ANNALS OF ECONOMICS AND FINANCE **4**, 275–341 (2003)

# Urbanization and Economic Development

J. Vernon Henderson

*Brown University*

This paper provides a survey and guide to the literature relevant to urbanization and economic development. The paper starts with some basic facts and trends about urbanization worldwide. It then reviews the traditional two-sector urban-rural model, but focuses on the modern version, Krugman's core-periphery model. However, two sector models do not capture the notion of an economy composed of many cities; nor do they represent modern agglomeration economies. Models and empirical evidence on agglomeration economies are reviewed. Then the paper turns to empirical evidence on the evolution of the size distribution of cities. It reviews the large literature on systems of cities models, focusing on an endogenous growth version. This part of paper concludes with a review of recent work integrating systems of cities models with the new economic geography. The final section reviews urbanization in China, focusing on policy issues such as migration, under-agglomeration and spatial biases in the FDI policy. 

*Key Words*: Urbanization; China reginal development; Systems od cities.
*JEL Classification Numbers*: O0, R0.

## 1. INTRODUCTION

Urbanization occurs as countries switch sectoral composition away from agriculture into industry and as technological advances in domestic agriculture release labor from agriculture to migrate to cities. Given this well accepted process, the study of urbanization with development focuses on three issues. For each of these, this paper will review key empirical facts and evidence and explain the key theoretical models used in analysis. In the last section, I turn to China, using the impacts of China's urbanization policies, to illustrate aspects of the first three sections.

The first issue concerns whether the urbanization process involving rural to urban migration within countries is reasonably efficient, or whether it is subject to forms of market failure or distortionary government policies. Part of the literature on the subject looks at the basic overall rural-urban



1529-7373/2002

divide to ask whether countries are over- or under-urbanized. That particular narrow question is not what the recent economics literature has focused on, for reasons we will see. Rather the literature has focused on the form that urbanization takes. In some writings form means the development and then perhaps subsequent reversal of a core-periphery spatial, or regional structure. In other writings, it means the development and then subsequent reversal of a high degree of urban primacy, or the degree of dominance of one city over other cities in a region. How does spatial concentration, in terms of, say, the share of the core region or primate city in the economy evolve with development? What are the efficiency implications of more or less, or of too much or too little spatial concentration?

The second issue concerns why industrialization involves urbanization. What market and non-market interactions lead economic activity to spatially cluster, or agglomerate into entities we call cities? There are a variety of papers which model the form of localized scale externalities such as information spillovers in output and input markets and backward and forward linkages which lead to agglomeration; and there is a large body of empirical work trying to measure the nature and extent of scale externalities. Finally there is a more recent literature examining dynamic externalities and localized knowledge spillovers.

The third issue concerns how cities form and interact with each other, in an urban system in both static and dynamic contexts. Rather than a simple core-periphery regional structure an economy is composed of an endogenous and potentially large number of cities of different sizes and types. The country's urban system can be viewed as a whole, or there can be core and periphery regions each with their own system of cities. Empirical evidence shows that over long periods of time within countries there tends to be a "wide" and very stable relative size distribution of cities. The natural questions then are what is the role of big versus small cities in a country – i.e., in what do they tend to specialize and how do they interact with each other? Second, what is the inter-relationship between national economic growth and growth of both individual cities and the overall urban system? The theory papers attempt to model all these questions, and the underlying facts about urban systems. Apart from providing a link between national and city growth, from a development perspective this literature indicates how national urban development evolves. This has implications for national policy governing the spatial allocation of public infrastructure investments, fiscal decentralization, internal migration policies and the like.

## 2. URBANIZATION AND ITS FORM

Urbanization, or the shift of population from rural to urban environments, is a transitory process, albeit one that is socially and culturally

traumatic. It moves populations from traditional-cultural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. It spatially separates families, particularly intergenerationally as the young migrate to cities and the old stay behind. By upper middle income ranges countries become "fully" urbanized, with 60-90% of the national population living in cities, with the actual percent urbanized varying with geography, role of agriculture, and national definitions of urban.

The idea that urbanization is a transitory phenomenon is born out by the simple statistics in Figure 1, comparing different regions of the world in 1960 versus 1995. While urbanization increased in all regions of the world over those 35 years, among developed countries there is little change since 1975. Soviet bloc and Latin American countries have almost converged to developed country urbanization levels.

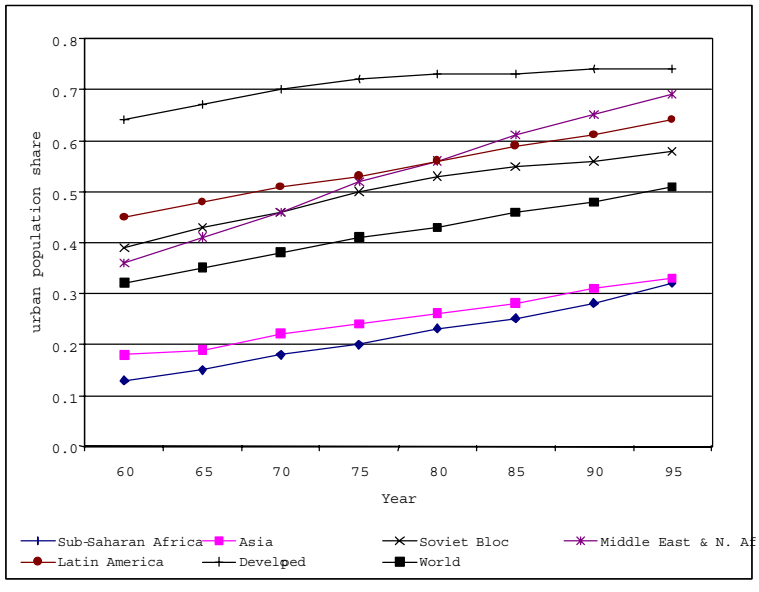
Despite this notion of urbanization being a transitory phenomenon, we don't actually have a good conceptual model of the dynamic transitory process. Models of urbanization per se are, oddly, static. The traditional versions focus on the question of urban "bias", or the effect of government policies on the urban-rural divide, or the efficient rural-urban allocation of population at a point in time. These models are the long-standing dual economy models, that date back to Lewis (1954). They are two sector models with an exogenously given sophisticated urban sector and a "backward" rural sector (Rannis and Fei (1961), Harris and Todaro (1970) and others as now well exposited in textbooks (e.g., Ray (1998)).

Dual sector models presume an exogenously given situation where the productivity of labor in the urban sector exceeds that in the rural sector. Arbitrage in terms of labor migration is limited by inefficient labor allocation rules such as farm workers being paid average rather than marginal product or artificially limited absorption in the urban sector (e.g., formal sector minimum wages). The literature focuses on the effect on migration from the rural to urban sector of policies such as rural-urban terms of trade, migration restrictions, wage subsidies, and the like.

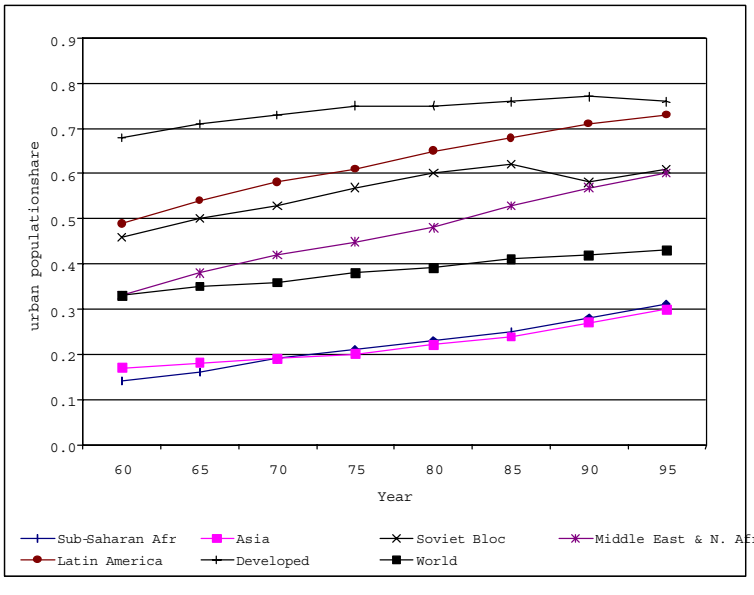
The final and most complex version of the models are the Kelley and Williamson (1998) and the Becker, Mills, and Williamson (1984), which are CGE models which introduce dynamic elements. They have savings behavior and capital accumulation, population growth, and multiple economic sectors in the urban and rural regions. Labor markets within sector and across regions are allowed to clear. The multiple economic sectors allow consideration of the effects of a wider array of policy instruments, including sector specific trade or capital market policies for housing, industry, services and the like. However the starting point is again an exogenously given initial urban-rural productivity gap sustained initially by migration costs and exogenous skill acquisition. On-going urbanization is the result

**FIG. 1.** Share of Urban Population in Total Population.

(a) Average over Countries




(b) Weighted Average, Using Country Population.



of exogenous forces – technological change favoring the urban sector or changes in the terms of trade favoring the urban sector.

As models of urbanization, these dual economy ones were a critical step but they suffer obvious defects, apart from their rather static nature. First how the dual starting point arises is never modeled. Second, and related to the first as we will see, there are no forces for agglomeration that would naturally foster industrial concentration in the urban sector. Finally although the models have two sectors there is really little spatial or regional aspect to the problem. There is a new generation of two-sector models, the core-periphery models, which attempt to address to differing degrees these three defects. However core-periphery models are not really about urbanization per se, since in many versions including Krugman's (1991a) initial piece the agricultural population is fixed. The models ask under what conditions in a two-region country, both regions versus only one region industrializes or urbanizes. In application to the development process, I interpret these models as starting to analyze the form urbanization takes. Before turning to these models, I review the limited empirical evidence first on urbanization and then on the form of urbanization in terms of core-periphery structures. Then I turn to the theoretical literature in economic geography on core-periphery structures.

### 2.1. What Do We Know About Urbanization and Its Form?

There are several important facts that we know about the urbanization process. We briefly review these and then turn to the bulk of the literature devoted to the form that urbanization takes. That literature leads to the core-periphery models.

#### *2.1.1. Urbanization*

The dual-economy models typically take as given the desirability of ongoing urbanization. They then ask what types of market failures or government policies work to hinder the needed migration. The focus has been on "urban bias". Renaud (1981) makes the simple point that, in general, government policies bias, or influence urbanization through their effect on national sectoral composition. So policies affecting the terms of trade between agriculture and modern industry or between traditional small town industry (textiles, food processing) and high tech large city industry affect the rural-urban or small-big city allocation of population. Such policies include tariffs, and price controls and subsidies, and are analyzed in the system of cities models discussed in section 3.

The idea that (1) urbanization reflects changes in sector composition and (2) government policies affect urbanization primarily through their effect on sector composition is a key point of empirical studies of urbanization

by Fay and Opal (1999) and Davis and Henderson (2001). These studies argue that urbanization which occurs in the early and middle stages of development is determined largely by changes in national economic sector composition and government policies tend to affect urbanization indirectly through their effect on sector composition. Of course it is also possible that with or without sector distortions, migration from rural to urban areas can be influenced by wage policies as in the dual-economy literature or by migration restrictions, as in former planned economies such as China (Au and Henderson (2002)).

A second point about urbanization is that writers such as Gallup, Sacks and Mellinger (1999) suggest that urbanization may "cause" economic growth, rather than emerge as part of the growth, sectoral change process. The limited evidence so far suggests urbanization doesn't cause growth. Henderson (2002a) finds no econometric evidence linking the extent of urbanization to either economic or productivity growth or levels, per se. That is if a country increases its degree of urbanization per se, typically it doesn't grow faster. In a more refined version of growth and urbanization links, so far we have been unable to quantify for different levels of development, the "optimal" degree of urbanization. For each level of development there should be an optimal degree of urbanization where either over- or under-urbanization detract from growth. While that may make sense, econometric evidence doesn't support the idea, perhaps because the data are problematical or because in sub-Saharan Africa, rapid urbanization over the last thirty years is correlated with negative or zero economic growth.

Finally there is an informal notion (World Bank (2000)) that urbanization follows the same stages as population growth (the "demographic" transition between falling death rates and falling fertility rates) – an S-shaped relationship where population growth is slow at low levels of development, then there is a period of rapid acceleration in intermediate stages, followed by a slowing of growth. These differential growth population rates imply an S-shaped relationship between population levels and GDP per capita. These ideas do not seem to carry over to the urbanization process. Davis and Henderson (2001) find a simple concave relationship between the level of urban population and GDP per capita (with or without controlling for national population), at least over the last 35 years. Urbanization is most rapid at low income levels, tapering off from there until a country is fully urbanized.

Figure 2 illustrates where the percent urban is a concave function of income per capita. In Figure 3 a similar relationship is posited. There the relationship between total national urban population and income per

**FIG. 2.** Percent Urban and Development Level, 1965-95

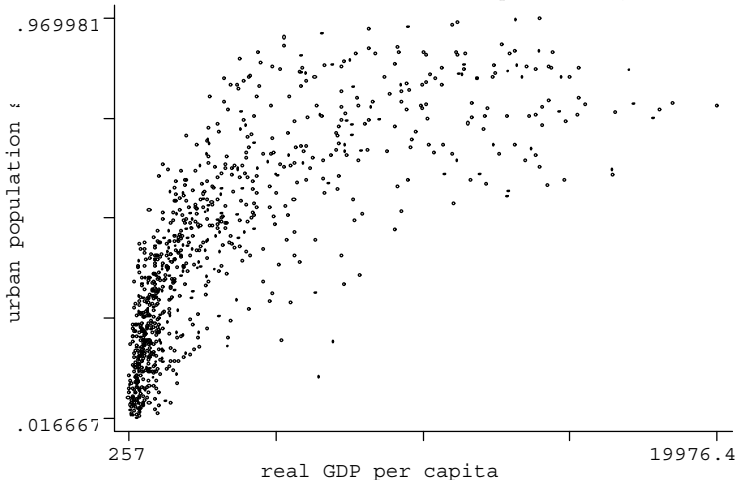
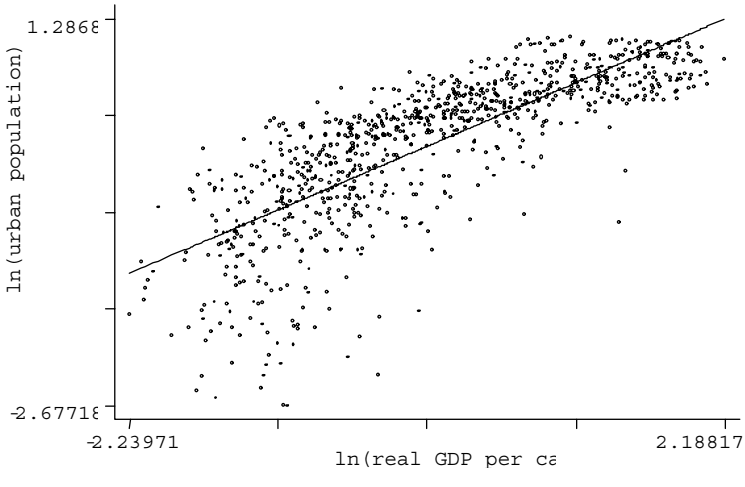



**FIG. 3.** Partial Correlation Between ln(urban population) and ln(real GDP per capita), Controlling for ln(national population).




capita is explored after parcelling out the effect of national population, or country size. In Figure 3 the log of national urban population is an increasing concave function of the log of income per capita, so national urban population will generally also be a concave function of income per capita.

*2.1.2. The Form of Urbanization: The Degree of Spatial Concentration*

In 1965, Williamson published a key paper based on cross-sectional analysis of 24 countries in which he argued that national economic development is characterized by an initial phase of internal regional divergence, followed by a phase of later convergence. That is, a few regions initially experience accelerated growth relative to other (peripheral) regions, but later the peripheral regions start to catch up. Barro and Sala-i-Martin (1991 and 1992 present extensive evidence on this for the USA, Western Europe, and Japan, by examining the evolution of inter-regional differences in per capita incomes. While inter-regional out-migration from poorer regions plays a role in catch-up, it may not be critical. In fact for Japan, the authors argue that later convergence of backward regions occurred in the absence of a real role for migration. Instead, productivity improved in backward regions.

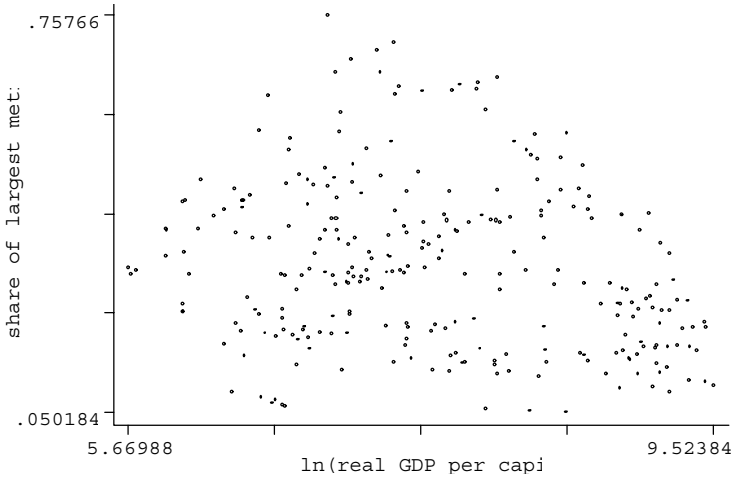
The urban version of this divergence-convergence phenomenon looks at urban primacy. Following Ades and Glaeser (1995), conceptually the urban world is collapsed into two regions – the primate city versus the rest of the country, or at least the urban portion thereof. Like dual sector models the focus is on how government policies and institutions affect primacy, with strong political-economy considerations. The basic question concerns to what extent urbanization is confined to one (or a few) major metro areas, relative to being spread more evenly across a variety of cities. That is, to what extent is urbanization concentrated? Primacy is the simplest measure, where a common measure of primacy is the ratio of the population of the largest metro area to all urban population in the country (Ades and Glaeser (1995), Junius (1999), and Davis and Henderson (2001)). A more comprehensive measure might use a Hirschman-Herfindal index [HHI] from the industrial organization literature, which is the sum of squared shares in national urban population of every metro area. That is a tremendous data gathering exercise, so far attempted only by Wheaton and Shishido (1981) for a single year.

What these papers find is an inverted $U$-shape relationship where urban-concentration first increases, peaks, and then declines with economic development. Despite different concentration measures and methods, Wheaton and Shishido (1981) examining a HHI using cross-section non-linear OLS and Davis and Henderson (2001) examining primacy using panel data methods and IV estimation find that urban primacy rises, peaks in the $2000-4000 range (1985 PPP dollars), and then declines. Junius (1999) finds a peak at somewhat higher income levels, but still the inverted $U-$shape. As

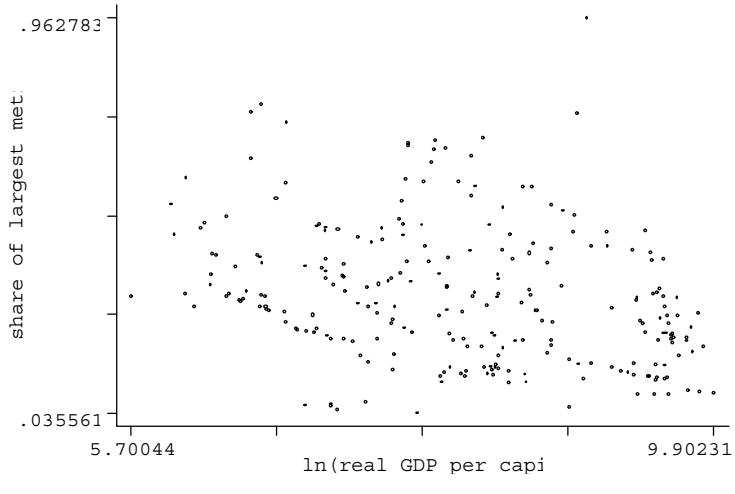
Figure 4 illustrates however the inverted $U$-relationship is noisy and more relevant in earlier (1965-75) than later (1985-95) time periods.

**FIG. 4.** Primacy and Economic Development.

(a) Early period: 1965-75.




(b) Recent Period: 1985-95.




Lee (1997) explores a case study of Korea. Seoul's urban primacy peaked around 1970 and while Seoul's absolute population has continued to grow, its share has declined steadily. What is of particular interest, especially in thinking about later core-periphery models is the role of manufacturing. At the urban primacy peak in 1970, Seoul had a dominant share of national

manufacturing although Pusan and Taegu had also developed large shares. During the next 10-15 years as Lee (1997) shows, manufacturing first suburbanized from Seoul to satellite cities in the rest of Kyonggi province (its immediate hinterland), as well as to satellite cities surrounding Pusan and Taegu. Such suburbanization of manufacturing has been also documented for Thailand (Lee (1998)), Colombia (Lee (1989)), and Indonesia (Henderson, Kuncoro and Nasution (1996)). But the key development following the early 1980's in Korea is the spread of manufacturing from the three major metro areas (Seoul, Pusan, and Taegu) and their satellites to rural areas and other cities. The share of rural areas and other cities in manufacturing in 1983 is 26%; by 1993 it is 42%, in a time period where (1) national manufacturing employment is fairly stagnant and (2) rural areas and other cities actually continue to experience modest absolute population losses. That is, manufacturing deconcentrated both relatively and absolutely to hinterland regions, where population levels were at best stagnant. This manufacturing deconcentration coincided with economic liberalization, enormous and widespread investment in inter-regional transport and infrastructure investment, and fiscal decentralization (Henderson, Lee, and Lee (2001)).

Apart from documenting the concentration-deconcentration process this empirical literature focuses on two critical sets of issues. First concerns the role of political economy and government policies in the process, building upon the concerns from the dual economy literature (Ades and Glaeser (1995)). Second is the issue of the relation of spatial concentration to growth. On the first set of issues the basic idea is that national policy makers favor the national capital (or other seat of political elites such as São Paulo in Brazil) for reasons of personal gain or beliefs about its inherent productivity advantage. For example, restraints on trade for hinterland cities favor firms in the national capital. Policy makers and bureaucrats may gain as shareholders in such firms or they may gain rents from those seeking licenses or other exemptions to trade restraints. What sort of restraints operate? Henderson and Kuncoro (1996) for Indonesia discuss the spatially centralized allocation mechanism for export and import licenses and for the granting of large bank loans. Centralization means hinterland bureaucrats can't grant such items and hence can't compete in the rent seekings process; the benefits of rent seeking for those items is the monopoly of central bureaucrats and officials. Trade protection for the primate city can also involve under-investment in hinterland transport and communications infrastructure.

Whether as true beliefs or as a justification to cover rent-seeking behavior, policy makers in different countries articulate a view that large cities

are more productive and thus should be the site for government-owned heavy industry (e.g., São Paulo or, Beijing-Tianjin historically). Later we will point out that it may be true that output per worker in heavy industries is higher in the productive external environment of large metro areas. It just isn't high enough to cover the higher opportunity costs of land and labor in those cities, which is one reason why those state-owned heavy lose money in such cities. Additionally, there is the environmental issue of putting heavy industry in the midst of the largest number of potential pollution victims (Tolley, Gardiner and Graves (1979)).

Favoritism of a primate city creates a non-level playing field in competition across cities. The favored city draws in migrants and firms from hinterland areas, creating an extremely congested high cost-of-living metro area. If such cities are of excessive size, in theory that affects national productivity, draining resources away from productive and innovative activity into shoring up the quality of life in cities like Bangkok, Jakarta, Karachi or Mexico City. Policy makers can try to resist the migration response to primate city favoritism. Former planned economies, most notably China, institutionally can and do limit migration. In most countries while explicit migration restrictions are not possible, primate cities can refuse to provide legal housing development for immigrants and to provide basic public services in immigrant neighborhoods. Hence the development of squatter settlements, bustees, kampongs and so on. But still, favored cities tend to draw in enormous populations.

Is there econometric evidence indicating that these forces seem to be important and the stories relevant? The most recent studies examine the political economy of the issue. Favoritism of a primate city is first documented. Ades and Glaeser (1995) based on cross-section analyses find that if the primate city in a country is the national capital it is 45% larger. If the country is a dictatorship, or at the extreme of non-democracy, the primate city is 40-45% larger. The idea is that representative democracy gives a political voice to the hinterland regions limiting the ability of the capital city to favor itself. Apart from representative democracy, fiscal decentralization helps to level the playing field across cities, by giving political autonomy for hinterland cities to compete with the primate city.

Davis and Henderson (2001) explore these ideas further, examining in a panel context the impact upon primacy of democratization and fiscal decentralization from 1960-1995. Using a panel approach with IV estimation, they find smaller effects than Ades and Glaeser but still highly significant ones. Examining both democratization and fiscal decentralization together they find moving from the extreme of least to most democratic form of

government reduces primacy by 8% and from the extreme of most to least centralized government reduces primacy by 5%. Primate cities which are national capitals are 20% larger and primate cities in planned economies with migration restrictions are 18% smaller. Finally transport infrastructure investment in hinterlands which opens up international markets to hinterland cities reduces primacy. A one-standard deviation increase in roads per sq. kilometer of national land area or in navigable inland waterways per sq. kilometer, ceteris paribus, each reduce primacy by 10%.

The second set of issues concerning the degree of spatial concentration is the "so-what" question. The first examination of this is Henderson (2002a), which asks whether, for any level of development, there is an optimal degree of urban concentration as measured by primacy, and, if so, whether significant deviations detract from productivity growth. The idea is that optimal primacy for any level of development derives from a trade-off from increasing primacy of enhancing scale economies contributing to productivity growth versus accentuating the extent of resources diverted to shoring up the quality of life in primate cities. Using panel data and IV estimation for 1960-1990, the paper finds that there is an optimal degree of primacy at each level of development that declines as development proceeds. That is, initial relative agglomeration is most important at low levels of development when countries have low knowledge accumulation, are importing technology, and have limited capital to invest in widespread hinterland development. Error bands about optimal primacy numbers are quite tight. Second, large deviations from optimal primacy strongly affect productivity growth. An 33% increase or decrease in primacy from a typical best level of .3 reduces productivity growth by 3% over five years. There is a modest tendency internationally to excessive primacy, with the usual suspects such as Argentina, Chile, Peru, Thailand, Mexico, and Algeria having extremely high primacy.

### 2.2. Core-Periphery Models

Are there models which explain the development of a core-periphery structure across regions of a country? Can these models be used to also explain reversal of a core-periphery structure? The answer is a limited yes. The models are mostly static and the driving force is exogenous technological change. But they address interesting issues.

With Krugman's (1991) paper on the "new" economic geography, a new brand of two-region models appeared. Krugman's paper and the multitude of papers which followed distinctly differ from the dual-economy literature. First there are explicit scale economy forces that foster endogenous agglom-

eration. Second while there are two regions, no starting point is imposed where one region is assumed to start off ahead of the other. Urbanization, or more specifically industrialization, may occur in both regions or in only one region. One region can become "backward" (under certain assumptions), or, if not backward (lower real incomes) at least relatively depopulated. But these are outcomes solved for in the model. Third the models have some notion of space represented as transport costs of goods between regions. Finally the models are focused on a key developmental issue – the initial development of a core (say, coastal) region and a periphery (say, hinterland) region as technology improves (transport costs fall) from a situation starting with two identical regions. Some papers (Puga (1999), Helpman (1998), and Tabuchi (1998)) also analyze how under certain conditions, with further technological improvements, there can be reversal. Some industrial resources leave the core; and the periphery also industrializes/urbanizes, either partially or to the same extent as the core.

The drawback of the models, as regional models, is they are almost exclusively unidimensional in focus: what happens to core-periphery development as transport costs between regions decline. They are not focused on other forms of technological advance, let alone endogenous technological development. With two exceptions, Fujita and Thisse (2002) and Baldwin (2001), the models are static. But even in these exceptions, the focus is on the effect of exogenous changes in transport technology on the regional allocation of population, within an endogenous growth context. Compared to the older dual economy literature there are generally no typical policy considerations of interest to development economists, such as the impact of wage subsidies, rural-urban terms of trade, or capital market imperfections. An exception is that some papers have examined the impact on core-periphery structures of reducing barriers to international trade, such as tariff reduction.

However the examination of core-periphery development or of core region urbanization/industrialization makes the new economic geography literature of interest in any review of urbanization and development. An excellent summary of the key elements is in Neary (2001) and Fujita and Thisse (2000) have an excellent review of the now enormous literature on new economic geography. Fujita, Krugman and Venables (1999) stands as the basic reference on detailed modeling. My examination here is limited to the regional version of the model (as opposed to the two country version), where labor migration across regions occurs, as in Krugman (1991).

In Krugman (1991) there are two regions, each with an identical number of farmers who are completely immobile and who each produce a fixed amount of farm output. Only manufacturing and the fixed population of manufacturing workers are mobile across regions. National scale economies arise from Dixit-Stiglitz (1977) diversity in manufacture's output, given firm level scale economies (fixed costs). Relative to the traditional location literature, Krugman's key insight is that when manufacturing firms choose a location, they employ workers who reside and consume at the location, creating local backward and forward linkages. The more workers in a region, the more varieties result, and real incomes rise, and more workers are attracted to the region – a "virtuous" circle. Rather than presenting the Krugman model per se, I outline the structure of Puga's (1999) variant, since it has a key element of interest – possible reversal of the core-periphery structure – and its assumptions are perhaps more palatable. Puga allows for inter-sectoral (farming-manufacturing) as well as inter-regional labor mobility; and, building on Venables (1996), he also allows for national scale economies in the production process, as well as in consumption.

Here I present the primitives and key relationships of a core-periphery model, discuss the key analytical tool, discuss the key insights about agglomeration versus dispersal forces, and present basic core-periphery results. I do not do full derivations given the limited space and the fact that a number of good reviews and summaries already exist. The idea is to articulate the forces at work.

There are two regions, each endowed with an equal amount of land, $K_i$ for region $i$. Agricultural output in region $i$, $y_i$ is produced with land and labor in agriculture in region $i$, $L_{Ai}$ so

$$y_i = K_i^{1-\theta} L_{Ai}^{\theta} \tag{1}$$

The agricultural sector is perfectly competitive, its output is transported costlessly across regions (a very weird assumption made throughout this literature), and consequently the numeraire is usually the price of agricultural products.

Preferences exhibit Dixit-Stiglitz (1977) returns from varieties of manufacture's $x$, so for any individual

$$U = y^{1-\gamma} x^{\gamma} \tag{2a}$$

$$x = \left[\sum_{k=1}^{M} (x(k))^{(\sigma-1)/\sigma}\right]^{\sigma/(\sigma-1)} \tag{2b}$$

where $M$ is the number of manufacturing varieties nationally (given a closed economy). The elasticity of substitution $\sigma > 1$. As $\sigma$ falls to 1 having varieties is increasingly important since they are less substitutable in consumption. Under this formulation, at equal resource cost (which is not the case with scale economies to firms), consumers would always prefer another variety to more of a given variety. National returns to scale in population arise since a larger economy, as we will see, can support a greater number of varieties.

Given (2), the indirect utility of a worker in region $i$ may be written as

$$V_i = q_i^{\gamma} w_i \tag{3}$$

where $w_i$ is the wage in region $i$ and $q_i$ is a price index for the composite of manufactures for a person in region $i$, imposing symmetry (which is an endogenous outcome) in manufacture's output and pricing decisions. Given the functional form in (2b) for the composite good, the corresponding price index, $q_i$, from standard Dixit-Stiglitz results has the form

$$q_i = \left[\sum_{k=1}^{N_i} (p_i(k))^{1-\sigma} + \sum_{d=1}^{N_j} (p_j(d)\tau)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}$$

where with symmetry

$$q_i = (p_i^{1-\sigma} N_i + (\tau p_j)^{1-\sigma} N_j)^{\frac{1}{1-\sigma}} \tag{4}$$

$N_i$ and $N_j$ are the number of varieties produced in regions $i$ and $j$. $p_i$ and $p_j$ are the local prices of a variety in respectively regions $i$ and $j$. But items shipped from $j$ to $i$ are subject to transport costs; $\tau > 1$ is the number of units of a good needed to be shipped from $j$ in order for a one unit to arrive in $i$. With this form of iceberg transport costs, producers in one region would never choose to duplicate varieties offered in other regions. Note given $\sigma > 1$, $q_i$ is increasing in $p$ and decreasing in varieties $N$.

Manufacturers have identical technologies for each variety and by assumption each produce only one variety sold under monopolistic competition. Manufacturers employ labor and the composite of all manufactured products. For simplicity that composite has the same form as (2b), with a

production relationship $Ax^u l^u = \alpha + \beta x_k$, where $x$ is a composite, $l$ labor, and $x_k$ the output of variety $k$. We get the following total cost function (with appropriate normalization of $A$) for a firm in region $i$

$$TC_i = q_i^u w_i^{1-u}(\alpha + \beta x_i) \tag{5}$$

where $x_i$ is output of any single firm in region $i$. Note the fixed cost $\alpha$ plays a critical role. Firm scale economies limit the efficient and/or equilibrium number of firms in an economy. Increasing the population of the economy allows it to pay more $\alpha$'s and have more firms and varieties. That, per se, increases per resident welfare as an economy's size grows.

The rest is standard. Firms mark-up over marginal cost $\beta q_i^u w_i^{1-u}$ by $\sigma/(\sigma-1)$ under monopolistic competition and under zero profits with free entry produce $x = \alpha(\sigma-1)/\beta$. Demand for output of any firm in region $i$ producing variety $k$ is

$$x(k) = p_i(k)^{-\sigma}[e_i q_i^{(\sigma-1)} + e_j q_j^{(\sigma-1)} \tau^{(1-\sigma)})] \tag{6}$$

$e_i$ and $e_j$ are the demand bases for any variety (total expenditures on manufactures) in regions $i$ and $j$. This base, $e$, is the share of $x$ in consumption, $\gamma$, times all local income (wage, land rents, and profits if any) plus the producer share parameter for manufactures, $u$, times total costs of all local manufactures. Market clearing conditions are threefold. First demand equals supply. Second workers move to equalize utility across regions and third firms relocate until profits are equal (to zero) in both regions. With free inter-regional firm and labor mobility, in the literature, either workers always have equal utility across regions ($q_i^{-\gamma} w_i = q_j^{-\gamma} w_j$) with instant migration and firms move across regions and change in number according to differential profits; or firms adjust instantly so profits remain zero everywhere and workers adjust through inter-regional migration to inter-regional utility differences.

The key point in these models is always the following. If regions are of equal size, then a symmetric outcome with identical regions will always solve the first order and market clearing conditions. However is such a candidate for an equilibrium actually an equilibrium? In particular is it stable (or in other contexts is it a Nash equilibrium in location choice)? That is, if a new firm is added to region $j$ (or moves from $i$ to $j$) will firm profits in $j$ then exceed those in $i$, inducing further agglomeration into $j$, with the typical final outcome being complete agglomeration of manufacturing in region $j$? What are the forces at work?

First is a force promoting stability of a symmetric outcome. An extra firm in $j$ lowers the price index in that region, which lowers (eq. (6)) the demand facing each firm for any variety. That is a competition effect. There are two forces promoting instability of a symmetric outcome, or promoting a core-periphery structure. First are demand or backward linkages, increasing profitability for existing firms. The new firm in hiring in the labor and manufacturing input markets increases demand for labor directly and indirectly, which induces in-migration. Thus the demand for any local variety in the home market is increased (relative to the other region) due to more labor income and increased demand for inputs. Second are cost and forward linkages. An extra firm lowers $q_j$ by providing more varieties, which in turn lowers input costs for firms. With $q_j$ declining, real wages rise inducing in-migration to equalize real wages, which causes nominal wages to decline, lowering production costs. The question is what is the net effect of these three forces.

In general, for any values of $\sigma, \gamma$, and $u$, there are three regions of parameter space corresponding to different values of transport costs $\tau$. The size of these regions vary as $\sigma, \gamma$, and $u$ vary; but the key experiment is to always vary $\tau$. In the first region of parameter space with relatively high values of $\tau$, only a symmetric equilibrium is stable. Farming satisfies the Inada conditions so there is farming in both regions for any parameter values in equilibrium. With high $\tau$, if we start from a symmetric equilibrium, if a firm moves to $j$, the competition forces dominate and profits decline. Since competition is mostly in the local own market given protection offered by transport costs, local competition effects are enhanced and firms can't gain by moving from $i$ to $j$, so symmetry is maintained. With high $\tau$, if we start from a core-periphery structure, if a firm moves from the core to the periphery, it increases its profits, given its market is protected by transport costs (i.e., demand effects dominate). That induces more firms to move, causing the core-periphery structure to fall apart and a symmetric outcome to occur. Of course intuitively the point is simple. With high transport costs, manufacturers locate in both regions to sell to farmers.

At the other extreme with very low transport costs, we can have only a core-periphery structure. A symmetric outcome is unstable, because with low transport costs, backward and forward linkages dominate (even though they weaken as transport costs fall) to ensure (1) agglomeration of all manufacturing in the core and (2) relative agglomeration of farming in the core. The fall in transport cost so weakens the protection of local markets, that local production disappears in one region.

At intermediate values of $\tau$, there are multiple equilibria. Symmetric equilibria are stable, as are core-periphery structures.

The development twist is to view changes in $\tau$ as technological progress. As technology improves so $\tau$ falls, a country moves from a symmetric outcome to a core-periphery outcome. To this Krugman-type result, Puga adds an interesting twist, which raises the potential for reversal of a core-periphery structure. If, instead of being mobile, labor is immobile across regions as $\tau$ declines, we still progress from a symmetric to core-periphery outcome. However now when $\tau$ gets very low, firms will leave the core to move back to the periphery with its low wage costs (given a "surplus" of labor in the periphery). Once trade become minimally expensive, linkage effects no longer work to ensure the core's dominance. This of course is the typical suggested scenario. Core-periphery structures start to reverse themselves once transport costs fall, so firms can utilize cheap hinterland labor. In Puga, the reversal can be partial with more manufacturing in the core than the periphery or it can be complete with again a symmetric outcome. Puga gets this result under a special case with forced labor immobility across regions. However one could then conceive of a situation with limited labor mobility where as technology improves ($\tau$ falls) we move through the various regions of parameter space with some agglomeration of labor in the core. However the importance of the core for manufacturing could first become almost exclusive, followed by decentralization in the latter stages, where industry moves back to employ the remaining cheap labor in the hinterland.

There are easier modeling ways to get the core-periphery reversal, as noted by Helpman (1998), Junius (1999), and Tabuchi (1998), while maintaining (some) labor mobility. Tabuchi (1998), for example, follows the Krugman structure of immobile agricultural workers but perfectly interregionally mobile manufacturing workers. The key element of reversal is congestion, represented in Tabuchi and Helpman as rising housing costs, either because there is commuting or fixed land for housing. In this context we have the same stages where as transport costs fall from very high levels a core-periphery structure develops.[1] But then when transport costs are very low, linkage effects in the Dixit-Stiglitz model become unimportant. From a core-periphery structure, firms move to the periphery where wage costs are low because housing costs are low, resulting in industrial dispersion.

[1]Tabuchi (1998) as well as others (Fujita and Thisse (2002)) note that even with high transport costs we can have a core-periphery structure if manufacturing's share in consumption is very high.

*2.2.1. Extensions of the Core-Periphery Model*

Two extensions of the core-periphery model are of interest to the development-growth literature. First is the reformulation of the model in a growth context, by Baldwin and Forslid (2000), Baldwin (2001), and Fujita and Thisse (2002). We start with the Fujita-Thisse version, where there are two regions, three economic sectors and two types of workers. Immobile unskilled workers are employed in the traditional and modern sectors; mobile (at a cost) skilled workers are employed in the innovative sector; and there is an international capital market where the two regions face an exogenously given cost of capital. The core-periphery structure depends on the migration decisions of skilled workers and agglomeration in the innovative sector. Overall scale economies are in the modern sector, where the number of consumer varieties equals the number of (infinite length) patents in the innovative sector. Consumers have infinite horizons in a continuous time model.

The productivity of skilled workers equals the knowledge capital in each region; and the number of patents developed in a region each instant is proportional to the knowledge capital in that region. Knowledge capital in a region is the sum of all human capital of skilled workers in that region plus knowledge spillovers proportional to the human capital of skilled workers in the other region. Finally human capital of any worker is proportional to the number of patents in the country (not region). If knowledge spillovers across regions are limited, then that becomes a powerful force for agglomeration, since knowledge capital nationally, which determines the level of patent development and hence the rate of human capital, increases with agglomeration.

The authors consider several situations, which differ by whether modern firms (and the patent they each hold) are mobile across regions (equalizing profits in the modern sector). Assuming firm mobility, again the issue is whether a core-periphery structure emerges for different exogenous values of the cost of transporting modern goods across regions. The results do differ from the static model since limited knowledge spillovers across regions mean the innovating sector is always agglomerated, if firms (and hence issued patents) are perfectly mobile. Thus for high transport costs, the innovative sector is in the core and the variety demands by those skilled workers in the core draw in a disproportionate share of modern sector firms. But for high transport costs some modern sector activity exists in the periphery to serve the demands of unskilled workers there. As transport costs fall at some point, the modern sector agglomerates entirely in the core, given the

transport cost of serving the periphery from the core is low enough.[2] This analysis of the role of transport costs is not really different than in the static core-periphery models, especially given the "technological change" – drop in transport costs – is exogenous. However there is an aspect of growth of interest – evolving spatial inequality.

Because unskilled workers are immobile, a core-periphery structure generates inequality between core-periphery workers. However Fujita and Thisse show that if overall growth is fast enough, periphery workers will be absolutely better off under a core-periphery structure than an (unstable) symmetrical structure. Agglomeration in the innovative sector spurs development of varieties nationally and if the rate of variety expansion is sufficient, periphery workers are absolutely better off.

Baldwin and Forslid (2000) have a simpler model – non-forward looking workers and two rather than three explicit sectors. But they focus less on the role of transport costs and more on growth. In their framework inter-regional knowledge spillovers are a force encouraging a non-core-periphery structure in the sense that as knowledge flows more freely that reduces the costs of a symmetric outcome and permits stable symmetric outcomes over a larger range of (relatively high) transport costs of trade.

The core-periphery model has also been used to analyze the impact on peripheral, or hinterland regions of "globalization", or reduced barriers to international trade (Krugman and Venables (1995), Krugman and Livas (1996), Puga and Venables (1999)). This literature argues that globalization helps peripheral regions (at least in certain regions of parameter space) either because it redefines the focal points in the economy away from the traditional core to border regions or because it opens up markets for hinterland producers. While peripheral producers may be relatively non-competitive in domestic markets, once international markets open to the whole country, the relative competitive advantage of the core over the periphery in distant international market may be quite modest.

### *2.2.2. Urbanization and the Core-Periphery Model*

As a regional model, the core-periphery model suffers often cited limitations. The location-resource bound good – agriculture – has no transport costs; surely the cost of transporting agricultural products from fertile hinterland regions (e.g., U.S. mid-West) is a force for dispersion, as well documented historically (see section on geography below). The assumption

[2]If firms and patents are not perfectly mobile, stable symmetric equilibria exist for high values of transport costs.

of iceberg transport costs are also a force against dispersion. With linear transport costs, at some distance trade ceases and peripheral regions need to produce their own varieties (potentially duplicating coastal varieties). But these are details, reflecting choices on essential modeling ingredients.

More critically the core-periphery model is sufficiently complex that to date almost no welfare and policy analysis has been carried out with it. Part of the reason is that we know little about the welfare properties of the equilibria that can result. Policy analysis would be in an $n$-best context and one in which the role of government has not been modeled. As noted earlier, the first generation dual economy models were completely policy focused, examining the impact of input and output market distortions. Since much of development economics focuses on policy issues, this is a severe limitation. Second, to date with the exception of work by Hanson (1996, 2000) and Holmes and Stevens (2002), little empirical work has been done to test the core-periphery model and its key aspects.

However, the key issue in terms of urbanization is that the core-periphery model is more a regional model, with limited urban implications. What are the key distinctions? Urban models are focused on the city formation process, where economies are composed of numerous cities, in which both the number and sizes of cities are endogenous. An important issue is the extent of market completeness in the national land market in which cities form and the role of land developers, city governments and inter-city competition in that formation process. A second key distinction is that there are distinct city "types", where within a region there is a wide size distribution of cities. Each city type is relatively specialized in a particular product or range of products, so one research question is the inter-relationship between, say, large more diverse metro areas and smaller more specialized metro areas. A third distinction as we will see involves a focus on welfare, policy, and institutional issues.

Finally the details differ. Urban models utilize Marshall's scale externalities such as localized information spillovers, as well as local knowledge accumulation as the basis of agglomeration, rather than market linkages. As we will see that becomes a basis to link urban and national economic growth. Urban models also account for the internal structure of cities where commuting and congestion and other negative externalities associated with crowding are a force for dispersion. Finally while urban models can incorporate an agricultural sector, they de-emphasize the role of agriculture given in developed countries such as the USA only 2-3% of the local force is actively engaged in agriculture. The focus is on footloose production.

## 3. SCALE ECONOMIES

In this section, we examine the forces for agglomeration that are intrinsic to urban model. We examine models of the micro-foundations of scale economies in the urban literature and review the empirical evidence on the subject. Understanding the nature of scale externalities which in a modern economy are viewed as the key spatial agglomerating force is important to understanding the inter-relations across cities and the production structure of cities. Since these scale externalities are also the basis of endogenous growth theory, it is useful to see how they play at the sub-national level in cities, where close spatial proximity makes the idea of spillovers most relevant.

### 3.1. Scale Externalities: Microfoundations

In the original urban systems model (Henderson (1974)), the basis for agglomeration is localized own industry scale externalities, usually modeled as being Hicks' neutral. In a typical specification, following Chipman (1970) firms are competitive, constant returns to scale producers where output of firm $i$ in city $j$ is

$$x_{ij} = A(N_j)x(k_{ij}, n_{ij}) \tag{7}$$

$x(\cdot)$ is CRS with firm inputs of capital ($k_{ij}$) and labor ($n_{ij}$). The $A(\cdot)$ function is a Hicks' neutral shifter factor where $A^{'} \succeq 0$ and $N_j$ can be total employment in the own industry in city $j$, or total employment overall in city $j$. Also the scale measure, rather than being local employment, can be local output or local number of firms. The relevant arguments in $A(\cdot)$ are the subject of a large body of empirical work, discussed later.

Starting in 1982, urban economists worked on the micro-foundations to the block-box process in eq. (7), examining Marshall's (1890) hypothesized urban externalities such as (in modern words) information spillovers, search and matching externalities in labor markets, and intra-industry plant specialization. In path-breaking piece Fujita and Ogawa (1982) modelled firms along a line as being subject to exogenous information spillovers from other firms where information decays with distance. If the line runs from $b_1$ to $b_2$ and firms are uniformly distributed on the line and information decays expotentially with distance, the $A\,(\cdot)$ function in (7) for a firm at $y$ has the form

$$A(y) = \int_{b_1}^{b_2} e^{-\alpha|y-s|}ds = \frac{1}{2}[2 - e^{-\alpha(y-b_1)} - e^{-\alpha(b_2-y)}] \tag{8}$$

$\alpha$ is the rate of spatial decay of information and each firm's contribution to the $A(y)$ function of the firm at $y$ is exogenous and not dependent on the size of the operations of other firms. If firm output $x(y)$ is simply $A(y)$ (times one unit of labor) then total output of all firms over the $b_1$, $b_2$ interval can be shown to be (integrating in (8) over $y$)

$$X = 2/\alpha \left[N - \alpha^{-1}[1 - e^{-\alpha N}]\right] \tag{9}$$

where $N \equiv b_2$ - $b_1$ is the measure of city employment. Note $dX/dN$, $d^2X/dN^2 > 0$, so the marginal product of labor is increasing in city employment.

There are two interesting extensions to this model. First Kim (1988) endogenizes the spillovers in eq. (7), where firms choose the amount of information they receive given the cost of information acquisition rises with distance. Second Lucas and Rossi-Hansberg (2001) and Rossi-Hansberg (2001) redo Fujita-Ogawa in a circular city where the density of firms in the central business district is endogenous. As Rossi-Hansberg shows this raises the issue of equilibrium versus optimal land use patterns and optimal spatial structure. In equilibrium configurations, firms don't recognize the impact of increasing land consumption decisions on reducing the proximity of firms to each other, thus contributing to excessive decay of spillovers. Optimal land use configurations tend to be of higher overall density, or in a more compact business district, with less overall spatial decay of spillovers.

For labor market search and matching models, Helsley and Strange (1990) assume workers in a city are heterogenous in (unranked) skills, and are drawn from a uniform distribution of skills over the unit circle. Firms must commit to a technology which is an address, $s$, on this unit circle before knowing the actual drawing (addresses) of workers. The value ex post value of a match is

$$\max[0, \alpha - \beta|s - y|] \tag{10}$$

between a firm at $s$ and a worker at $y$. If there are $m$ firms in the city a Nash equilibrium in location choices is for firms to uniformly distribute so the expected distance between any firm and workers is $(4m)^{-1}$. For $N$ workers in the city, firm profits are the expected number of employees $(N/m)$ multiplied by expected output per worker $((\alpha - \beta(4m)^{-1})$, assuming $\alpha > \frac{1}{2}\beta$, so $\alpha - \beta|s - y| > 0$ for all possible address combinations), all minus a fixed cost per firm of $C$. If, for example, for any city labor force the number of firms, $m$, is chosen to maximize total expected output in the

city, or $N(\alpha - \beta(4m)^{-1}) - Cm$, then we can show total city output is

$$X = \alpha N - \beta^{\frac{1}{2}} C^{\frac{1}{2}} N^{\frac{1}{2}} \tag{11}$$

where again $dX/dN$, $d^2X/dN^2 > 0$ or the marginal product of labor increases with city sizes.

Other models include ones based on Dixit-Stiglitz diversity of intermediate inputs which are non-traded across cities or relatedly on local intra-industry specialization. These are extreme versions of linkages where each city must produce its own varieties. So, for example, following Abdel-Rahman and Fujita (1990), suppose $y$ is firm output of the city's export good (say, computers) produced by CRS competitive firms with labor $n_y$ and varieties of intermediate inputs $x$, according to

$$y = n_y^{\alpha} (\sum_{i=1}^{n} x_i^{\rho})^{(1-\alpha)/\rho} \tag{12}$$

Then under the usual cost function for any variety $N_x^i = f + cX_i$ and a full employment constraint, if for simplicity $m$ and $X_i$ are chosen optimally, we can show[3] that total city output is

$$Y = C_0 N^{\frac{1-\alpha+\alpha\rho}{\rho}} \tag{13}$$

where again $dY/\alpha N, d^2Y/dN^2 > 0$, for $N$ total city employment and $C_0$ a constant. Similar to this model Becker and Henderson (2000) adapt the Becker and Murphy (1992) model of Adam Smith specialization where firms specialize in sets of contiguous heterogenous tasks needed for industry output to result. Again the marginal product of labor is increasing in urban scale.

Note that all these micro-foundation models have a reduced "black-box" form with rising marginal product of labor to the city (but not firm). Scale improves productivity, but the reasons could be quite different, as the different models indicate. These are models of "static externalities" – information spillovers today increasing local industry efficiency today. There are also specifications in dynamic contexts, but these are also black-box ones. The shift factor $A(\cdot)$ in equation (7) can be made to depend on the local stock of knowledge (say, local human capital) or the level of local

[3]Given full employment, symmetry and aggregation in the CRS $Y$ sector, $N = N_y + mN_x$. Given the cost function for $X$, then $N_y = N - m(f + cX)$. Substituting for $N_y$ into $Y = N_y^{\alpha}(mX^{\rho})^{(1-\alpha)/\rho}$ and optimizing with respect to $X$ and $m$ and then substituting back into the $Y$ function yields (13).

industry activity in the past $(N_j/(t-1))$ contributing to a stock of local trade secrets, or growth in $A(\cdot)$ can be made to depend on the local stock of knowledge. We will examine such formulations in both the review of empirical evidence and the presentation of the endogenous growth model.

### 3.2. Scale Externalities: Evidence

The tradition issue in evaluating scale externalities concerns the relevant arguments in the $A(\cdot)$ function in eq. (7) for particular industries. We consider both "static" externalities and the more recent literature on "dynamic" externalities.

#### *3.2.1. Static Externalities*

For some industries such as standardized manufacturing, the literature starting with Hoover (1948) argues that scale economies are ones of localization, meaning they are strictly internal to the own industry and dependent on scale of the own industry locally. Jacobs (1969) on the other hand argues that, for some industries where innovation and marketing are important, what is relevant is the overall scale and diversity of the local environment. In static form such economies are ones of urbanization, where scale externalities depend on the overall size or potential diversity of the local environment.

Early empirical work (e.g., Sveikauskas (1975, 1978), Nakamura (1985) and Henderson (1986, 1988)) examined the effect on productivity at the 2-3-digit (SIC) industry level of various scale measures estimating either a primal or dual (cost) form to eq. (7). Work was cross-sectional and industry-specific data were aggregated to the metropolitan area level, so the unit of observation was the city-industry. Despite different approaches and data sets (USA, Brazil and Japan), these three sets of studies concluded there are significant degrees of localization economies in most manufacturing industries such as primary metals, machinery, apparel, textiles, pulp and paper, food processing, electrical machinery, and transport equipment, and little evidence of urbanization economies. Below I will argue that scale economies being ones of localization, or internal to the own industry, helps promote urban specialization. Only in industries such as high fashion apparel or glossy publishing did strong evidence of urbanization economies emerge. However since these studies focus on productivity of manufacturing plants, they leave open the question that urbanization economies apply to situations envisioned by Jacobs (1969), such as R&D and perhaps the service sector. Locational evidence in Fujita and Ishii (1994) in electronics

suggests that R&D activities are drawn to large, diverse metro areas, while standardized production is decentralized to smaller cities.

These early productivity studies, even in their own terms face three major issues. First are location "fixed effects" or relatively time invariant unmeasured aspects of the local environment that affect both productivity and right-hand side variables such as industry scale or the capital to labor ratio, resulting in OLS estimates being biased. Such omitted variables include local human capital variables, infrastructure measures, and the local regulatory environment. Second, there is a selection problem. Firms and plants are heterogenous. Perhaps high (or low) productivity plants are disproportionately drawn to locations where there are relatively large clusters of own industry firms. Finally firms may be subject to a contemporaneous locational shock affecting both productivity and inputs, including local industry scale.

The early literature attempted to deal with the first and third problems through IV estimation, but as always with aggregate cross-section data there is the issue of valid instruments – ones not affecting productivity but still (in some conceptual framework) influencing right-hand side covariates. More recent work using panel city-industry data on Korea (Henderson, Lee, Lee (2001)) attempts to deal with the first problem by use of city fixed effects; and has the same findings as the older literature – scale externalities in manufacturing production are ones of localization.

Recent work on this issue has two innovations. First is the use of plant level data. Second is investigating what types of plants benefit from scale externalities. Third is a start on investigating the nature of spatial decay of externalities. Henderson (2002b) uses plant level productivity data in a panel context to difference out both city fixed effects and a plant unobserved heterogeneity term (that operates as a Hicks' neutral shift factor) to try to deal with both selectivity and fixed effects. He finds that high tech industries benefit more from localization externalities than traditional machinery industries. Plants in single plant firms benefit more than plants of multi-plant firms, who have a corporate information network to rely on. Finally externalities appear to derive from the number of own industry plants locally representing, say, the count of sources of local information spillovers, rather than total local employment in the own industry. This last item could indicate that information spillovers are the underlying force for externalities, rather than, say, labor market and search externalities. But none of this empirical literature delves into the micro-foundations of scale externalities to effectively distinguish information spillover, labor market externalities, intra-industry specialization, and the like.

On the extent of spatial externalities, Ciccone and Hall (1996) using aggregate cross-section data argue that density of local activity is important. Henderson (2001) argues that scale effects are internal to the own county and don't result from activity in nearby counties. But the most direct work is that of Rosenthal and Strange (2002) who look at how localization scale effects decay with distance, although their work is based on indirect productivity inferences from birth patterns. They have data by zip codes on births and argue that relative to adding plants within a 1 mile radius, adding plants in a 1-5 mile radius improves productivity by only 7-50% as much depending on the industry, with effects generally dying out at ten miles. Small plants benefit more than big plants from these localization effects.

These studies still face problems with controlling for contemporaneous location shocks which influence both productivity and hence scale. Henderson (2002b) and Rosenthal and Strange (2002) try controlling for time-metro area fixed effects (i.e., contemporaneous metro area shocks) while investigating scale effects at a more detailed geographic level (county or zip code). However that leaves open doors – what about zip code or county shocks. A potential solution to find good instruments for local scale measures involves looking at a location decision framework to model agglomeration (Arthur (1990)). There potential instruments for local country scale, would be (exogenous to own county) attributes of competitor counties. Improved attributes in those counties draw plants away from the own county without directly affecting own county productivity (Bayer and Timmins (2001)).

### *3.2.2. Dynamic Externalities*

There appear to be two sets of working definitions of dynamic externalities. First is that either the history of economic activity in a location affects productivity levels today or base period variables affect productivity growth. The second set concerns the effect of "knowledge" (rather than information) spillovers on productivity levels. Knowledge is typically measured by average education and the issue is whether average education in a city affects productivity. It isn't clear this is a dynamic effect per se. It could be static in the sense that average education could simply enhance static productivity levels (but not on-going growth rates of productivity), but as we will see later that is sufficient to enhance overall urban scale and promote endogenous growth.

For the knowledge accumulation framework, Rauch (1993a) estimates that average education in a city enhances individual wages, although he has no control for location effects, sorting effects, or contemporaneous shocks affecting both wages and education. Moretti (1999) in an important piece merges plant level productivity data for 1982 and 1992 with individual education data from the Population Census (PUMS) for 1980 and 1990 to test whether average educational attainment outside the own industry affects plant level productivity, controlling for own industry education. Controlling for overall location fixed effects, he finds that a 1-year increase in average education in the city outside the own industry increases plant productivity by 5%. He also finds that the effect for multi-plant firms is zero, while for single plant firms it is 7.7%. This is very suggestive work. Clearly it would be interesting to combine an analysis of knowledge accumulation with scale externalities.

For the productivity growth framework, there is a growing empirical literature on city-industry growth, dating to Glaeser, Khalil, Scheinkman, and Shleifer (1992), with a variant in Henderson, Kuncoro, and Turner (1995). The idea is that base period variables such as local own industry scale or diversity of the local industrial environment encourage local industry growth, by promoting local productivity growth which attracts more firms to a city. Glaeser et al. (1992) find evidence of "dynamic" diversity effects which they call Jacobs economies. Henderson et al. (1995) find these for high tech industries but find only "dynamic" locationalization economies, called MAR (Marshall-Arrow-Romer) externalities in traditional capital goods industries. There is a lot of controversy about what these estimations really say especially since issues of endogeneity (to location fixed effects) are typically overlooked. Looking at net growth of employment of plants combines two processes, plant births and plant deaths. Davis, Haltiwanger and Schuh (1996) present convincingly evidence that deaths tend to be related to plant and firm idiosyncratic shocks, rather than location attributes. Thus the typical location literature analyzes patterns of births to make inferences about profit functions and scale effects(Carleton (1983)). Another issue is that, while diversity affects location choices it may not affect productivity (as Henderson (2002b) finds in looking at lagged diversity or own industry scale effects on productivity). So diversity may affect, say, the price, availability, and quality of intermediate inputs drawing firms into a city without affecting scale externalities. While industries co-locate to "trade" (reduce transport costs of intermediate inputs) so that growth in industry $B$ at a location is correlated with growth of support industries $X$ to $Z$, that doesn't mean the degree of diversity of $X$ to $Z$ affects pro-

ductivity in the sense of affecting the $A(\cdot)$ function in eq. (7). Another key issue concerns how to put all this in a framework of agglomeration over time, with stochastic components Arthur (1989). As we will note below, individual industry agglomerations at a location tend to change quickly over time. We have no developed model of that.

## 4. ECONOMIES COMPOSED OF CITIES

This section examines empirical evidence and models for larger countries or regions with urban systems comprising dozens or even hundreds of cities. There is an emerging set of well documented facts about the size distribution, production patterns, evolution of the sizes and numbers of cities over time, and the role of geography in urban systems. Once we have examined the empirical evidence, we will turn to modeling systems of cities, with a focus on city specialization and trade patterns and the growth in sizes and numbers of cities over time. Finally we will examine very recent theoretical work focused on the role of large metro areas versus smaller ones in an economy and how to integrate traditional urban systems models with key aspects of the new economic geography.

### 4.1. Empirical Facts About Urban Economic Geography

In this section, we examine the evolution of the size distribution of cities in countries, accounting for city size growth and entry of new cities. We examine patterns of specialization in production by cities. Finally we turn to attempts to account for explicit geographic factors on urban development.

#### *4.1.1. The Size Distribution of Cities and Its Evolution*

Work by Eaton and Eckstein (1997) on France and Japan and by Dobkins and Ioannides (2001) on the USA with later work by Black and Henderson (2002) and Ioannides and Overman (2001) on the USA establish some basic facts about the development of urban systems in France, Japan, and the USA over the last century or so. In general, there is a wide size distribution of cities in any large economy, where relative size distributions are remarkably stable over time. In this sub-section we examine facts about the evolution of the size distribution of cities and city growth. In the next we ask why there is a wide size distribution, where relatively big and small cities coexist indefinitely.

The empirical work looks at the decade by decade development of urban systems. In doing so, there are critical choices researchers must make when assembling data. First is to define what is described by the all-purpose term

"city". The usual definition is the "metro area" where from a conceptual point of view one is trying to capture all contiguous urban economic activity around an urban core, or central city. Large metro areas like Chicago comprise over 100 municipalities, or local political units, and are defined to cover the entire metro area labor market and to geographically cover all contiguous manufacturing, service and residential activity radiating out from the Loop (Chicago city center) until activity peters out into farm land or very low density development. Of course many problems arise, such as how to treat two or more neighboring and expanding metro areas that at some point start to overlap. A second problem concerns how to do these definitions over time. One approach is to use whatever contemporaneous definition the country census/statistical bureau uses but one problem with that is that metro area (vs. municipality) concepts only start to be applied after World War II. Another approach is to take current metro area definitions and follow the same geographic areas back in time, focusing on non-agricultural activity.

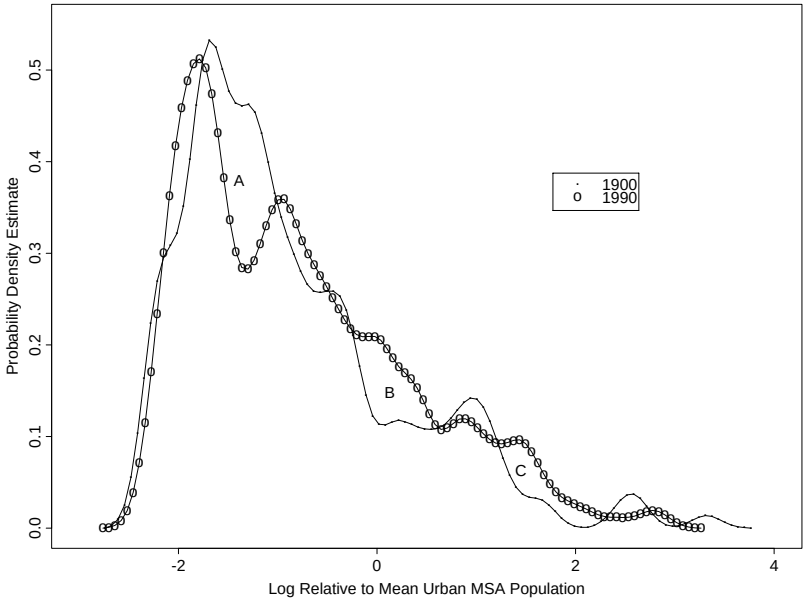
A third problem concerns how to define "consistently" when an agglomeration becomes a city, or metro area over time, especially since the economic nature, population density, and spatial development of metro areas have changed so much over time. Some authors use an absolute cut-off point (e.g., urban population of 50,000 or more); some use a relative cut-off point (e.g., the minimum size city included in the sample should be .15 mean city size); and others look at a set number (e.g., 50 or 100) of the largest cities. For these three issues whatever choices researchers make can strongly affect specific results. Nevertheless a variety of findings emerge that qualitatively are consistent across studies.

In the research, an initial focus was on studying the evolution of the size distribution of cities, applying techniques utilized by Quah (1993) in examining cross-country growth patterns. Cities in each decade are divided by relative size into 5-6 discrete categories, with fixed relative size cut-off points for each cell (e.g., $<$.22 of mean size, .22 to .47 of mean size, ... $>$ 2.2 mean size). A first order Markov process is assumed and a transition matrix calculated. Typically stationarity of the matrix over decades can't be rejected, so cell transition probabilities are based on all transitions over time. If $\mathbf{M}$ is the transition matrix, $i$ the average rate of entry of new cities in each decade (in a context where in practice there is no exit), $\mathbf{Z}$ the (stationary) distribution across cells of entrants (typically concentrated on the lowest cell), and $\mathbf{f}$ the steady-state distribution, then

$$\mathbf{f} = [\mathbf{I} - (1-i)\mathbf{M}]^{-1} i\mathbf{Z} \tag{14}$$

In the data decade relative size distributions are remarkably stable over time and steady-state distributions tend to be close to the most recent distributions. Most critically there is no tendency of distributions to collapse and concentrate in one cell, or for all cities to converge to mean size; nor generally is there a tendency for distributions to become bipolar. Plots of relative size distributions for the U.S.A. in 1900 versus 1990 look almost identical as Figure 5 illustrates; and Lorenz curves for Japan (1925-1985) and France (1876-1990) in Eaton and Eckstein (1997) almost overlap. In Figure 5 the relative size distribution of cities is plotted for 1900 and 1990, where relative size is actual size divided by mean size in the corresponding time period. The density functions for 1900 and 1990 almost coincide.

**FIG. 5.** Density Functions for MSA Size Distributions




Another finding is that, for larger cities, over time there is little change in relative size rankings. In Japan and France, the 39-40 largest cities in 1925 and 1876 respectively all remain in the top 50 in 1985 and 1990 respectively; and, at the top, absolute rankings are unchanged. The USA displays more mobility due to substantial entry of new cities. However, while smaller cities do move up and down in rank, the biggest cities tend to remain big over time. So, for example, cities in the top decile of ranking stay in that decile indefinitely, with newer cities joining that decile as the total number of cities expands. Alternatively viewed based on the Markov

transition process, the mean first passage time for a city to move from the bottom cell to the top cell is typically 1/10 of the mean first passage time to go from the top to the bottom , where in Black and Henderson (2002) the later mean first passage time is 545 decades, beyond any horizon of the data.

Why do big cities stay big? A common answer, in part modeled in Henderson and Ioannides (1981), is physical infrastructure. Large cities have a huge historical capital stock of streets, buildings, sewers, water mains and parks that are cheaply maintained and almost infinitely lived in, that gives them a persistent comparative advantage over cities without that built-up stock. A second answer is modeled in Arthur (1990) and Rauch (1993b) where, with localized scale externalities in production, large cities with an existing fertile externality environment for a particular set of industries have a comparative advantage in attracting new firms over cities with small representation of those industries. We will return to this issue below.

Within these relative size distributions of cities, as urbanization and growth proceed, both the absolute sizes and numbers of cities have tended to grow historically, as a country urbanizes and grows in total population. City sizes in the USA, Japan, and France over the past century have grown at average annual rates of 1.2 - 1.5%, depending on countries and sample choices, rates which involve city sizes rising 3.3 - 4.5 fold every century. A small city today which is 250,000 would have been a major center in 1900. In the USA there has also been a large increase in the number of cities. Over 1900-90, using a relative cut-off point to define city entry (minimum size is .14 of mean size), Black and Henderson (2002) find a 50% increase in the number of cities, while under an absolute cut-off point (50,000) in Dobkins and Ioannides (2001) the number of cities triples.

However we count cities, it is clear they have grown in population on an on-going basis over the last century, even in developed countries. The next section will model this as related to technological change induced by knowledge accumulation. Glaeser, Scheinkman, and Shleifer (1995) in a cross-section city growth framework estimate that controlling for 1960 population, cities in 1990 are 7% larger if they have a one-standard deviation increase in median years of schooling. Black and Henderson (1999) place the issue in a panel context for 1940-1990 controlling for city fixed effects and examining the impact of percent college educated (which has enormous time variation). They find a one-standard deviation increase in percent college educated increases city size by 20%.

**Zipf's Law.** In considering the size distribution of cities, especially in a cross-sectional context, there is an enormous literature on what is termed Zipf's Law (Rosen and Resnick (1980), Clark and Stabler (1991), Mills and Hamilton (1994), Ioannides and Overman (2001)). City sizes are postulated to follow a Pareto distribution, where if $R$ is rank from smallest, $r$, to largest, 1, and $n$ is size

$$R(n) = An^{-a} \tag{15}$$

Under Zipf's Law $a = 1$, or we have the rank size rule where, for every city, rank times size is a constant, $A$. Putting (15) in log-linear form, empirical work produces $a$'s that vary across countries, samples, and times but are "close" to one (ranging, say, from .7 to 1.3) and equations with very high explanatory power. This empirical regularity has drawn considerable attention. While Black and Henderson (2002) show that, with (15) in logs, (1) $a < 1$ and (2) a quadratic in $\ln(n)$ better fits the data for the USA for 1900-90, so that the relationship does not precisely follow a Pareto distribution, the rank size rule may be a good first approximation.

Where would such a relationship come from? Urban economists have not focused on that issue, but in a major development, Gabaix (1999a, 1999b) starts to formalize the underlying stochastic components which might lead to such a relationship, building on Simon (1955). Gabaix shows that if city growth rates obey Gibiat's Law where growth rates are random draws from the same distribution,[4] so growth rates are independent of current size, Zipf's Law emerges as the limiting size distribution. Growth is scale invariant, so the final distribution is and we have a power law with exponent 1. Gabaix sketches an illustrative model. Cities face on-going amenity shocks (bounded away from zero) in an overlapping generations model where only the fraction of people who are young are mobile. The young move to equalize utility which is real income multiplied by the (scale invariant) amenity shock. Real income is subject to local scale (dis)economies which net to zero in large cities. This formulation leads to Zipf's Law for the size distribution of cities.

In a recent draft paper, Duranton (2002) illustrates a similar process in a more developed model. He has "first nature" (immobile given natural resource location) production and "second nature" (mobile, or footloose) production in $m$ cities. There are $n$ ($m >> n$) products, in a Grossman-Helpman (1991) product quality ladder model. Investment in innovation to try to move the next step up in the ladder in industry $k$, can also

[4] Actually the requirement is that they face the same mean and variance in the drawing.

lead to the next step up in a different industry – i.e., there can be cross-industry innovation. To partake of a winning innovation occurring for industry $k$ in city $i$, requires industry $k$ production to locate in city $i$ for footloose industries, which underlies the stochastic process. The result is a approximation (quadratic form to $\ln(R)$ in $\ln(n)$ in (15)) to Zipf's Law.

Duranton's formulation has the advantage over Gabaix's as an urban framework in that cities have patterns of production specialization which change over time (see next sub-section). Second the paper starts to try to more explicitly add urban agglomeration benefits and crowding costs. Both papers pass over issues of city formation and economic growth, as well as issues of stability of static allocation.[5]

While Gibrat's Law is a neat underlying stochastic process, does it hold up empirically? Black and Henderson (2002) test whether in the relationship, $\ln n_{it} - \ln n_{it-1} = a + \delta t + \alpha \ln n_{it-1} + \varepsilon_{it}$, $\alpha = 0$ as hypothesized under the Law. The Law requires $\varepsilon_{it}$ to be i.i.d., so simple OLS suffices. Black and Henderson find $\alpha < 0$ under a variety of circumstance and sub-samples, under appropriate statistical criteria, which rejects Gibrat's Law. Ioannides and Overman (2001) examine the issue in a more non-parametric fashion, characterizing the mean and variance of the distribution from which growth rates are drawn. The mean and variance of growth rates do seem to vary with city size but bootstrapped confidence intervals are fairly wide generally, allowing for the possibility of (almost) equal means.

#### *4.1.2. Geographic Concentration and Urban Specialization*

Geographic concentration refers to the extent to which an industry $k$ is concentrated at a particular location or, more generally concentrated at a few versus many locations nationally. The measure of concentration of industry $k$ at location $i$ might be $l_{ik} = X_{ik} / \sum_i X_{ik}$. $X_{ik}$ is location $i$'s employment or output of industry $k$. Thus $l_{ik}$ is location $i$'s share of, say, national employment in industry $k$. On the other hand specialization refers to how much of a location's total employment is found in industry $k$, or $s_{ik} = X_{ik} / \sum_k X_{ik}$. As Overman, Redding and Venables (2001) demonstrate, if we normalize $l_{ik}$ by location $i$'s share of national employment ($s_{ik} \equiv \sum_k X_{ik} / \sum_k \sum_i X_{ik}$) and $s_{ik}$ by industry $k$'s share of national employment ($s_k \equiv \sum_i X_{ik} / \sum_k \sum_i X_{ik}$) we get the same measure – a location

[5] With scale effects and, say, an inverted $U-$shape to city real income in urban scale, equalized utilities can occur on upward and downward sloping portions of the $U-$shape – only the later are stable, or are Nash equilibria in population location decisions (see below).

quotient, or

$$q_{ik}^{k} = X_{ik} \frac{(\sum_k \sum_i X_{ik})}{\sum_k X_{ik} \sum_i X_{ik}} \tag{16}$$

The distribution of $q_{ik}$ across industries, $k$, compared over time for a city would tell us about how city $i$'s specialization patterns are changing over time. And the distribution of $q_{ik}$ across locations, $i$, over time would tell us whether industry $k$ is becoming more or less concentrated over time. As Overman et al. (2001) point out, in a practical application looking at many industries and cities over time or across countries, the issue concerns how to produce summary measures to describe how overall concentration for one industry compares with another or how one city's degree of specialization compares with another. Another issue concerns how to account in measuring specialization or concentration for different forces that cause these phenomena. The literature uses a variety of approaches.

Evidence on a variety of countries such as Brazil, U.S.A., Korea, and India (Henderson (1988), and Lee (1997)) indicate that cities are relatively specialized. The traditional urban specialization literature going back to Bergsman, Greenston and Healy (1972) uses cluster analysis to group cities into categories based on similarity of production patterns – correlations (or minimum distances) in the shares of different industries in local employment, $S_{ik}$. Cluster analysis is an "art form" in the sense that there is no optimal set of clusters, and it is up to the researcher to define how fine or how broad the clusters should be and there are a variety of clustering algorithms.

Using 1990 data on the U.S.A. Black and Henderson (2002) group 317 metro areas into 55 clusters, "defining" 55 city types based on patterns of specialization for 80 2-digit industries. They define textile, primary metals, machinery, electronics, oil and gas, transport equipment, health services, insurance, entertainment, diversified market center, and so on type cities, where anywhere from 5-33% of local employment is typically found in just one industry. They show that production patterns across the types are statistically different and that average cities and educational levels by type differ significantly across many of the types. Specialization especially among smaller cities tends to be absolute. At a 3-digit level many cities have absolutely zero employment in a variety of categories. So in 1992 for major industries like computers, electronic components, aircraft, instruments, metal working machinery, special machinery, construction machinery, and refrigeration machinery and equipment, respectively, of 317 metro

areas 40%, 17%, 42%, 15%, 77%, 15%, 14% and 24% have absolutely zero employment in these industries.

Kim (1995) in looking at the USA examines how patterns of specialization have changed over time, by comparing for pairs $(i, j)$ of locations $\sum_k |s_{ik} - s_{jk}|$ and by estimating locational Gini's for industry concentration (Krugman (1991b). He finds that states are substantially less specialized in 1987 than in 1860, but that localization, or concentration has increased over time. For Korea, as part of the deconcentration process noted earlier, Henderson, Lee, and Lee (2001) find that from 1983 to 1993, city specialization as measured by a normalized Hirschman-Herfindahl index

$$g_j = \sum_k (s_{jk} - s_j)^2 \tag{17}$$

rises in manufacturing, while a provincial level index declines. Cities become more specialized and provinces less so. Clearly the geographic unit of analysis matters as do the concepts. City specialization as exposited in the models presented below is consistent with regional diversity, when regions are composed of a large number of cities.

Henderson (1997) for the USA and Lee (1997) for Korea show that the $g_j$ index of specialization in manufacturing declines with metro area size. Smaller cities are much more specialized than larger cities in their manufacturing production. More generally, Kolko (1999) demonstrates that larger cities are more service oriented and smaller ones more manufacturing oriented. For six size categories (over 2.5 million, 1 - 2.5 million, ... $<$ .25 million, non-metro counties) he shows that the ratio of manufacturing to business service activity rises from .68 to 2.7 as size declines, where manufacturing and business services account for 35% of local private employment. The other 65% of local employment is in "non-traded" activity whose shares don't vary across cities – consumer services, retail, wholesale, construction, utilities.

What about concentration of industry – the extent to which a particular industry is found in a few versus many locations? In an extremely important paper Ellison and Glaeser (1997) model the problem using USA data, to determine to what extent there is clustering of plants within an industry due to either industry-specific natural advantages (e.g., access to raw materials) or spillovers among plants, where plants locate across space so as to maximize profits and profits depend on area specific natural advantage, spillovers, and an i.i.d. drawing from Weibul distribution. The

idea is to explain the joint importance of spillovers and natural advantage in geographic concentration.

Geographic concentration for industry $j$ is $G_j = \sum_i (s_{ji} - x_i)^2$, where $s_{ji}$ is the share of industry $j$ in employment in location $i$ and $x_i$ is location $i$'s share in total national employment (to standardize for location size). Where $0 \leq \gamma^{na} \leq 1$ represents the importance of natural advantage (where the variance in relative profitability of a location is proportional to $\gamma^{na}$) and $\gamma^S$ represents the fraction of pairs of firms in an industry between which a spillover exists, Ellison and Glaeser show that

$$\begin{aligned} E[G_j] &= (1 - \sum_i x_i^2)(\gamma_j + (1 - \gamma_j)H_j) \\ \gamma &\equiv \gamma^{na} + \gamma^s - \gamma^s \gamma^{na} \end{aligned} \tag{18}$$

where $H_j$ is the standard Hirschman-Herfindahl index of plant industrial concentration in industry $j$. So $E[G_j]$ equals $\gamma_j$ adjusted for variations in location size $(1 - \sum x_i^2)$ and industry concentration $H$. The empirical part calculates $\gamma_j$ for all 3- or 4-digit manufacturing industries across states and countries. They show for 4-digit industries that $G > (1 - \sum x_i^2)H$ in 446 of 459 industries, where $G \leq (1 - \sum x^2)H$ only if $\gamma \leq 0$. That is almost all industries display some degree of spatial concentration due to either natural advantage or spillovers. Second they argue that 25% of industries are highly concentrated ( $\gamma > .05$) and 43% are not highly concentrated ( $\gamma < .02$). In a later article, Ellison and Glaeser (1999) argue that, based on econometric results relating location choices to natural advantage measures, 10-20% of $\gamma$ is accounted for by natural advantage. The rest is due to intra-industry spillovers, a rather critical finding in urban analysis indicating the importance of understanding the nature of scale externalities.

### *4.1.3. Geography*

A variety of recent studies have examined the role of geography, primarily natural features, in the spatial configuration of production and growth of cities. Rappaport and Sacks (2001) building on Sacks' general geography program herald the role of coastline location in the U.S.A., as a factor promoting city growth. In a related but more comprehensive study, Beeson, DeJong and Troeskan (2001) look at USA counties from 1840-1990. They show that iron deposits, other mineral deposits, river location, ocean location, river confluence, heating degree days, cooling degree days, mountain location, and precipitation all affect 1840 county population significantly. However for 1840-1990 growth in county population, only ocean location,

mountain location, precipitation, and river confluence matter, controlling for 1840 population. That is, first nature items strongly affected 1840 and hence indirectly 1990 populations; but growth from 1840-1990 is independent of many first nature influences. Ocean location as Sacks' suggests has persistent growth effects.

Both these studies ignore the geography of markets and the role of neighbors in influencing city evolution. Dobkins and Ioannides (2001) show that growth of neighboring cities influence own city growth and cities with neighbors are generally larger than isolated cities. Black and Henderson (2002) put neighbor and geographic effects together. They calculate normalized market potential variables (sum of distance discounted populations of all other counties in each decade, normalized across decades).They find climate and coast affect relative city growth rates; but market potential has big effects as well, although they are non-linear. Bigger markets provide more customers, but also more competition, so marginal market potential effects diminish as market potential increases. Market potential helps explain why North-East cities in the USA maintain reasonable growth given it is the most densely populated area from history, despite the natural advantages of the West.

Introducing market potential brings us full circle to the Krugman (1991) model exposited earlier. There is little empirical work on the model, with Hanson's work being a notable exception. Hanson (2000) examines wage relations across USA counties in an explicit Krugman monopolistic competition model, where scale derives from diversity of final consumption goods. By examining the effect on county wages and employment of surrounding economic activity, or market potential, by imposing the structure of the Krugman model, Hanson infers (1) that prices exceed marginal cost by 10-20%, (2) demand shocks attenuate quickly and disappear at about 400 miles, and (3) scale effects (diversity) are very strong relative to transport effects in driving geographic concentration.

### 4.2. Systems of Cities Models

Systems of cities models date back to Henderson (1974), with a variety of substantial contributors to further development (Hochman (1977), Kanemoto (1980), Henderson and Ioannides (1981), Abdel-Rahman and Fujita (1990), Helsley and Strange (1990), and Duranton and Puga (2002), to name a few). Here I outline the model in Black and Henderson (1999) which is an endogenous growth model of cities, that will thus lead directly to the growth-urban connection. The analysis is broken into two parts. The first examines the traditional static model, focused on city formation

and the determination of the sizes, numbers, and industrial composition of cities in an economy at a point in time. The second adds on the growth part.

*4.2.1. The System of Cities at a Point in Time*

Consider a large economy composed of two types of cities, where there are many cities of each type and each type is specialized in the production of a specific type of traded good. We will show why (when) there is specialization momentarily and the generalization to many types of goods and cities is straightforward. To simplify the growth story, each firm is composed of a single worker. In a city type 1, in any period, the output of firm $i$ in a type 1 city is

$$X_{1i} = D_1(n_1^{\delta_1} h_1^{\psi_1}) h_{1i}^{\theta_1} \; 0 < \delta_1 < \frac{1}{2} \tag{19}$$

$h_{1i}$ is the human capital of the worker and is his given input in the production process. A firm/worker is subject to two local externalities. First is own industry localization economies, the level of which depend on the total number of worker-firms, $n_1$, in this representative type 1 city. $n_1$ could represent the total volume of local spillover communications as in eq. (9), where $\delta_1$ is the elasticity of firm output with respect to $n_1$. The restriction $\delta_1 < \frac{1}{2}$ ensures a unique solution in an economy composed of many type 1 cities. The second externality, $h_1$, is the average level of human capital in the city and represents local knowledge spillovers, as in section 2.2.2. $h_1^{\psi_1}$ could be thought of as the richness of information spillovers $n_1^{\delta_1}$.

Given this simple formulation the wage of worker $i$ is simply

$$W_{1i} = X_{1i} \tag{20}$$

In an economy of identical individual workers in type 1 cities, individuals will all have the same human capital level (either exogenously in a static context, or endogenously in a growth context). Thus total city output will simply be

$$X_1 = D_1 h_1^{\sigma_1 + \psi_1} n_1^{1+\delta_1} \tag{21}$$

*Equilibrium City Sizes.*

Equations (19) and (21) embody the scale benefits of increases in local employment, where output per worker is an increasing function of local own industry employment. Determinant city sizes arise because of scale

diseconomies in city living, including per capita infrastructure costs, pollution, accidents, crime, and commuting costs. In Henderson (1974) those are captured in a general cost of housing function, but most urban models consider an explicit internal spatial structure of cities. All production occurs at a point – the center of the city. Surrounding the center in equilibrium in local land markets is a circle of residents each on a lot of unit size. People commute back and forth at a constant cost per unit (return) distance of $\tau$. That cost can be from working time, or here an out-of-pocket cost paid in units of $X_1$. Equilibrium in the land market is characterized by a linear rent gradient, declining from the center to zero at the city edge where rents (in agriculture) are normalized to zero. Standard analysis dating to Mohring (1961) gives us expressions for total city commuting and rents, in terms of city population where[6]

$$\text{total commuting costs} = bn_1^{3/2} \tag{22}$$

$$\begin{aligned} \text{total land rents} &= \frac{1}{2} bn_1^{3/2} \\ b &\equiv 2/3\pi^{-\frac{1}{2}}\tau. \end{aligned} \tag{23}$$

Equation (22) are the critical resource costs, where the marginal commuting costs of increasing city size are increasing in city population. Rents are income to, potentially, a city developer.

How do cities form and how are sizes determined? There are an unexhausted supply of identical city sites in the economy, each owned by a land developer in a nationally competitive urban land development market. A developer for an occupied city collects local land rents, specifies city population (but there is free migration in equilibrium), and offers any inducements to firms or people to locate in that city, in competition with other cities. Population is freely mobile. Helsley and Strange (1990) specify the city development game to determine how many cities will form and

[6] An equilibrium in residential markets requires all residents (living on equalize size lots) to spend the same amount on rent, $R(u)$, plus commuting costs, $\tau u$, for any distance $u$ from the CBD. Any consumer then has the same amount left over to invest or spend on all other goods. At the city edge at a radius of $u$, rent plus commuting costs are $\tau u_1$ since $R(u_1) = 0$; elsewhere they are $R(u) + \tau u$. Equating those at the city edge with those amounts elsewhere yields the rent gradient $R(u) = \tau(u_1 - u)$. From this, we calculate total rents in the city to be $\int_0^{u_1} 2\pi u R(u) du$ (given lot sizes of one so that each "ring" $2\pi u du$ contains that many residents) or $1/3\pi\tau u_1$. Total commuting costs are $\int_0^{u_1} 2\pi u(\tau u) du = 2/3\pi\tau u_1^3$. Given a city population of $n$ and lot sizes of one, $n_1 = \tau u_1^2$ or $u_1 = \pi^{-\frac{1}{2}} n^{\frac{1}{2}}$. Substitution gives us eqs. (20) and (21).

what their sizes will be. Given this game, Henderson and Becker (2001) show that resulting solutions (with multiple factors of production) are (1) Pareto efficient, (2) the only coalition proof equilibria in the economy, (3) unique under appropriate parameters (see below), and (4) free mobility ones where the developer specified populations are self-enforcing. They also show under appropriate conditions such outcomes arise (1) in a self-organized economy with no developers where city governments can exclude residents ("no-growth" restrictions) to maximize the welfare of the representative local voter, (2) in a growing economy where developers form new cities and old cities are governed by (even passive) local governments. Note for developing countries the key ingredients: either national land markets must be competitive with developers free to form new cities or atomistic settlements can arise freely and local autonomous governments can limit their populations as they grow. Without such institutions if, for example, cities only form through "self-organization", the result is enormously oversized cities (Henderson (1974), Henderson and Becker (2001)) where all net scale benefits are totally dissipated so the population is no better in cities than doing home production.[7]

In this context, the developer of a representative city chooses city population (or equivalently number of firms) and subsidies to locating firms/workers to maximize profits, or

$$\begin{aligned} \max_{n_1, T_1} \quad & \pi_1 = \frac{1}{2} b n_1^{3/2} - T_1 n_1 \\ \text{subject to} \quad & W_1 + T_1 - 3/2 b n_1^{\frac{1}{2}} = I_1 \end{aligned} \tag{24}$$

where $T_1$ is the per firm subsidy (e.g., in practice in a model with local public goods, a tax exemption). $I_1$ is the real income per worker available in equilibrium in national labor markets under free mobility, which a single developer takes as given. In the constraint, $I_1$ equals wages in (20) and (19), plus the subsidy, less per worker rents plus commuting costs paid. Maximizing with respect to $T_1$ and $n_1$ and imposing perfect competition in national land markets so $\pi_1 = 0$ ex post, yields

$$T_1 = \frac{1}{2} b n_1^{\frac{1}{2}} \tag{25}$$

[7] At the limit city sizes are so large with such enormous diseconomies that the population is indifferent between being in a rural settlement of size 1 (the size of a community formed by a defecting migrant) and an enormous oversized city. As we will see with an inverted-$U$ shape to real income $I_1$, self organization has cities at the right of the peak at $\tilde{n}$ where $I_1(n = 1) = I_1(n = \tilde{n})$ rather than where $I_1$ is maximized.

$$n^* = (\delta_1 2b^{-1} D_1)^{2/(1-2\delta_1)} h_1^{2\varepsilon_1} \tag{26}$$

$$\varepsilon_1 = \frac{\theta_1 + \psi_1}{1 - 2\delta_1} \tag{27}$$

This solution has a variety of properties heralded in the urban literature. First it reflects the Henry George Theorem (Flatters, Henderson, and Mieszkowski (1974), Stiglitz (1977)), where the transfer per worker/firm exactly equals the gap ($\delta_1 W_1$) between social and private marginal of labor to the city, and that externality subsidy is exactly financed out of collected land rents. That is, total land rents cover the cost of subsidies need to ensure Pareto efficient outcomes, as well as the costs of local public goods in a model where good goods are added in. Second the efficient size in (26) is the point where real income, $I_1$, peaks as an inverted $U$-shape function of city size (where $I_1 = W_1 + \frac{1}{2} b n_1^{\frac{1}{2}} - 3/2 b n_1^{\frac{1}{2}}$, where $3/2\ b n_1^{\frac{1}{2}}$ is per worker rents plus commuting costs and $\frac{1}{2} b n_1^{\frac{1}{2}}$ is per worker share in local land rents). If $\delta_1 < \frac{1}{2}$, we can show that $I_1$ is a single-peaked function of $n_1$, so $n_1^*$ is the unique efficient size. If $\delta_1 > \frac{1}{2}$, in essence there will only be one type 1 city in the economy, because net scale economies are unbounded. Given $n_1^*$ is the size where $I_1$ peaks, $n_1^*$ is a free mobility equilibrium – a worker moving to another city would lower real income in that city and be worse off. Finally city size, $n_1^*$ is increasing in technology improvements: $\tau$ declining, $\delta_1$ rising, $D_1$ rising, or local knowledge accumulation ($h_1$) rising.

*Other City Types and Specialization.*

In Black and Henderson, $X_1$ of city type 1 is an input into production of the single final good in the economy, $X_2$ (from which, hence in a growth context human capital is also "produced"). In many models all outputs of specialized city types are final consumption goods. Here $X_2$ is produced in type 2 cities where the output for worker/firm $j$ is correspondingly

$$X_{2j} = D_2 (n_2^{\delta_2} h_2^{\psi_2}) h_{2j}^{\theta_2} X_{1j}^{1-\alpha} \tag{28}$$

Here per worker output is also subject to own industry local scale externalities ($n_2^{\delta_2}$) and to local knowledge spillovers ($h_2^{\psi_2}$). However now there is an intermediate input, $X_{1j}$, which is the numeraire good with $X_{2j}$ priced at $P$ in national markets. The analysis of city sizes and formation for type 2 cities proceeds as for city type 1, with corresponding expressions other than the addition of an expression for $P$ in $n_2^*$ and $I_2$ and a restriction for an inverted $U-$shape to $I_2$ that $\delta_2 < \alpha/2$.

Two basic issues arise. Why do cities specialize and how are the equilibrium numbers of cities of each type and relative prices $P$ determined. On

specialization, in this model there are no costs of inter-city trade: no costs of shipping $X_1$ as inputs to $X_2$ types and shipping $X_2$ back as retail goods in $X_1$ type cities. All transport costs are internal to the city, given the relative greater importance of commuting costs in modern economies. Given that and given scale economies are internal to the industry, any specialized city out-competes any mixed city. The heuristic argument is simple. Consider any mixed city with $\tilde{n}_1$ and $\tilde{n}_2$ workers in industry 1 and 2. Split that city into two specialized cities, one with just $\tilde{n}_1$ people and the other with just $\tilde{n}_2$. Scale economies are undiminished ($\tilde{n}_1^{\delta_1}$ and $\tilde{n}_2^{\delta_2}$ in both cases in industries 1 and 2 respectively) but per worker commuting costs are lower in the specialized cities compared to the old larger mixed cities, so real incomes are higher in each specialized city compared to the old city. The rigorous argument is a little more subtle in the growth context where human capital levels, $h_1$ and $h_2$, differ endogenously across industries and affect incomes.[8] Having localization economies is a sufficient but not necessary condition for specialization. Industries can have urbanization economies so scale depends on total local employment. However if the degree of urbanization economies differs across industries (the corresponding $\delta_1 \neq \delta_2$) then each industry has a different efficient local scale and is better off in a different size specialized city than any mixed city. In fact mixed cities are more likely to emerge if each good has localization economies multiplied by separate spillovers from the other industry or sharing of some common public infrastructure (Abdel-Rahman (2000)).

A basic problem in the pre-economic geography urban models is the lack of nuance on transport costs. Either transport costs of goods across cities is zero ($X_1$ and $X_2$) or infinite (housing, and potentially other non-tradeables). A recent innovation is to have generalized transport costs (without a specific geography) where the cost of transporting a unit of $X_1$ to an $X_2$ city is $t_1$ and the cost of shipping $X_2$ back to an $X_1$ city is $t_2$, an innovation due to Abdel-Rahman (1996) in a model similar to the static one used here (one intermediate and one final good) and then generalized by Xiong (1998) and Anas and Xiong (2001). Now specialization as opposed to diversified cities depends on the level of $t_1$ and $t_2$. At appropriate points as $t_1$ or $t_2$ or both rise from zero, $X_1$ and $X_2$ will collocate (in developer run cities). More generally with a spectrum of, say, final products, we would expect that some products have low enough $t$'s to always be produced in specialized cities, some high enough $t$'s to be in all cities, and some in

[8] It raises issues of low education types potentially benefiting from high education type externalities, in a context where separation is desirable but a separating equilibrium costly to maintain (Black (2002)). See later.

middle range $t$'s are produced in some cities (ones with bigger markets) but not others (with smaller markets). No one has yet simulated this more complex outcome.

The second issue concerns how to close the model in a static context and solve for $P$ the relative price of $X_2$ and $m_1$ and $m_2$ the number of cities of each type. In a large economy integer problems are ignored and a full employment constraint imposed so

$$m_1 n_1 + m_2 n_2 = N \tag{29}$$

where $N$ is national population. The second equation (to solve the 3 unknowns $P$, $m_1$, and $m_2$) equates real incomes across cities ($I_1 = I_2$) only in static context. That is in a static context individual workers move across cities to equalize real incomes. Finally there is an equation where national demand equals supply in either the $X_1$ or $X_2$ market (i.e., the supply, $m_1 X_1$, equals the demand for $X_1$ as an intermediate input, $m_2 n_2 x_1$, and for producing commuting costs $m_1(bn_1^{3/2}) + m_2(bn_2^{3/2})$ from eq. (20)). In this specific model that will yield values of $m_1$, $m_2$ and $P$ that are functions of parameters and $h_1$ and $h_2$. In a static context of identical workers, one would impose $h = h_1 = h_2$. We will discuss momentarily the solution for $h_1$ and $h_2$ and the model in the growth context.

In the static context where labor mobility requires $I_1 = I_2$, in the larger type of city, say type 1, commuting and land rent costs will be higher. Thus, if real incomes are equalized, $W_1 > W_2$ as a compensating differential for higher living costs. Firms in type 1 cities are willing to pay higher wages because type 1 cities offer them greater scale benefits. Empirical evidence shows as cities increase from a small size (say, 50,000) to very large metro areas, both the cost-of-living and real wages double (Henderson (1988)).

In a static context, at the national level there are constant returns to scale or replicability. If we double national population, the numbers of cities of each type and national output of each good simply double, with individual city sizes and real incomes unchanged.[9] With two goods and two factors basic international trade theorems (Rybczynski, factor price equalization, and Stolper-Samuelson) hold (Hochman (1977), Henderson (1988)).

**Policy in a System of Cities.** The insight that large urbanized economies are replicable with CRS is important, since it simplifies pol-

[9]Here with $h_1$ and $h_2$ yet to be solved we would need to double the numbers of people with $h_1$ and $h_2$ respectively. Below we will see the solution with growth to $h_1$ and $h_2$ is national scale invariant.

icy analysis. Policy analysis of system of cities is not a focus of recent work, but Henderson (1988) considers the effects of a variety of policies. For example trade protection policies favoring industry $X$ produced in relatively large size cities will alter national output composition towards $X$ production and increase the number of large relative to small cities. National urban concentration will rise. Similarly subsidizing an input such as capital for a high tech product, $X$, again, say, produced in a larger type of city will cause the numbers of that type of city to increase and raise urban concentration. As another example, national minimum wage policies may not bite in large high wage cities but will bite in smaller low wage cities. In general cities subject to binding minimum wages will increase in size, but their numbers and overall production will decline. In order to pay the legislated higher wages, relative prices of those products rise (as supply declines) and greater city sizes generate greater local scale effects.

Another issue is that policy makers may favor large cities because they view them as "more productive". Indeed for an industry found in smaller towns, it may be that the $A(\cdot)$ they face in eq. (7), their technology level including whatever externalities, may be higher in a larger city. However that doesn't mean they locate there. Although the $A(\cdot)$ may be higher, in order for them to locate there, it must be sufficiently relatively higher to afford the higher wage and land rents, compared to a smaller city. If not, their profit maximizing or cost minimizing location is the smaller city.

### *4.2.2. Growth in a System of Cities*

Black and Henderson (1999) specify a dynastic growth model where dynastic families grow in numbers at rate $g$ over time starting from size 1. If $c$ is per person family consumption, the objective function is $\int_0^\infty (\frac{c(t)^{1-\sigma}-1}{1-\sigma}) e^{-(\rho-g)t} dt$ where $\rho(> g)$ is the discount rate. Dynasties can splinter (as long as they share their capital stock on an equal per capita basis) and the problem can be put in an overlapping generations context with equivalent results (Black (2000)), under a Galor and Zeira (1993) "joy of giving" bequest motive.

The only capital is human capital and as such there is no market for it. Intra-family behavior substitutes for a capital market. Specifically families allocate their total stock of human capital (H) and members across cities, where $Z$ proportion of family members go to type 1 cities (taking $Zh_1e^{gt}$ of the $H$ with them) and $(1-Z)$ go to type 2 cities taking $(1-Z)\ h_2e^{gt}$ with them). Additions to the family stock come from the equation of motion where the cost of additions, $P\dot{H}$, equals family income ($Ze^{gt}I_1 + (1 -$

$Z)e^{gt}I_2$ less the value of family consumption of $X_2$, or $Pce^{gt}$. Constraints prohibiting consumption of human capital, non-transferability except to newborns, and non-transferability within families across city types (either directly or indirectly through migration) are non-binding on equilibrium paths.

Families allocate their populations across types of cities, with low human capital types (say $h_1$) "lending" some of their share ($h = H/e^{gt}$) to high human capital types (say $h_2$). High human capital types with higher incomes ($I_2 > I_1$ if $h_2 > h_1$) repay low human capital types so $c_1 = c_2 = c$ (governed by the family matriarch). This in itself is an interesting development story, where rural families diversify migration destinations (including the own rural village) and remittances home are a substantial part of earnings. Fujita and Thisse (2001) model a life cycle version where workers migrate to the core region to accumulate savings to take back to the periphery to invest in physical capital there, under imperfect capital markets. In Black and Henderson if capital markets operate perfectly for human capital (i.e., we violate the "no slavery" constraint) or capital is physical and capital markets operational, one dynastic family could move entirely to, say, type 1 cities and lend some of their human capital to another dynastic family in type 2 cities. With no capital market, each dynastic family must operate as its own informal capital market and spread itself across cities.

In this context Black and Henderson show that, regardless of scale or point in the growth process, $h_1/h_2$ and $I_1/I_2$ are fixed ratios, dependent on $\theta_i$. As $\theta_1/\theta_2$ rises (relative returns to capital), $h_1/h_2$ and $I_1/I_2$ rise. $Z$ and $m_1/m_2$ are all fixed ratios of parameters $\theta_i, \delta_i, \alpha$ under equilibrium growth. Only $P$ is a function of human capital accumulation (increasing if $(\theta_1+\psi_1)/(1-2\delta_1) > (\theta_2+\psi_2)/(\alpha-2\delta_2)$. Equilibrium and optimal growth differ because the private returns to education in a city, $\theta_i$, differ from the social returns, $\theta_i + \psi_i$. But local governments can't intervene successfully to encourage optimal growth. Why? With free migration and "no slavery", if a city invests to increase its citizens' education, a person can take their human capital ("brain drain") and move to another city (be subsidized by another city to immigrate, given that city then need not provide extra education for that worker). This model hazard problem discourages such education subsidization.

**Growth properties: Cities.** From eq. (24), equilibrium (and efficient) city size in type 1 cities is a function of the per person human capital level, $h_1$, in type 1 cities. After solving out for $P$ the same is true of type 2 cities. City sizes grow as $h_1$ and $h_2$ grow, where under equilibrium growth given $h_1/h_2$ is a fixed ratio $\dot{h}/h = \dot{h}_1/h_1 = \dot{h}_2/\dot{h}_2$ where a dot represents a time

derivative. Then

$$\frac{\dot{n}_2}{n_2} = \frac{\dot{n}_1}{n_1} = 2\varepsilon_1 \frac{\dot{h}}{h} \tag{30}$$

where $\dot{n}_i/n_i$ is the growth rate of efficient sizes $n_i^*$.

For the number of cities, the issue is whether growth in individual sizes absorbs the national population growth, or

$$\frac{\dot{m}_1}{m_1} = \frac{\dot{m}_2}{m_2} = g - \frac{\dot{n}_i}{n_i} = g - 2\varepsilon_1 \frac{\dot{h}}{h} \tag{31}$$

The numbers of cities grow if $g > \dot{n}_i/n_i$. Note growth in numbers and sizes of cities is "parallel" by type, so the relative size distribution of cities is constant over time.

**Growth properties: Economy.** Ruling out explosive or divergent growth, there are two types of growth equilibria. Either the economy converges to a steady state level (where $\gamma^c \equiv \dot{c}/c = \frac{1}{\sigma}(Ah^{\varepsilon-1} - \rho)$), or it experiences endogenous steady-state growth. Convergence to a level occurs if $\varepsilon \equiv \varepsilon_1(1-(\gamma-2\delta_2))+\varepsilon_2(\gamma-2\delta_2) < 1$, where $\varepsilon$ is a weighted average of the individual city type $\varepsilon_i$. In that case at the steady-state $\bar{h}, \dot{n}_i/n_i = 0$ and $\dot{m}_i/m_i = g$, or only the numbers but not sizes of cities grow just like in exogenous growth (Kanemoto (1980), Henderson and Ioannides (1981)). If $\varepsilon = 1$ then there is steady-state growth, where $\bar{\gamma}^h = \dot{h}/h = \frac{A-\rho}{\sigma}$ (where the transversality conditions require $A > \rho$). In that case $\dot{n}_i/n_i = 2\varepsilon_1(\frac{A-\rho}{\sigma})$, or cities grow at a constant rate. and their numbers also increase if $g > 2\varepsilon_1(\frac{A-\rho}{\sigma})$.

### *4.2.3. Extensions*

There are two major extensions to the basic systems of cities models. First people may differ in terms of inherent productivity or in terms of endowments. Second, while we have discussed the issue of city specialization versus diversification we haven't really developed any insights into a more nuanced role of small highly specialized cities versus large diversified metro areas in an economy.

Turning to the first extension, Henderson (1974) had physical capital as a factor of production owned by capitalists who needn't reside in cities. Then equilibrium city size reflects a market trade-off between the interests of city workers who have an inverted $U-$shape to utility as a function of the size of the city they live in and capitalists whose returns to capital rise indefinitely with city size (for the same capital to labor ratio). There is a political economy story there where capitalists collectively in an economy

have an incentive to limit the number of cities, thus forcing larger city sizes. Helsley and Strange (1991) and then Becker and Henderson (2000) have matching models between the attributes of entrepreneurs and workers, as noted earlier. But again the two class model yields a market resolved conflict between what is the city size that maximizes the welfare of one versus another group.

In a different approach Abdel-Rahman and Wang (1997) (see also Abdel-Rahman (2000) for a synthesis) and later (Black (2000) look at high and low skill workers who are used in differing proportions in production of different goods. Black has a low skill traded production good and a second traded good produced with high skill workers and inputs of a low skill non-traded good, where high skill workers generate production externalities in the form of knowledge spillovers for all traded goods. In Black, urban specialization with high skill workers concentrated in one type of city is efficient, but a separating equilibrium, where low skill workers and low tech production stays in its own type of city (rather than trying to cluster with high tech production) is not easy to sustain. Black characterizes conditions under which a separating equilibrium will emerge.

Abdel-Rahman and Wang (1997) impose an urban core-periphery structure where a high tech good can only be produced by heterogeneous skilled workers but the low tech good by either those workers or homogeneous unskilled workers. Urban scale economies arise in public infrastructure provision as well as better matching of heterogeneous skilled workers. The low tech good is assumed to be produced in a system of hinterland (peripheral) cities and the high tech good in the core region metropolis. The focus is on determinants of income inequality, although much of the work revolves around the Nash bargaining process in the matching process between heterogeneous skilled workers and the firms which hire them, and less on endogenous properties of systems of cities.

It is important to note that there is a much more developed literature on inequality induced by neighborhood selection, where the characteristics of neighbors affect skill acquisition (e.g. family background of the class affects individual student performance). That leads to segregation of talented or wealthier families by neighborhood (Benabou (1993), Durlauf (1996)) and can help transmit economic status across generations, promoting intergenerational income inequality.

**Metro Areas**. Simple indices of urban diversity indicate that smaller cities are very specialized and larger cities highly diversified. So the question is what is the role of large metro areas in an economy and their relationship to smaller cities. Henderson (1988) and Duranton (2002) have a

first nature - second nature world, where every city has a first nature economic base and footloose industries cluster in these different first nature cities. Large metro areas are at the top of an urban hierarchy in Henderson (1988), with first nature activity benefiting most from local scale externalities and with the greatest varieties of footloose activity clustered in the metro area. The smallest cities are engaged in specialized first nature activity with minimal scale externalities, where the local market doesn't attract much footloose production. But it seems that today few metro areas have an economic base of first nature activity. Accordingly recent literature has focused on the role of large metro areas as centers of innovation, headquarters, and business services (Kolko (1999)).

The Dixit-Stiglitz model opened up an avenue to look at large metro areas as having a base of diversified intermediate service inputs, which generate scale-diversity benefits for local final goods producers. That initial idea was developed in Abdel-Rahman and Fujita (1990). That idea has led to a set of papers focused on the general issue of what activities, under what circumstances are out-sourced. Theory and empirical evidence (Holmes (1999) and Ono (2000)) suggests that as local market scale increases, final producers will in-house less of their service functions and out-source them more. That out-sourcing encourages competition and diversify in the local business service market, encouraging further out-sourcing.

In terms of incorporating this into the role of metro areas versus smaller cities, Davis (2000) has a two-region model, a coastal exporting region and an interior natural resource rich region. There are specialized manufacturing activities which, for production and final sale, require business service activities, summarized as headquarters functions. Headquarters purchase local Dixit-Stiglitz intermediate services such as R&D, marketing, financing, exporting, and so on. Headquarters activity is in port cities in the coastal region. The issue is whether manufacturing activities are also in these ports versus in specialized coastal hinterland cities versus in specialized interior cities. Scale economies in manufacturing and headquarters activities are different and independent of each other, so that, based on scale considerations, these activities would be in separate specialized cities. However if the costs of interaction (shipping manufactured goods to port and transactions costs of headquarters-production facility communication) between headquarters and manufacturing functions are extremely high, then both manufacturing and headquarters activities can be found together in coastal port cities. Otherwise they will be in separate types of cities. In that case, manufacturing cities will be in coastal hinterlands if costs of headquarters-manufacturing interaction are high relative to ship-

ping natural resources to the coast. However if natural resource shipping costs are relatively high, then manufacturing cities will be found in the interior.

Duranton and Puga (2001) have developed an entirely different and stimulating view of large metro areas. In an economy there are $m$ types of workers who have skills each specific to producing one of $m$ products. Specialized cities have 1 type of worker producing the standardized product for that type of worker subject to localization economies. Diversified cities have some of all types of workers. Existing firms at any instant die at an exogenously given rate; and, in a steady-state, new firms are their replacement. New firms don't know "their type" – what types of workers they match best with and hence what final product they would be best off producing. To find their type they need to experiment by trying the different technologies (and hence trying different kinds of workers). New firms have a choice. They can locate in a diversified city with low localization economies in any one sector. In a diversified city they can experiment with a new process each period until they find their ideal process. At that point they relocate to a city specialized in that product, with thus high localization economies for that product. Alternatively new firms can experiment by moving from specialized city to specialized city with high localization economies, but face a relocation cost each time. If relocation costs are high, the advantage during their experimental period is to be in a diversified city. This leads to an urban configuration of experimental diversified metro areas and other cities which are specialized in different standardized manufacturing products.

The Duranton and Puga model captures a key role of large diversified metro areas consistent with the data. They are incubators where new products are born and where new firms learn. Once firms have matured then they typically do relocate to more specialized cities. This also captures the product-life cycle for firms in terms of location patterns. Fujita and Ishii (1994) document the location patterns of Japanese and Korean electronics plants and headquarters. In a spatial hierarchy mega-cities house headquarters activities (out-sourcing business services) and experimental activity. Smaller Japanese or Korean towns house specialized, more standardized high tech production processes and low tech activity is off-shore.

## 5. URBAN ISSUES IN CHINA

In this last section I turn to a specific application of the urbanization and economic geography models to China. Chinese urbanization has some

special features driven by historical and current policies affecting urbanization. I first discuss features of Chinese urbanization and key policies. Then I turn to a review of analyses of the impacts of these policies on production, growth and efficiency of Chinese cities.

## 5.1. Some Key Features and Policies of the Chinese Urban System

### *5.1.1. Low Urban Concentration*

Chinese cities are relatively small and equal sized, compared to most countries. The UN puts the population of Shanghai metro area, the largest city, at 12.3m in 2000, well below the populations of the 10 largest metro areas in the world. More critically is that China only has 9 metro areas with populations over 3 million while it has another 125 or so metro areas with populations from 1-3 million; – a ratio of .072, compared to the worldwide ratio for the same size categories of .27 (Henderson (2002c)). To give a more common frame of reference for comparisons, we examine spatial Gini's.

For 1657 metro areas with populations over 200,000 in 2000 for the world, the spatial Gini is .564. The Gini is the usual one: rank cities from smallest to largest and plot the Lorenz curve of their accumulated share of total population for the sample (world cities in this case). The Gini is the share of area below the 45° line that lies between the 45° line and the Lorenz curve. China's Gini is .43 in 2000, way below the world, and compares to .65, .65,.61 ,.60, .60, .60, .59, .58, .56, .54 and .52 for other large countries respectively of Brazil, Japan, Indonesia, UK, Mexico, Nigeria, France, India, Germany, USA and Spain. Only former Soviet bloc countries have similarly low Gini's, Russia with .45 and Ukraine with .40.

In the second part of this section, we will argue that Chinese cities in general are too small, leading to significant efficiency losses. In fact we will argue more generally that there is insufficient spatial agglomeration throughout, in both the urban and rural sector.

### *5.1.2. The Hukou System*

In China, the geographic-urban dispersion of population is maintained by strong migration restrictions, under the hukou system. Migration restrictions play such a strong role in the society and economy, it is critical to describe them. The hukou system in China is similar to an internal passport system (see Chan (1994) for a detailed description). A person's local "citizenship" and residence is initially defined for a child as a birth right, traditionally by the mother's place of legal residence. The entitlements and details of the system differ for urban and rural residents. Legal residence

in a city entitles one to local access to permanent jobs, regular housing, public schooling, and public health care (where almost all health care is public) in that city. Until the early 1990's, it also entitled urban people to "grain rations" – rations of essentials such as grain and kerosene.

Legal residence in a village or rural township entitles residents to land for farming, township housing, job opportunities in rural industrial enterprises, and access to local health and schooling facilities in their town. Residents also have some degree of "ownership" in local enterprises; although distributed profits all go to the local public budget, which may be used to finance township housing and infrastructure. Again, until recent years, legal residence in a township also entitled a "peasant" to some share in locally produced (or allocated from the outside) grain and other essentials.

How does a person change their local citizenship? There are several common mechanisms. First is education. A smart rural youth may persist through the competitive school system to go to a college and then be hired into an urban job, with an urban hukou. Second, the state at times can open the gates, permitting factories to hire permanent workers from rural areas, permitting family reunification, or permitting legal migration from rural areas to nearby small cities. However the official changes in residence or hukou status average about 18 million (in under 1.3% of the population) a year over the last 20 years with little annual variation (Chan (2000)).

People can migrate to an area without local hukou, or an official change of residence "citizenship", either illegally ("unregistered") or legally as a temporary worker or as a "permanent resident" on a long-term permit. For example, a rural person may be hired as a "contract worker" in industry or services, for a term of three years. People may move illegally, without registering in the new location, and work in the informal sector for low pay, under poor conditions, with risk of deportation. Despite these possibilities and despite some recent relaxations of restrictions in particular provinces, the restrictions in migration remain tight.

Temporary migrants to larger cities typically have no, or very high priced access to health care and schooling facilities and regular, "legal" housing. In fact cities have strict national guidelines on conversion of agriculture to urban land; and institutional difficulties in housing markets in expanding supply makes it particularly difficult for migrants to find decent housing. All this means living and social conditions for migrants and their families are extremely difficult, since children face no or very high priced access to schooling and health care. But there are other restrictions. Legal temporary migration requires getting a permit from the city of in-migration and cities can impose various hurdles to getting a permit – permission from

the home location, proof of a guaranteed job and specific housing, and the like. Cities have also published job lists, citing jobs for which migrants are not eligible; in 2000, Beijing listed over 100 occupants as non-eligible ones. Migrants may still have to pay taxes to their rural home village for services they don't consume and on land left fallow. Finally migrants have traditionally faced direct fees (Cai (2000)). There is a license fee to work outside the home township paid to the township that can be equivalent to several months' wages. At the destination there can be fees for city management, for being a "foreign" worker, for city construction, for crime fighting, for temporary residence, and even for family planning if the migrant is female. All these restrictions sharply reduce the benefits and raise the costs of migration, particularly into large cities. Migration is limited and most migration is short-term, or "return" migration, as we will detail with data below. Overall the hukou system holds 100's of millions of people in locations where they are not exploiting their earning potential, as we will detail below.

### *5.1.3. Aspects of Urban Policy Since 1978*

As defined in part by the 1982 Sixth 5-Year Plan, as well as the Seventh 5-Year Plan, the post-Mao period has a set of initially defined urbanization policies that persist today. Good sources on aspects of these policies include Chan (1994), Kojima (1996), Fujita and Hu (2001), and Wei and Wu (2001). First urban population was to expand, but through the rapid growth of smaller cities relaxing hukou transfers at the level, while containing the sizes of larger cities. The 1990's witnessed the rapid growth in number of cities, as many places were recognized as having passed 100,000 urban population mark. However China's spatial Gini and degree of urban concentration remains very low by world standards and even lower in 2000 than in 1960 (.42 versus.47). General urban population expansion has also been fueled by rapid growth, particularly in coastal towns, of township populations, always pushing these towns towards (or past) the 100,000 mark to be a city (Ma and Fan 1994).

In the Sixth and Seventh 5-year plans there is a sense of hierarchy, played out both in governance structures and in economic policy. Larger cities are to lead smaller ones and rural areas; the coast is to lead the center and west. "Leading" has many dimensions. Larger cities focus on newer production – initially high tech and light industry and then business service development in recent years. Large cities receive new technologies and hand-down traditional activities to their hinterlands, in particular contracting-out parts

and components production to small cities, towns, and rural areas, and relocating heavy, polluting production to their ex-urban areas.

Another aspect of urbanization policy, implicit and as part of big cities leading the rest, is played out in the development of rural industry – the town and village enterprise sector [TVE's]. The rapid productivity growth in agriculture after 1978, coupled with prior restrained urbanization, meant a vast surplus of labor in agriculture. Given the desire to continue to restrain urbanization (although at a much higher rate after the 15 or so years before 1978), a policy of "leave the land but not the village" was crafted. So the rural sector was to industrialize, but generally not spatially agglomerate. TVE development was constrained by under-capitalization, an inability to spatially agglomerate, and in the 1980's policies restraining its competition with SOE's (followers are not supposed to out-compete leaders!).[10] However, TVE growth was rapid: starting from an initial miniscule level, by 1997 VA in the TVE sector was twice that in remaining SOE's (independent accounting units). TVE's had hard budget constraints, fewer regulations, and greater ability to respond to input (hiring and promotion, choice of sellers of intermediate inputs) and output (product demand) market conditions. By the early 1990's, Jefferson and Singhe (1999) document how TVE total factor productivity exceeded SOE's, ascribing that to the greater operational freedom and hard budget constraint of TVE's.

Still TVE sector development was constrained by the under-capitalization of the rural sector that has been a feature of modern China. Based on micro data, Jefferson and Singhe show that the rate of return on physical capital in the TVE sector in 1997 exceeded that of SOE's by 25%. In addition the higher wage and compensation returns to labor in the urban sector combined with college education being the key to permanent migration from rural to urban areas, means the more educated population is funneled into cities. An area of investigation is the very high social returns to education in the rural sector, improving township allocation decisions of resources between agricultural, animal husbandry and TVE activities (Yang and Au (1997)).

In addition to these policies governing rural-urban (and big city-small city) allocation of capital and labor there are other much more explicit policies with a spatial bias (Chan (1994), Naughton (2002), Fujita and Hu (2001)). While they have some big city-small city/town flavor, they also have a coastal versus rest of the country flavor. Arguably the key element

[10] Usually shifting restrictions on products that could be produced by TVE's were the competition restraint. Success by a TVE in competition could lead to its product line being terminated (Henderson (1988)).

is initial policies that directed FDI and trade to certain coastal cities. In the early (1979) reforms, 4 coastal special economic zones, centered on 4 prefecture level cities were created to encourage free market experimentation, an inflow of FDI, and development of international trade. In the mid-1980's, 14 more coastal cities were declared as open cities to foreign investors, with 2 more coastal cities added by 1990. In addition 10 cities (half overlapping with open status) were given separately listed status – economic decision making powers equal to the provincial cities of Beijing, Tianjin, and Shanghai.

Fujita and Hu (2001) show that 14 open cities and 4 spatial economic zones accounted for 42% of national FDI from 1984-1994. In 1990, the 24 "special status" cities (special economic zones, open, and separately listed) plus Beijing accounted for 65% of all FDI in prefecture level cities, while accounting for only 36% of value-added in non-agricultural production of prefecture level cities. This initial advantage persists, despite opening of the entire economy. For example, these 25 cities account for 63% of all FDI accumulated from 1990-1997 in all prefecture (or provincial) level cities.

Fujita and Hu (2001) argue persuasively that the agglomeration of electronics and light manufacturing in coastal areas such as the region around Guangzhou is due to these initial policies promoting FDI and trade in these favored coastal areas. The effect is reflected in the ratio of investment occurring in coastal versus interior regions: in 1984 the ratio is 1.12 while 10 years later it is 1.93 (see also Naughton, 2002). These policies and their impacts are deliberate spatial policies of the Sixth and Seventh 5-year plans, favoring development in a spatial hierarchy of the coastal region. On a more positive note, Wei and Wu (2001) do show the expanding trade within these favored regions, tended to reduce rural-urban income inequality, because trade helped the TVE sector in the urban fringe (rural) areas.

An entirely different aspect of this spatial policy bias involves transportation. On a world scale, China has an anemic road system. Its ratios of national roads to land, roads to population, or paved roads to land or to population in 1995 are very low by international standards. For example, its ratio of roads to population (which is better than using land as the normalization or looking at paved roads) in 1995 is 1.2, to be compared to 2.1 for India, 1.5 for Pakistan, 1.9 for Indonesia, or 2.7 for Mexico. Half of these roads are paved in India but only 15% in China. For a country with far-flung populations, this places hinterlands at great disadvantage in their ability to secure inputs and truck products to coastal and international

markets. Only now is a highway being built to link Chengdu, Sichuan's capital, and the 100m. people in the Sichuan region to the coast.

The final aspect of spatial bias involves governance and fiscal relations. Fiscal rules and inter-governmental relations in China are not well defined and transparent. Revenue redistribution contracts send monies coming from the center back to provinces and cities; up to the mid-1990's these appeared to favor bigger and richer cities. But much official public expenditure is extra-budgetary – local revenues retained within localities (Jin and Zou (2002)). What is retained and the specifics of a city's fiscal allocations from above, whatever the rules, are in part the result of bargaining. And in the hierarchy of big city versus small or coast versus interior, the bigger and the coastal have greater bargaining power. Actual results depend on the personalities and power of local leaders, with an interesting literature on China documenting this (Cheung, Chung, and Lin (1998)). Bigger cities have more effective fiscal autonomy and more control over key appointments (e.g., heads of local state-owned banks which become a source of funds and subsidies of local industries). Cities compared to rural areas are favored with the ability to offer lower tax rates on FDI firms. The issue is a difficult one and there has emerged no clear way to quantify the fiscal advantage of one city or set of cities over others. But the spatial bias and lack of transparency is a key feature of China's urban sector.

### 5.2. Some Effects of Policy

This section focuses on agglomeration economies and city sizes in China – the extent to which cities in China may be too small. It examines the welfare losses from under-sized cities and the extent of rural-urban migration that is actually occurring . At the end we will turn to the issue of spatial bias and history of pro-coastal policies.

#### *5.2.1. Under-agglomeration in Cities*

Using data for 1996 and 1997, Au and Henderson (2002) estimate city production functions for 212 prefecture level (or above) cities. Output is value-added per worker in the non-agricultural sector of the city proper. Determinants include the capital stock to labor ratio, share of accumulated FDI in capital stock, distance to the coast, education and scale measures. With respect to the last item, real output per worker following traditional systems of cities analysis outlined in section 3 is postulated to be an inverted U-shape function of local scale, as measured by total local non-agricultural employment. At low scale, the marginal benefits in terms of increased productivity of increased local scale from enhanced scale externalities and local

**TABLE 1.**

Efficient City Sizes

| A. City Employment at Peak of VA per Worker | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| manufacturing to service ratio | .6 | .8 | 1.0 | 1.2 | 1.4 | 1.6 | 1.8 | 2.0 |
| $L^*$ | 2730 | 2380 | 2030 | 1670 | 1320 | 970 | 620 | 270 |
| 95% confidence interval | | | | | | | | |
| - lower | 1880 | 1680 | 1420 | 1090 | 670 | 180 | | |
| - upper | 3590 | 3080 | 2630 | 2260 | 1980 | 1760 | 1580 | 1430 |
| B. Gain from moving to $L_i^*$ | | | | | | | | |
| percent current size is below peak | | 50 | | 40 | | 30 | | 20 |
| percent gain in $VA$ per worker | | 35% | | 20% | | 9.5% | | 4.1% |

backward and forward linkages outweigh the marginal costs from increased congestion and commuting costs and environmental degradation. So at low city scale, real value-added per worker is increasing in scale, then at some city size it peaks, and after that declines with further increases in city scale.

However cities are in an economic "hierarchy" where cities relatively and absolutely specialize in different products. In general in that context, the manufacturing to service ratio of cities declines as city scale rises, or cities move up the hierarchy. To capture this, Au and Henderson (2002) postulate that the inverted U-shape shifts right as the manufacturing to service ratio drops. They estimate a relationship for city $i$ where

$$\ln(VA_i/L_i) = \beta X_i + \alpha_1 L_i + \alpha_2 L_i^2 + \alpha_3 L_i \cdot MS_i \tag{32}$$

In (32), $X_i$ are controls on technology, capital-labor ratio, access and the like. $VA_i$ is value-added; $L_i$ is total (non-agricultural) employment, and $MS_i$ is the manufacturing to service $VA$ ratio (seconday to tertiary sector $VA$). In eq. (32) output per worker peaks where

$$L_i^* = \frac{\alpha_1 + \alpha_3 MS_i}{-2\alpha_2} \tag{33}$$

where $\alpha_2 < 0$, $\alpha_1 > 0$, $\alpha_1 + \alpha_3 MS_1 > 0$ and $\alpha_3 < 0$. The last reflects the economic hierarchy idea: bigger cities are more service oriented, so $L_i^*$ declines as $MS_i$ rises.

Estimation of (32) in Au and Henderson (2002) is by instrumental variables using 1990 (planning period) variables as instruments. Estimation

produces a tight fit with excellent specification test results. Table 1A shows relevant manufacturing to service ratios, the peak points ($L_i^*$). where value-added per workers is maximized and the 95% confidence interval for peak scale. Note scale is in thousands of workers. Most Chinese cities (85%) lie to the left of their peak points and 43% are below the 95% confidence interval on $L_i^*$. That is, 43% of cities are significantly to the left of $L_i^*$, or significantly undersized. Table 1B shows the percent gain in $VA$ per worker from moving below the peak to the peak. About 50% of cities are 50% or more below their peak size, with resulting large productivity losses.

For county-level cities, Au and Henderson are unable to quantify an inverted-U, instead finding unbounded scale effects (for these smaller city sizes). Similarly for TVE's across provinces, local scale economies (average township TVE employment by province over three years) are unbounded and very large – a 10% increase in local scale increases value-added by worker by 3%. This is the same order of magnitude found by Jefferson and Singhe (1999) to TVE scale, using micro data.

The conclusion is that throughout China there is under-agglomeration, held in place by the hukou system, and also property right issues in rural areas. For the latter, there is no ability to readily transfer TVE ownership and location, for township residents to sell their "shares" in local TVE's so as to liquidate and relocate, or for township residents to shift location to another town. That makes rural agglomeration difficult. However here we focus more on rural-urban migration. But free migration in China would change the landscape – some prefecture and county-level cities would experience huge population inflows over a period of years. Some townships would also experience huge inflows and transform into major cities. Conversely, these flows imply some cities and towns would experience large population losses.

### *5.2.2. Extent of Actual Rural-Urban Migration*

In the popular press, there is sometimes a sense that there is already enormous migration in China, with the transformation well underway. Certainly a transformation is underway, and may be more advanced in provinces such as Guangdong; but the issue is the extent of overall population movements. In 1998, the commonly accepted number for the "floating" population – those currently outside their town of residence for more than 1-3 days – was about 100m of 1.2b or so people. From Chan (2000), several factors are apparent concerning these 100m. First the number of annual permanent residence changes has been constant at about 18m for the prior

15 years. About 15% of the population relocates every 10 years, including, as we will suggest, a substantial portion of rural-rural and urban-urban moves. Second, in general, most temporary migration in China is return migration – migrants move for a few months or years and then return home, rather than remaining as temporary migrants in a destination indefinitely. Third, most of even this temporary migration is short distance.

**TABLE 2.**

Migration in China

A. Stocks of the Population

| | |
|---|---|
| floating population (outside of township of residence ) | 100m (estimated) |
| temporary migrants (outside of township of residence for more than 6 months) in 1998 | 62.4 m |
| percent of temporary migrants living outside home county (1995) | 59% |

B. Flows of Population 1990-1995 Rural/Urban (origin $\longrightarrow$ destination) Percent

| | |
|---|---|
| U$\longrightarrow$U | 35.4 |
| U$\longrightarrow$R | 4.8 |
| R$\longrightarrow$R | 23.8 |
| R$\longrightarrow$U | 36.0 |
| | |
| Out-of Province destination | 32.1% |

Source: Abstracted from Chan (2000).

Table 2 covers this short distance aspect, as well as an overview of temporary migration. Of the 100m floating population in 1998, only 62.4m had been out of residence for over 6 months. Of these (based on 1995 survey results), 41% moved just within their home county. So in 1998 only about 37m people had been living outside their official county of residence for more than 6 months. For these, what about rural versus urban destinations?

Based on flows for 1990-1995, for migrants moving for over 6 months, 40% of moves involve urban residents and 60% rural. For these 60% rural, 60% go to cities, as opposed to other rural areas. Finally for all movers with $6^+$ months stay, only 32% move outside of province. If we apply these numbers to 1998 and assume urban and rural movers have equal out of province propensities, in 1998 of the 62.4m temporary migrants, only 12m were rural folks moving out of province (62.4m * 32 *.6). Of the 12m temporary long distance rural migrants only some portion (60% suggested in Table 2) go to cities.

Whatever the exact numbers and the fact that we are past 1998, the analysis suggests that the permanent urban populations are only modestly supplemented with rural migrants on a nationwide basis. Table 2B suggests of the 62.4 temporary migrants, under 15m (62.4 *.36) involve rural-urban migration, both within and outside provinces. Even if we triple that number to adjust for increased migration and to add in some of the floating population staying less than 6 months, that still means only 10% of the official 450m urban residents are temporary migrants from rural areas.

*5.2.3. Spatial Discrimination and The Coast Versus The Hinterlands*

China has subsidized FDI (through tax breaks) in prefecture level cities and encouraged FDI and trade development in certain coastal cities, as part of a general program emphasizing coastal development, over hinterland development. The question is whether the FDI policy is efficient. On the subsidization question, the argument is that FDI brings in technology transfer, as well as creating job opportunities for low cost Chinese labor. The counter-argument is that FDI is not particularly high tech, compared even to more sophisticated domestic production, and FDI may discourage, or divert funds from local R&D. The evidence is not conclusive. For example Au and Henderson (2002) find that, ceteris paribus (same total capital to labor ratio) that cities with a one-standard deviation higher FDI/capital ratio have 8% higher output per worker. In Henderson (2002), FDI also enhances city growth rates. And in Fujita and Hu (2001) as noted earlier, FDI is associated with coastal agglomeration.

Assessing the issue of the efficiency of coastal versus hinterland development is less straightforward. On FDI, in Au and Henderson (2002), there is no evidence that FDI interacts with distance to the coast or city size – returns to FDI occur in the same degree for all cities regardless of size or location. But there is a more general question of coastal versus hinterland development. The Rappaport and Sacks (2001) story is that hinterlands are inherently inferior locations for economic development, compared to coastal locations. Démurger, Sacks, Woo, et al (2002) amend the story for China to argue that favored provinces tend to be coastal provinces so that the faster growth of coastal provinces is explained by a combination of policy-bias and inherent advantage.

A limitation in the analysis of coastal advantage is the failure to control for market potential of cities, a control fundamental in the analysis of economic geography (Overman, Redding and Venables (2001)). Statistically the issue is that in many countries (e.g. USA), historically populations have

agglomerated on coasts (including in Rappaport and Sacks for the USA the Great Lakes). So access to the coast captures both greater domestic market potential effects, and pure coast effects. In Au and Henderson (2002), distance to the coast on its own in eq. (1) significantly reduces productivity. However introduction of market potential eliminates the effect of access to the coast, and produces large significant effects for market potential. Similarly in Henderson (2002) access to the coast is not associated with higher growth per se, once FDI and market potential differentials are accounted for. If we consider Sichuan in Western China, its 100m residents have enormous market potential. With proper modern highway links to the coast, it in some sense will become "coast", with easier access to the coast. While coastal provinces still have better access to international markets, Sichuan may be domestically competitive, relatively specializing in domestic products. Its "disadvantage" may reflect policy disadvantage in terms of transport development, FDI, and loosening of planning constraints, more than an inherent disadvantage of hinterland location.

## REFERENCES


Abdel-Rahman, H., 1996, When do cities specialize in production. *Regional Science and Urban Economics* **26**, 1-22.

Abdel-Rahman, H., 2000, City systems: General equilibrium approaches. In: J-M Huriot and J-F Thisse (eds.), *Economics of Cities: Theoretical Perspectives*, Cambridge University Press, 109-37.

Abdel-Rahman, H. and M. Fujita, 1990, Product variety, marshallan externalities, and city sizes. *Journal of Regional Science* **30**, 165-85.

Abdel-Rahman, H. and P. Wang, 1997, Social welfare and income inequality in a system of cities. *Journal of Urban Economics* **41**, 462-83.

Ades, A.F. and E.L. Glaeser, 1995, Trade and circuses: Explaining urban giants. *Quarterly Journal of Economics* **110**, 195-227.

Anas, A. and K. Xiong, The formation and growth of specialized cities. *Mimeo.* State University of New York at Buffalo.

Arthur, 1990, Silicon valley locational clusters: When do increasing returns to scale imply monopoly. *Mathematical Social Sciences* **19**, 235-51.

Au, C.C. and J.V. Henderson, 2002, How migration restrictions limit agglomeration and productivity in China. *NBER Working Paper* No. 8707.

Baldwin, R.E., 2001, Core-Periphery model with forward-looking expectations. *Regional Science and Urban Economics* **31**, 21-49.

Baldwin, R.E. and R. Forslid, 2000, The Core-Periphery model and endogenous growth: Stabilizing and de-stabilizing integration. *Economica* **67**, 307-42.

Barro, R. and X. Sala-i-Martin, 1991, Convergence scross states and regions. *Brookings Papers on Economic Activity* **1**, 107-82.

Barro R. and X. Sala-i-Martin, 1992, Regional growth and migration: A Japan-USA comparison. *Journal of Japanese and International Economics* **6**, 312-46.

Bayer, P. and C. Timmins, 2001, Identifying social interactions in endogenous sorting models. Yale University mimeo.

Becker, G., E. Mills, and J.G. Williamson, 1992, *Indian Urbanization and Economic Growth since 1960*. Johns Hopkins Press.

Becker, G. and K. Murphy, 1992, The division of labor, coordination cost, and knowledge. *Quarterly Journal of Economics* **107**, 1137-60.

Becker, R. and J.V. Henderson, 2000, Intra-industry specialization and urban development. In: J-M Huriot and J-F Thisse (eds.), *Economics of Cities*, Cambridge University Press, 138-66.

Beeson, P.E., D.N. DeJong, and W. Troeskan, 2001, Population growth in US counties, 1840-1990. *Regional Science and Urban Economics* **31**, 669-700.

Benabou, R., 1993, Workings of a city: Location, education, and production. *Quarterly Journal of Economics* **108**, 619-52.

Bergsman, J., P. Greenston, and R. Healy, 1972, The agglomeration process in urban growth. *Urban Studies* **9**, 263-88.

Black, D., 2000, Local knowledge spillovers and inequality. *Mimeo*. University of California Irvine.

Black, D. and J.V. Henderson, 1999, A theory of urban growth. *Journal of Political Economy* **107**, 252-84.

Black, D. and J.V. Henderson, 2002, Urban evolution in the USA. *Journal of Economic Geography*, forthcoming.

Cai, Fang, 2000, *The Mobile Population Problem in China*. Henan People's Publishing House: Zhengzhou.

Carleton, D., 1983, The location and employment choices of new firms: An economic model with discrete and continuous endogenous variables. *Review of Economics and Statistics* **65**, 440-49.

Chan, K.W., 1994, *Cities With Invisible Walls*. Oxford University Press: Hong Kong.

Chan, K.W., 2000, Internal migration in China: Trends, determiation, and scenarios. University of Washington. Report prepared for World Bank (April).

Cheung, Peter T.Y., J.H. Chung, and Z. Lin (eds.), 1998, *Provincial Strategies of Economic Reform in Post-Mao China: Leadership, Politics, and Implementation*. Armonk, N.Y.: M.E. Sharpe.

Chipman, J.S., 1970, External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* **85**, 347-85.

Ciccone A., and R.E. Hall, 1996, Productivity and the density of economic activity. *American Economic Review* **86**, 54-70.

Clark, J.S. and J.C. Stabler, 1991, Gibrat's law and the growth of Canadian cities. *Urban Studies* **28**, 635-39.

Davis, James, 2000, Headquarter service and factory urban specialization with transport costs. Brown University.

Davis, S., J. Haltiwanger, and S. Schuh, 1996, Job Creation and Destruction. MIT Press.

Davis, J. and J.V. Henderson, 2001, Evidence on the Political Economy of the Urbanization Process. *Mimeo. Journal of Urban Economics*. Forthcoming.

Démurger, S., J.D. Sacks, W.T. Woo, S. Bao, G. Chang, and A. Mellinger, 2001, Geography, economic policy, and regional development. *Asian Economic Papers* **1**, forthcoming.

Dixit A. and J. Stiglitz, 1977, Monopolistic competition and optimum product diversity. *American Economic Review* **67**, 297-308.

Dobkins, L.H. and Y.M. Ioannides, 2001, Spatial interactions among U.S. cities: 1900-1990. *Regional Science and Urban Economics* **31**, 701-32.

Duranton, G., 2002, City size distribution as a consequence of the growth process. LSE *Mimeo*.

Duranton, G. and D. Puga, 2001, Nursery cities: Urban diversity process innovation, and the life cycle of products. *American Economic Review* **91**, 1454-77.

Durlauf, S.N., 1996, A theory of persistent income inequality. *Journal of Economic Growth* **1**, 75-93.

Eaton, J. and Z. Eckstein, 1997, Cities and growth: Evidence from France and Japan. *Regional Science and Urban Economics* **27**, 443-74.

Ellison, G. and E. Glaeser, 1999, The geographic concentration of US manufacturing: A dartboard approach. *Journal of Political Economy* **105**, 889-927.

Ellison, G. and E. Glaeser, 1999, The geographic concentration of industry: Does natural advantage explain agglomeration. *American Economic Association Papers and Proceedings* **89**, 311-16.

Fay, M. and C. Opal, 1999, Urbanization without growth: Understanding an African phenomenon. *Mimeo*. World Bank.

Flatters, F., J.V. Henderson, and P. Mieszkowski, 1974, Public goods, efficiency, and regional fiscal equalization. *Journal of Public Economics* **3**, 99-112.

Fujita, M. and H. Ogawa, 1982, Multiple equilibria and structural transition of non-monocentric configurations. *Regional Science and Urban Economics* **12**, 161-96.

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

Fujita, M. and T. Ishii, 1994, Global location behavior and organization dynamics of Japanese electronic firms and their impact on regional economies. Paper presented for Prince Bertil Symposium on the Dynanmic Firm, Stockholm.

Fujita, M. and J-F Thisse, 2000, The formation of economic agglomerations. In: J-M Huriot and J-F Thisse (eds.) *Economies of Cities*, NY, Cambridge University Press.

Fujita, M. and J-F Thisse, 2001, Agglomeration and growth with migration and knowledge externalities. *Kyoto University Institute for Economic Research* WP #531.

Fujita, M. and D. Hu, 2001, Regional disparity in China 1995-1994: The effects of globalization and economic liberalization. *The Annals of Regional Science* **35**, 3-37.

Fujita, M. and J-F Thisse, 2002, *Economics of Agglomeration*. Cambridge University Press.

Gabaix, X., 1999a, Zipf's law and the growth of cities. *American Economic Association and Proceedings* **89**, 129-32.

Gabaix, X., 1999b, Zipf's law for cities: An explanation. *Quarterly Journal of Economics* **114**, 739-67.

Gallup, J.L., J.D. Sacks, and A. Mellinger, 1999, Geography and economic development. *International Regional Science Review* **22**, 179-232.

Galor, O. and J. Zeira, 1993, Income distribution and macro economics. *Review of Economic Studies* **60**, 35-52.

Glaeser, E. H. Khalil, J. Scheinkman, and A. Shleifer, 1992, Growth in cities. *Journal of Political Economy* **100**, 1126-52.

Glaeser, E., J. Scheinkman, and A. Schelifer, 1995, Economic growth in a cross-section of cities. *Journal of Monetary Economics* **36**, 117-34.

Grossman, G. and E. Helpman, 1991, Quality ladders in the theory of growth. *Review of Economic Studies* **58**, 43-61.

Hanson, G., 1996, Localization economies, vertical organization, and trade. *American Economic Review* **86**, 1266-75.

Hanson, G., 2000, Market potential, increasing returns, and geographic concentration. *Mimeo*. University of Michigan (November).

Harris, J. and M. Todaro, 1970, Migration, unemployment, and development: A two sector analysis. *American Economic Review* **40**, 126-42.

Helpman, E.,1998, The size of regions. In: D. Pines, E. Sadka and I. Zilcha (eds.), *Topics in Public Economics: Theoretical and Applied Analysis*, Cambridge University Press, 33-54.

Helsley, R. and W. Strange, 1990, Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* **20**, 189-212.

Henderson, J.V., 1974, The sizes and types of cities. *American Economic Review* **61**, 640-56.

Henderson, J.V., 1986, Efficiency of resource usage and city size. *Journal of Urban Economics* **19**, 47-70.

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

Henderson, J.V., 1997, Medium size cities. *Regional Science and Urban Economics* **27**, 583-612.

Henderson, J.V., 2002a, The urbanization process and economic growth: The so-what question. *Journal of Economic Growth*. Forthcoming.

Henderson, J.V., 2002b, Marshall's scale economies. *Journal of Urban Economics*. Forthcoming.

Henderson, J.V., 2002c, City growth, worldwide 1960-2000. *Mimeo*. Brown University.

Henderson, J.V. and R. Becker, 2001, Political economy of city sizes and formation. *Journal of Urban Economics* **48**, 453-84.

Henderson, J.V. and Y. Ioannides, 1981, Aspects of growth in a system of cities. *Journal of Urban Economics* **10**, 117-39.

Henderson, J.V., A. Kuncoro, and M. Turner, 1995, Industrial development in cities. *Journal of Political Economy* **103**, 167-90.

Henderson, J.V. and A. Kuncoro, 1996, Industrial centralization in Indonesia. *World Bank Economic Review* **10**, 513-40.

Henderson, J.V., A. Kuncoro, and P. Nasution, 1996, Dynamic development in Jabotabek. *Indonesian Bulletin of Economic Studies* **32**, 71-96.

Henderson, J.V., T. Lee, and J.Y. Lee, 2001, Scale externalities in Korea. *Journal of Urban Economics* **49**, 479-504.

Hochman, O., 1977, A two factor three sector model of an economy with cities. *Mimeo*.

Holmes, T., 1999, Localization of industry and vertical disintegration. *Review of Economics and Statistics* **81**, 314-25.

Holmes, T. and J.J. Stevens, 2002, Geographic concentration and establishment scale. *Review of Economics and Statistics*, forthcoming.

Hoover, E.M., 1948, *The Location of Economic Activity*. NY: McGraw-Hill.

Ioannides, Y.M. and H.G. Overman, 2001, Zipf's law for cities: An empirical examination. *Mimeo*. Tufts University.

Jacobs, J., 1969, *The Economy of Cities*. NY: Random House.

Jefferson, G. and I. Singhe, 1999, *Enterprise Reform in China: Ownership Transition and Performance*. Oxford University Press: New York.

Jin, J. and H-F. Zou, 2002, Soft budget constraint on local governments in China. *Mimeo*. World Bank.

Junius, K., 1999, Primacy and economic development: Bell shaped or parallel growth of cities. *Journal of Economic Development* **24(1)**, 1-22.

Kanemoto, Y., 1980, *Theories of Urban Externalities*. Amsterdam: North-Holland.

Kelly, A.C. and J.G. Williamson, 1984, *What Drives Third World City Growth? A Dynamic General Equilibrium Approach.* Princeton University Press.

Kim, H.S., 1988, Optimal and equilibrium land use pattern in a City: A nonparametric approach. *Ph.D. Thesis*. Brown University.

Kim, S., 1995, Expansion of markets and the geographic distribution of economic activities: The trends in US manufacturing structure, 1860-1987. *Quarterly Journal of Economics* **95**, 881-908.

Kojima, R., 1996, Breakdown of China's policy of restricting population movement. *The Developing Economies* **34**, 370-401.

Kolko, J., 1999, Can I get some service here? Information technology service industries, and the future of cities. *Mimeo* Harvard University.

Krugman, P., 1991a, Increasing returns and economic geography. *Journal of Political Economy* **99**, 483-99.

Krugman, P., 1991b, *Geography and Trade*. MIT Press, Cambridge.

Krugman, P. and E, Livas, 1996, Trade policy and the third world metropolis. *Journal of Development Economics* **49**, 137-50.

Krugman, P. and A.J. Venables, 1995, Globalization and the inequality of nations. *Quarterly Journal of Economics* **110**, 857-80.

Lee, K.S., 1988, Infrastructure constraints on industrial growth in Thailand. *World Bank INURD Working Paper* No. 88-2.

Lee, K.S., 1989, *The Location of Jobs in a Developing Metropolis*. Oxford University Press.

Lee, T.C., 1997, Industry decentralization and regional specialization in Korean manufacturing. *Ph.D. Thesis*. Brown University .

Lewis, W.A., 1954, Economic development with unlimited supplies of labor. *Manchester School of Economic and Social Studies* **22**, 139-91.

Lucas, R.E., 1988, On the mechanics of economic development. *Journal of Monetary Economics* **12**, 3-42.

Lucas, R.E. and E. Rossi-Hansberg, 2001, On the internal structure of cities. *Econometrica*. Forthcoming.

Ma, L. and M. Fan, 1994, Urbanization from below: The growth of towns in Jiangsu, China. *Urban Studies* **31, 10**, 1625-1645.

Marshall, A., 1890, *Principles of Economics*. London: MacMillan.

Mills, E. and B. Hamilton, 1994, *Urban Economics*. Scott-Foresman.

Mohring, H., 1961, Land values and measurement of highway benefits. *Journal of Political Economy* **49**, 236-49.

Moretti, E., 1999, Worker education, externalities, and technology adoption: Evidence from plant-level production functions. University of California Berkeley Center for Labor Economics, WP No. 21.

Nakamura, R., 1985, Agglomeration economies in urban manufacturing industries: A case study of Japanese cities. *Journal of Urban Economics* **17**, 108-24.

Naughton, G., 2002, Provincial economic growth in China: Causes and consequences of regional differentiation. *Mimeo*. University of California, San Diego.

Neary, J.F., 2001, Of hype and hyperbolas: Introducing the new economic geography. *Journal of Economic Literature* **49**, 536-61.

Ono, Y., 2000, Outsourcing business service and the scope of local markets. CES Discussion Paper CES 00-14.

Overman, H., S. Redding, and A. Venables, 2001, The economic geography of trade, production, and income: A survey of empirics. *Mimeo*. London School of Economics .

Puga, D., 1999, The rise and fall of regional inequalities. *European Economic Review* **43**, 303-34.

Quah, D., 1993, Empirical cross section dynamics and economic growth. *European Economic Review* **37**, 426-34.

Rannis, G. and J. Fei, 1961, A theory of economic development. *American Economic Review* **51**, 533-65.

Rappaport, J. and D. Sacks, 2001, The US as a coastal nation. *Mimeo*, RWP 01-11. Federal Reserve Bank of Kansas City.

Rauch, J.E., 1993a, Productivity gains from geographical concentration of human capital: Evidence from the cities. *Journal of Urban Economics* **34**, 380-400.

Rauch, J.E., 1993b, Does history matter only when it matters a little? The case of city-industry location. *Quarterly Journal of Economics* **108**, 843-67.

Ray, D., 1998, *Development Economics*. Princeton: Princeton University Press.

Renaud, B., 1981, *National Urbanization Policy in Developing Countries*. Oxford University Press.

Rosen, K. and M. Resnick, 1980, The size distribution of cities: An examination of the Pareto law and primacy. *Journal of Urban Economics* **81**, 165-86.

Rosenthal, S. and W. Strange, 2002, Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, forthcoming.

Rossu-Hansberg, E., 2001, Optimal urban land use and zoning. *Mimeo*, November. University of Chicago.

Simon, H., 1995, On a class of skew distribution functions. *Biometrika* **44**, 425-40.

Stiglitz, J., 1977, The theory of local public goods. In: M.S. Feldstein and R.P. Inman (eds.), *The Economics of Public Services*, London: MacMillan, 273-334.

Sveikauskas, L., 1975, The productivity of cities. *Quarterly Journal of Economics* **89**, 393-413.

Sveikauskas, L., 1978, The Productivity of cities. U.S. Bureau of Labor Statistics, mimeo. J-94.

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

Tolley, G. J. Gardner, and P Graves, 1979, *Urban Growth Policy in a Market Economy*. NY: Academic Press.

Venables, A.J., 1996, Equilibrium locations of vertically linked industries. *International Economic Review* **37**, 341-59.

Wei, S-J. and Y. Wu, Globalization and inequality: Evidence from China. *CEPR discussion paper* No. 3088.

Wheaton, W. and H. Shishido, 1981, Urban concentration, agglomeration economies, and the level of economic development. *Economic Development and Cultural Change* **30**, 17-30.

Williamson, J., 1965, Regional inequality and the process of national development. *Economic Development and Cultural Change* **June**, 3-45.

World Bank, 2000, *Entering the 21st Century: World Development Report 1999/2000*. Oxford University Press.

Xiong, K., 1998, Intercity and intracity externalities in a system of cities: Equilibrium, transient dynamics, and welfare analysis. *Ph.D. Thesis*. State University of New York at Buffalo.

Yang, D.T. and M. Y. An, 1997, Human, capital entrepreneurship, and farm household earnings. *Mimeo*. Duke University.