regional Science and Urban Economics 26, 645–670.
Panzar, J.C., Willig, R.D. (1981). "Economies of scope". American Economic Association Papers and Proceedings 71, 268–272.
Papageorgiou, Y., Pines, D. (1999). An Essay on Urban Economic Theory. Kluwer Academic, Boston.
Papageorgiou, Y., Pines, D. (2000). "Externalities, indivisibilities, nonreplicability and agglomeration". Journal of Urban Economics 48, 509–535.
Pines, D. (2000). "On alternative urban growth patterns". Unpublished manuscript. The Eitan Berglas School of Economics, Tel-Aviv University.
Pines, D. (2001). "New economic geography: Revolution or counter revolution?" Journal of Economic Geography 1, 139–146.
Rivera-Batiz, F.L. (1988). "Increasing returns, monopolistic competition and agglomeration economies in consumption and production". Regional Science and Urban Economics 18, 25–153.
Romer, P. (1986). "Increasing returns and long-run growth". Journal of Political Economy 94, 1002–1037.
Romer, P. (1987). "Growth based on increasing returns due to specialization". American Economic Review 77, 56–62.
Rosenthal, S.S., Strange, W.C. (2004). "Evidence on the nature and sources of agglomeration economies". In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics, vol. 4. Elsevier, Amsterdam, pp. 2119–2171. This volume.
Scotchmer, S. (1986). "Local public goods in an equilibrium: how pecuniary externalities matter". Regional Science and Urban Economics 16, 463–481.
Stiglitz, J.E. (1977). "The theory of local public goods". In: Feldstein, M.S., Inman, R.P. (Eds.), The Economics of Public Services. MacMillan, London.
Tabuchi, T. (1998). "Urban agglomeration and dispersion: a synthesis of Alonso and Krugman". Journal of Urban Economics 44, 333–351.
United Nations (1996). United Nations Human Development Report. United Nations Development Program, Washington, DC.
Upton, C. (1981). "An equilibrium model of city sizes". Journal of Urban Economics 10, 15–36.
Wilson, J.D. (1987). "Trade in a Tiebont economy". American Economic Review 77, 431–441.