Urbanization and City Growth[1]

J. Vernon Henderson

Hyoung Gun Wang

Brown University

April 19, 2004

Abstract. This paper models the urbanization process and how urbanization in a country is accommodated by increases in numbers versus population sizes of cities, in an endogenous growth context. It then estimates the key equations of the model describing growth in city numbers in a country and growth in individual city sizes. The paper shows that the relative size distribution of cities has been constant over time, with no increases in overall relative urban concentration. Growth in city numbers is a process well explained by national population growth and national technological progress which first affects rural-urban migration and second leads to increasing absolute sizes of existing cities. Political variables reflecting the degree of autonomy of local governments affect the process of city formation and the ability of an economy to generate new cities, both in theory and in practice.

As countries develop, they urbanize. Technological advances release labor from agriculture into production of urban goods, which are relatively income elastic in demand. This paper studies this shift of rural population to urban areas and the resulting spatial and economic transformation of countries. That transformation involves both the growth of existing cities and the creation of new cities, as some rural towns grow to become major urban centers. To study this process we assemble a worldwide data set for metropolitan areas over 100,000 from 1960-2000, and establish some basic facts about urban evolution. The paper provides an endogenous growth model of urbanization and the split between urbanization through existing city size growth versus development of new cities.

In the model, growth is based on human capital accumulation and related technological progress. Countries start from low human capital levels where all economic activity is in agriculture and at some critical value of human capital accumulation, urbanization starts. With urbanization, new cities don't simply pop-up. The growth in city numbers, in both theory and in empirical estimation, is driven by three underlying factors. First is national population growth and second is the extent of rural-urban migration as determined by economic growth. These two forces create a need for cities to accommodate increasing populations. The extent to which the urbanizing population is accommodated by growth of existing cities versus development of new cities, is largely determined by the third

[1] This research was supported by a grant from the Research Committee of the World Bank. Rupa Ranganathan of the World Bank prepared the original data and documentation.

factor: endogenous technological change, which drives the growth rate of existing cities. This is the first paper we know of that models this process, as well as estimates the key relationships. The empirical work starts by looking at the extent to which urbanization is accomplished through growth in city numbers versus city sizes and then estimates the equations describing growth in city numbers and in city sizes for countries.

In examining the data, one issue is whether rapid urbanization involves a transformation of existing urban systems in a country, so that they become dominated by mega-cities. The popular press (e.g. Economist, Time, Newsweek, National Geographic) and international agencies tend to assume that this is the case especially for developing countries, and then focus on the "dark side" of urbanization. The UN with its mega-city projects asks (UN, 1993) how bad "the negative factors associated with very large cities" need to get "before [it is in the] self-interest of those in control to encourage development of alternative centers" and discusses the perils of the development of “unbalanced urban hierarchies”. The World Development Report (2000, Chapter 7) similarly has a strong emphasis on the grim side of life in mega-cities. This focus is unfortunate, because the underlying premise is wrong. As the data will indicate, urbanization is not concentrated in mega-cities and urban hierarchies do not become increasingly unbalanced with urbanization. Much of urbanization occurs through the development of new cities and growth of smaller metro areas.

Given the underlying economic fundamentals driving urbanization, another important issue is the extent to which urbanization is affected by institutional settings and public policy. For example, do increasing democratization and corresponding increasing regional representation and fiscal decentralization within countries affect the urbanization process? We will argue and present evidence to show that increasing democratization increases the development of new cities. Similarly we ask what the effect is of trade policies promoting openness. Understanding the forces determining the development of new metro areas and the growth of existing ones and the role of public policies and institutions in the process is critical for developing countries undergoing rapid urbanization. To better plan infrastructure investments, to understand issues of "regional inequality", and to understand the impact of the myriad of policies with implications for the spatial allocation of resources, we need to understand the evolution of national economic geographies. For example, this would inform decisions about the extent to which infrastructure resources should be poured into expansionary policies for existing cities, as opposed to developing new cities. And knowing the effect of political-institutional changes on the form of urbanization and national economic geography would help in policy formulation more generally.

In Section 1 of the paper we present some basic facts about urbanization, growth in numbers versus sizes of cities, and the evolution of city size distributions over time. In Section 2, we present a simple model of the urbanization process; and in Section 3 we estimate the two key equations of the model, examining the determinants of growth in numbers versus sizes of cities. The focus is on growth in city numbers since that equation captures all the key elements of the process. City growth will be driven by both national forces and individual city circumstances.

1. Facts About Cities and Urbanization

It is well documented that the urbanization rate, defined as increases in the percent of the national population that is urbanized, is most rapid at low income levels, and then tails off as countries become "fully urbanized" (e.g., World Development Report, 2000 or Davis and Henderson, 2003). The definition of full urbanization and the rural/urban division under full urbanization varies across countries, so fully urbanized usually ranges from 65-85% of the population being urbanized. This paper looks at the numbers and sizes of cities comprising the portion of the urban population in larger cities.

Data: Our base sample is the metro areas of the world with populations over 100,000 every 10 years from 1960 to 2000. Data on metro area populations are from a variety of sources, cited in the Appendix B. The 100,000 cut-off is chosen for practical reasons—it is the cut-off employed by many countries. None have a higher cut-off and most do not provide consistent data over time on cities below 100,000. Even USA metro areas which in theory have a cut-off of 50,000, in practice only include comprehensively urban counties with over 85,000 (Black and Henderson, 2003). For the largest cities (typically metro areas over 750,000 in 1990) the UN World Urbanization Prospects data are utilized for 1960-2000. For smaller cities, populations for defined metro areas (especially post-1970) are available from the UN Demographic Year Book and country annual statistical yearbooks. Also utilized are www.citypopulation.de and www.World-Gazetteer.com for recent years. In some countries there is a mixture of populations based on metro area definitions for bigger cities or for later years and based on municipality jurisdictions for smaller cities or earlier years. As described in the documentation for this data (see Appendix B), we draw different samples. The best data year is 2000 and for that we can present a fairly comprehensive picture of world urbanization in cities over 100,000. The next best year (often pre- or near independence in many developing countries) is 1960 and we can give a fairly comprehensive comparison of 1960 and 2000.

When we move to decade-by-decade growth analysis, in examining growth in the numbers of cities in a country, we generally have good counts of cities. But some countries are missing urban data for particular decades and the country-city number panel is unbalanced. For data on individual city growth, the panel is unbalanced first because of arrival of new cities. But also some individual cities have one or two years of "bad data" (e.g., a 1970 number that is 1/4 or fourfold the 1960 number and a 1975 or 1980 number that is 30% higher than the 1960 number). If a country has over 25% of its cities with at least one bad observation in a given decade the whole country is dropped for that decade; otherwise data are utilized for the sample of cities with sequences of good data.[2] All calculated decade growth rates are consistent so if data are for 1962 to 1970, 1962 is extrapolated back to 1960 based on the annual 1962-1970 growth rate.[3]

1.1 Changes in the Numbers, Sizes, and Size Distribution of Cities Worldwide

a) The 2000 Distribution

Table 1 and Figure 1 show the size distribution of cities over 100,000 in the year 2000. The first observation is that although there is a wide spread in the size distribution, most cities are smaller. 84% of cities over 100,000 people have under 1m population in size and those cities account for 37.2% of the population of all cities over 100,000. In the medium size category, say 1-3m people, are 12% cities accounting for 28.9% of the population of all cities over 100,000. After that we are into bigger cities -- the 94 cities over 3m. Of these, if we draw the line for mega-cities at 12 million, there are 11 such cities accounting of 9.7% of the total population of cities over 100,000. If we draw the line at 10m, there are 19 such cities accounting for 14.4% of this population.

Wherever the line is drawn, despite popular notions noted in the introduction, most of the world's 1.8b people, who live in significant size cities (over 100,000) represented in Table 1, live outside mega-cities. And if we put the world's total urban population in 2000 at 2.9b (WDR, 2000), the 1.8b in our sample of cities over 100,000 account for only 62% of the total urban population of the world. The rest are in even smaller cities. Mega-cities over 12m only account for 6% of the total urban population of the world and under 3% of its total population. The vast

[2] As will be discussed below we utilize a measure of market potential, which describes a city's access to other cities' populations. This variable is calculated on as broad a base as possible including city populations under 100,000 when available by country and including cities with possibly flawed numbers (where bad numbers are replaced by interpolations) that are dropped otherwise from the sample.
[3] All that means in estimation is that estimates of the 10-year growth rate may be based on 8 or 9 (or 11 or 12) years.

majority of the world's total urban population lives in smaller and medium size cities under 2 to 3m. While this is not to say international agencies should ignore mega-cities, what is missing is a focus on developing institutions and policies for the smaller and medium size metro areas that most of the world lives in.

b) Comparing 1960 With 2000

The same size categories of cities are given in Table 2 for 1960. By comparing Tables 1 and 2, two facts emerge, with details for regions of the world given in Table A1. First, the number of cities over 100,000 has increased by 120% from 1960-2000. Second, the sizes of cities have increased. The average size has risen by 36%. The number of cities over 12m has increased from 1 to 11. The number of cities from 3-12m has increased from 24 to 83. More critically what we consider "large" or "medium" has changed substantially. In 1960 a city of 1m would be considered fairly large; today it would be at best a medium size city.

With development, there is increasing absolute spatial concentration of national populations as people urbanize and as cities grow in average size with technological change. Given that, we want to compare the shape of relative size distributions of cities over time to determine if there are changes in the degree of relative spatial concentration across metro areas -- the proportion of people living in relatively smaller versus relatively larger cities. In the overtime comparisons, there are two issues. First, absolute size distributions are shifting right; and second, our data refer just to the cities in the upper part of size distributions. In comparisons over time we want to try to look at the same portion of the size distribution. Given, as we will show, that the shape of relative size distributions is stable over time, based on the theory presented below (Section 2.5a), we proceed in a way that covers the same portion of the size distribution over time. First, to compare relative size distributions, we normalize city sizes by the average size of cities in the relevant sample in that time period (Eaton and Eckstein, 1997). Second, we alter the relevant sample in each period, raising the minimum size absolute cut-off point to keep the same relative size slice of pie.

To keep the same relative standard to be a city, we take the 1960 ratio of the minimum (100,000) to mean (500,340) for the 1960 versus 2000 comparison sample, and apply that ratio $(.1999)$ to all years (see Black and Henderson (2003)). That is, we draw the cut-off point to be a city in the sample for a particular year to be the first s cities (ordered by size) such that the $s+1$ city would be below that relative size; or we choose s such that in time t

$$\min\left\{s(t);\ N_{s+1}(t) / \sum_{i=1}^{s+1} N_i(t) / (s+1)\ \le\ .1999\right\}$$

where $N_i(t)$ is the population of city i in time t. For the year 2000, out of a possible 2,657 cities in the world over 100,000, this gives us 1,673 cities with an average size of 994,717 and a minimum absolute size city of 198,874.[4]

Table 3a gives relevant comparison data for the 1960 versus 2000 sample, using the relative cut-off point samples which for example then cover all 1,197 cities over 100,000 in 1960 and 1,673 cities over 198,874 in 2000. Numbers for the absolute cut-off point of 100,000 in all years are also given. In Table 3a, in this sample, mean city size increases by 99% from 1960 to 2000 and the numbers of cities increase by 40%. Urbanization is accommodated by both increasing numbers and sizes of relatively large cities. In Table 3b, we decompose the world (and region) growth in total population in metropolitan areas, into the share of that increase in total population found in new cities -- ones present in 2000, but not in 1960. For the world, about 26% of the increase is accommodated in new cities and the rest in growth of existing 1960 cities. Note that for developed or more fully urbanized countries, the share of new cities is much smaller; while for developing and especially former Soviet bloc countries, the role of creation of new cities in accommodating urban population growth is more important. In thinking about the magnitudes, it is important to recognize that new cities are small per se, but have rapid growth rates, so that over time they accommodate increasing proportions of urbanization (see below).

We do two main comparisons of city size distributions in 1960 with 2000. First, we compare plots of size distributions and then spatial Gini coefficients. Graphically we compare the 1960 and 2000 relative size distributions (city sizes relative to world mean city size). In Figures 2-4, we plot for 20 cells the share of number of world cities in each cell. Cells divide the line of length ln (100,000/mean city size 1960) to ln (max city size 1960/mean city size 1960) into 20 equal length cells, so that there is approximately an equal percent change in city size as we move up the size distribution. Figure 2 does the world comparison. Note the overlap of the 1960 and 2000 relative size distributions; there is no clear pattern of one relative size distribution differing from the other. In Figure 3 for developed countries and in Figure 4 for other countries, the strong overlap continues. In Figure 5 for comparison, we show the 1960 versus 2000 absolute size distributions (not normalized by mean population) for the world, to be compared to Figure 2. Note the strong rightward shift in the absolute size distribution.

Over the last forty years there has been little change in the relative size distribution of cities: we have roughly the same proportions of small, medium and relatively large size cities. One basic point is that cities are not converging to some common (growing) size. The spread of relative city sizes remains constant over time, suggesting

[4] Note we can't draw the line at the 100,009/678,218 = .1475 minimum to mean for 2000, since that would require relative cut-off

on-going roles for cities of all relative sizes. As we will argue below, this is consistent with systems of cities models, where there are different types of cities having different equilibrium sizes and producing or specializing in different types of manufacturing and service products. The second basic point is that urban hierarchies are not becoming increasingly unbalanced, with a greater role for mega-cities. The hierarchy is rock stable in terms of size distribution.

Another way to examine spatial concentration is to calculate spatial Gini coefficients, which give an overall measure of spatial inequality for the entire distribution. To calculate the Gini for a relevant sample we rank all cities from smallest to largest on the x-axis and on the y-axis we calculate their Lorenz curve -- the cumulative share of the total sample population. If all cities were of (almost) equal size, the plotted line would be (approximately) the 45° line. The Gini is the share of the area between the 45° line and the plotted curve, relative to the area below the 45° line. The greater the area, the "less equal" the size distribution, since smaller cities account for a smaller (cumulated) share of the sample population.[5]

In Table 4, we give Gini's for the 1960 versus 2000 for the world, developed countries, (former) Soviet bloc countries, and the rest (developing countries). Table 4 reinforces Figures 2-4. Gini coefficients for the world, developed countries, and less developed countries are very similar in 1960 and 2000. With rapid world economic growth, relative urban concentration, or spatial inequality has not increased over time. Moreover, if anything, spatial inequality is slightly higher in the developed world, than in the rest, or less developed world. And Soviet bloc countries, as is commonly perceived, have lower Gini's reflecting a more even spread of population across cities; and these Gini's have fallen over time. Table 4 also shows that while the number of cities above the minimum <u>relative</u> size has increased by 100% in the developing world, they have declined by 8% in the developed world. Relatively stagnant city sizes in the developed world means some cities in 1960 between 100,000 and 200,000 did not grow fast enough to meet the 199,000 cut-off in 2000.[6]

---

points in 1960 of around 50,000; and we don't have data for cities under 100,000 in 1960 on a consistent basis.

[5] If cities are ranked 0 to s where s is the largest size city, the x-axis goes from 0 to s and the y-axis from 0 to $\sum_{i=1}^{s} N_i(\bar{N}s)^{-1}$ (= 1) where $\bar{N}$ is average size. The area under the 45° line is $s/2$. The area between the 45° line and the Lorenz curve is $\frac{s+1}{2}$ - $\frac{1}{s\bar{N}} \sum_{i=1}^{s} (s-i+1)\, N_i$ and the Gini is this area divided by $s/2$.

[6] Using a relative cut-off point is also critical to the calculations. If the 2000 cut-off were 100,000, that would add many relatively small city sizes with tiny shares of world city population, increasing spatial inequality and the Gini. For example for 2000, the world Gini for cities over 100,000 is .625, compared to .564 when we use a relative cut-off plant.

We look also in Table 4 at how the 1960 versus 2000 degree of spatial inequality has changed for 7 large countries. The sample is cities in these countries above the world relative cut-off point. Again for these countries Gini's are similar in 1960 versus 2000. Russia and China have distinctly lower Gini's than other countries, and Brazil and Japan higher, again conforming to common perceptions.

c) Other Types of Comparisons

We note that in assessing overall spatial inequality many researchers approximate Zipf's Law and estimate rank size coefficients, by regressing ln (rank) on ln (city size) (where largest is rank 1). Lower slope coefficients, or flatter lines (for rank on the y -axis) imply greater inequality. Indeed for 15 countries in 1960 and 2000 for which we calculated Gini's and rank size coefficients, the two are strongly negatively correlated $(R^2 = .53)$. We footnote the 1960 and 2000 slope coefficients for these countries and note that the absolute values of the slope coefficient for the regression for the world in 1960 and 2000 are respectively 1.083 and 1.103.[7] However we rely on Gini's here for several reasons. They are not based on a specific size distribution (Pareto for rank size coefficients). OLS estimation of rank size coefficients and standard errors is problematical (Gabaix and Ioannides, 2003); and as a practical matter, individual country coefficients in our data are significantly influenced by country sample size per se. Recent work by Duranton (2002) and Rossi-Hansberg and Wright (2003) suggest significant deviations of actual distributions from Zipf's Law.

Finally, according to Gabaix and Ioannides (2003), the emergence of Zipf's Law is based on Gibrat's Law for the upper tail of the city size distribution (such as we are looking at), which says that city growth rates are a random walk and independent of base period size.[8] While in section 4 below, we more fully develop the error structure for city growth equations, here we simply report on a test for unit roots under the hypotheses that $\beta = 0$ in a model where $\Delta \ln \text{ city pop}_i(t) = d_i + \beta \ln \text{ city pop}_i(t-1) + \theta_t + \varepsilon_{it}$ for city i (Im, Pesaran and Shin, 1997, and Levin, Lin and Chu 2002). We strongly reject the hypothesis $\beta = 0,$ for this specification, as well as one where data is demeaned at the individual country level. We find a β of -.037 (t = -25.7) without city specific constants, a β of -.218 (t = -66.9) with city specific constants, and a β of -.880 with city specific constant terms and a time

[7] The 2000 (and 1960) absolute value of slope coefficients by country are Brazil .87 (.69), China 1.32 (1.10), India 1.07 (1.14), Indonesia .90 (.94), Mexico 1.04 (.96), Nigeria .98 (1.53), France .97 (.93), Germany .74 (.74), Italy .76 (.83), Japan 1.06 (1.18), Spain .98 (1.01), UK .83 (.82), USA 1.11 (1.08), Russia 1.34 (1.14), Ukraine 1.31 (1.03).

[8] In fact, from Gabaix (1999), it is necessary to place lower bounds on how small cities can get. Otherwise a log normal distribution arises, rather than a Pareto.

trend. These negative β's also show a mean reversion process where, on average, smaller cities grow faster than bigger cities in accommodating urbanization, especially once the overall trend in growth in sizes of all cities is accounted for.

Second, we note that our comparisons focus on the size distribution of larger metro areas. Many authors look at just the very extreme right tail, as represented by "urban primacy", the share of the largest (or 2-4 largest) city(ies), in national urban population. That literature consistently finds that as countries develop, urban primacy first tends to rise and then later decline. But the turning point is fairly early on in the development process and the pattern does not really appear in the data in recent years (Davis and Henderson, 2003). While primacy itself remains an issue, here we focus on the overall set of metro areas.

Finally one can look at the evolution of the size distribution by decade, city transitions through the size distribution, and emergence of any steady-state distributions. To do so requires us to have consistent information on cities in all decades; and our sample shrinks somewhat as we lose all cities in a few countries and some other cities as well. The resulting sample we will use in estimation in Section 3, so we note some features of it in Table 5. There the number of cities grows between decades by 17%, 20%, 14%, and 1.4% for 1960-70, 1970-1980, 1980-1990, and 1990-2000, while average sizes increase by 15%, 9.2%, 13% and 20%. In early years, the numbers grow at a faster rate than sizes, while in the 1990-2000 time period the reverse is the case, which will be consistent with empirical and theoretical results below. Growth through city numbers is greater when countries are less urbanized, as is the case in 1960, compared to 2000. Overall in this new sample for 1960-2000, the numbers of cities rise by 62% and average sizes by 70%. Again clearly both dimensions to urban growth are critical. For this sample the relative minimum to average size cut-off is .18. Minimum absolute sizes by construction grow at the same rate as average sizes, so for this sample by 2000, the minimum size is 169,682.

For these data we experimented with transition analysis, following Dobkins and Ioannides (2001), Eaton and Eckstein (1997), and Black and Henderson (2003), dividing the size distribution into 5 cells in 1960 containing approximately 35%, 30%, 15%, 10% and 10% of all cities starting from the bottom, with fixed relative cell cut-off points.[9] We calculated entry rates of new cities and show two typical patterns. With entry of new cities in the bottom cells, many existing cities get "pushed up" into higher cells (so transition probabilities of moving up are high relative

[9] Upper cut-off points are .32, .66, 1.16, 2.00, end open.

to moving down). Second, cities in the top cell have extremely low probabilities of moving down. The relatively big, with their enormous long-lived public capital stocks and scale externalities, stay big.

Although worldwide transition matrices are not stationary (unlike large country ones) and applying transition analysis worldwide, as opposed to by country is suspect, we did calculate predicted the 2000 distribution and the steady-state one.[10] Starting from the actual 1960 distribution by cell of .35, .30, .15, and .10, the 2000 actual is .35, .29, .14, .12, and .10 while the 2000 predicated and steady-state are respectively .33, .31, .14, .12, and .10 and .32, .30, .14, .12, and .12. This rock stability just mirrors Figures 2-4.

2. An Urbanization-Growth Model.

In this section, we develop a very simple growth model of an economy that is urbanizing, drawing on urban growth models (Lucas 1988), in particular adapting the model in Black and Henderson (1999) to incorporate a rural sector. The presentation here is limited to cover the essentials, looking to econometric implementation. More details of the model are in Henderson and Wang (2004). In the model there is a rural sector producing agricultural products and an urban sector comprised of an endogenous number of cities of endogenous size producing urban goods. We start by assuming that there is just one type of city, or one type of urban export good. The exposition starts by describing technology and the allocation of resources at any instant, noting that ultimately economic growth in the model will be driven by human capital accumulation.

2.1 Urban and Rural Sectors

In the rural sector output per worker is $D_a h_a^{\theta_a}$. D_a describes the level of technology to be made endogenous below and h_a the level of human capital per worker in the agricultural sector. This makes the rural sector constant returns in labor employment. This non-critical simplification avoids having to distribute rural land returns and we comment on its implications later. Rural wages are thus

$$W_a = D_a h_a^{\theta_a} \tag{1}$$

where agricultural products are the numeraire. We choose agricultural products as the numeraire so as to be able to comment on the transition from a pre-urban to urban society.

In the urban sector, the value of output per worker and hence wage is

$$W = x = pDh^{\theta}n^{\delta} \tag{2}$$

D is the level of technology in the urban sector (to be made endogenous), h is the per person level of human capital, θ is the private return to human capital, n is the size of the city where the worker lives and works, and p is the relative price of x. δ represents the degree of local urban agglomeration economies, from, say, local information spillovers. Then total city output is

$$X = Dh^{\theta}n^{1+\delta} \tag{3}$$

How do cities achieve equilibrium sizes? The δ scale parameter is a basis for local agglomeration, or a centripetal force. To have more than one city, we need a centrifugal force, which in the literature is vested in the housing and commuting sector. Cities occupy space in this framework.

All production in a city occurs at a point, the central business district (CBD). Surrounding the CBD is a circle of residences, where each resident lives on a lot of unit size and commutes to the CBD (and back) at a constant cost per unit distance of τ (paid in units of city output). This formulation of commuting costs being out-of-pocket is not innocent, as we will discuss momentarily. Equilibrium in the city land market is characterized by a rent gradient, declining linearly from the CBD to the city edge, where rents are zero. As city population expands, city spatial size, average commuting distances, and the rent gradient rise. Standard analysis gives us expressions for total city commuting costs and rents in terms of city population, where[11]

$$\text{total commuting costs} = pb\tau n^{3/2}, \tag{4}$$

$$\text{total land rents} = ½\ pb\tau n^{3/2}, \tag{5}$$

and

---

[10] For M the transition matrix, i entry rates overall of cities and Z the vector distribution of entrants across cells, under stationarity the steady state distribution is $[I-(1-i)M']^{-1}iZ$.

[11] An equilibrium in residential markets requires all residents (living on equal-sized lots) to spend the same amount on rent, $R(u)$, plus commuting costs, $p\tau u_1$, for any distance u for 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_1, rent plus commuting costs are $p\tau u_1$ since $R(u_1)=0$; elsewhere they are $R(u)+p\tau u$. Equating those at the city edge with those amounts elsewhere yields the rent gradient $R(u) = p\tau(u_1 - u)$. 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 $1/3 p\pi\tau u_1^3$. Total commuting costs are $\int_0^{u_1} 2\pi u(p\tau u)du = 2/3 p\pi\tau u_1^3$. Given a city population of n and lot sizes of one, $n = \pi u_1^2$ or $u_1 = \pi^{-½}n^{½}$. Substitution gives us eqs. (4) and (5).

$$b \equiv 2/3\pi^{-½}$$

Eq. (4) is a critical resource cost to the city, where average commuting costs $(p\tau b n^{1/2})$ rise with city size, with an elasticity of ½. That is the force limiting city sizes. Eq. (5) constitutes the gross rental income of the city developer.

City Formation, Size, and Income

That leads to the question of how cities form. For the moment we assume there are freely functioning competitive national land development markets. We model each city "as if" it is operated by a land developer, who collects urban land rents, offers inducements to firms to locate in the city, and specifies city population (although people are free to move in equilibrium and equilibria are free mobility proof). Nationally, there are an unexhausted number of potential identical sites on which cities can form, and each developer controls only one site. This is a traditional formulation (e.g., Hamilton, 1975, Scotchmer, 1986), but it is absolutely essential to understand that the resulting solutions can be obtained in other ways. In a static context, Henderson and Becker (2000) show that this solution (1) is the only coalition-proof equilibrium and (2) will occur also in a model with only "self-organization," where each existing city is governed by an autonomous local government. That is, autonomous local governments acting to maximize the welfare of residents by setting optimal subsidy and zoning or land development policies will duplicate the developer solution. Of course in taking the model to data, we recognize that the political and land market institutions that exist in a country like America to implement such solutions exist in much weaker forms in most developing countries. Below that will lead to a discussion of how the city formation solutions vary under different institutional arrangements and how changes in political systems and institutions affect city sizes and urbanization.

For a representative city, we specify the developer's optimization problem as a succession of contemporaneous profit maximization problems. A developer's profits are residential land rents (eq. (5)) less any transfer payments, T, to each worker-firm. The developer faces a free-migration constraint that each worker's net income (after rents and commuting costs are paid) equals the prevailing net real income from working available in national labor markets to workers in other cities, I (as we will see below I is net of capital income which city developers see as fixed). The developer chooses city population, n, and transfer payments, T, to maximize current profits, or

$$\max_{n,T} \prod = 1/2pb\tau n^{3/2} - Tn \qquad (6)$$
$$\text{subject to } W + T - 3/2\,pb\tau n^{1/2} = I,$$

where from (2) $W = Dpn^{\delta}h^{\theta}$. In the constraint the first term is private income per worker-firm and the third term is rent plus commuting costs per resident anywhere in the city from (4) and (5), so the total left-hand side is net real income earned in the city. Solving (6), substituting for T back into $\prod$, and setting $\prod = 0$ (through folk theorem free entry of developers/cities in national land markets) together yield basic urban results:

$$T = 1/2\,pb\tau n^{1/2} \qquad (7)$$

and

$$n = Q^{\frac{1}{\delta}}(Dh^{\theta}\tau^{-1})^{\frac{2}{1-2\delta}} \qquad (8)$$
$$Q \equiv (\delta 2b^{-1})^{\frac{2\delta}{1-2\delta}}$$

The first result in eq. (7) reflects the Henry George theorem (Flatters et al 1974; Stiglitz 1977). Total transfers to firms (Tn) equal total urban land rents $(1/2\,pb\tau n^{3/2})$, where the transfer per firm equals the gap between private marginal product and social marginal product (δW) of a city worker due to enhanced scale benefits when a worker-firm enters a city. Developers (or local governments) have the incentive to subsidize entry to their cities, internalizing the benefits of local scale externalities.

The second result in eq. (8) tells us that equilibrium city size is a function of scale and other parameters. It is also a function of human capital per worker and technology, D. Second-order conditions and eq. (7) reveal the parameter restriction $\delta < 1/2$ is necessary to have multiple cities in the economy. Given this restriction, eq. (8) shows that city sizes increase as (1) the scale elasticity, δ rises, (2) per worker human capital, h, or technology, D, improve and (3) τ, unit commuting costs decline. Size rises as D or h rise because output efficiency (enhanced by scale economies, or the term $2/(1-2\delta)$) improves relative to fixed commuting costs, τ. We note, for later econometric implementation, that, if we reformulated the model so commuting costs are only time costs, or $W_i = pDh^{\theta}n^{\delta}(1-\tau u_i)$ for a person living at u_i distance from the CBD, with a total labor supply of one, then h and D do not appear in eq. (8). City size would vary just with τ and δ. In a time cost formulation, while D and h raise productivity they correspondingly raise opportunity time costs of travel. We use the out-of-pocket cost formulation, in part because it is easier in terms of growth modeling (so each worker's optimal human capital level doesn't vary with distance from the city center[12]). But the formulation also emphasizes technology gains relative to commuting costs in enhancing city size, although an alternative gain from technological improvement would be a decline in τ. We will return to this point later in the empirical section.

From (6), (7), (8) and (3), we can solve for net working income and wages in a city as

$$I = (1-2\delta)W = pQ(1-2\delta)(Dh^{\theta}\tau^{-2\delta})^{\frac{1}{1-2\delta}} \tag{9}$$

Note in (9), with their scale economies, cities are fertile environments for technological improvement and human capital investments. The private return to human capital on income, θ, is augmented by urban scale economies, or divided by $1-2\delta<1$.

Preferences and Urbanization

In comparing the rural and urban sectors the key element driving urbanization is the specification of preferences. Instantaneous utility for any person in the economy is given by

$$V = (X_c + a_c^{\gamma})^{\alpha} \qquad \gamma,\ \alpha<1 \tag{10}$$

This specification means agricultural goods are income inelastic; so as incomes grow, there is a natural push for urbanization, as demand for urban relative to agricultural goods increases.. This quasi-linear specification also means that per person demand for agricultural products is simply

$$a_c = \gamma^{\frac{1}{1-\gamma}} p^{\frac{1}{1-\gamma}} \tag{11}$$

To close the model completely, we now need to turn to savings behavior, migration decisions, and the urban-rural allocation of human capital. To simplify the presentation here, we examine growth under exogenous savings behavior and note the generalization to endogenous savings behavior.

2.2 <u>The Human Capital Market, Migration, Savings and Urban Growth</u>

We assume, as is common, an explicit market in human capital; but then describe how outcomes can be mimicked through intra-family resource allocation decisions without a human capital market. For now each person in the economy has $\tilde{h}$ in human capital which can be used in production or loaned out. The capital market equalizes the private returns in agriculture and urban production so that the human capital rental rate $r = p\theta Dh^{\theta-1}n^{\delta} = \theta_a D_a h_a^{\theta_a -1}$. Substituting in (8) for n^{δ} yields

[12] With a time cost formulation, human capital need not vary with distance if human capital is disembodied, like physical capital,

$$p = \frac{\theta_a}{\theta} Q^{-1} \tau^{\frac{2\delta}{1-2\delta}} D_a D^{-\frac{1}{1-2\delta}} h_a^{\theta_a - 1} h^{1-\frac{\theta}{1-2\delta}} \tag{12}$$

Free migration requires urban incomes gross of human capital returns to equal the same for agriculture, or $\tilde{I} = I + r(\tilde{h} - h) = \tilde{W}_a = W_a + r(\tilde{h} - h_a)$. Note the addition of rentals $r(\tilde{h} - h)$ or $r(\tilde{h} - h_a)$ does not change the land developer's problem in (6), where the developer treats $h, \tilde{h}, I$ and $\tilde{I}$ all as fixed. Thus from (6) and (7) we have

$$W - W_a = p\tau b n^{1/2} + r(h - h_a) \tag{13}$$

Urban wages exceed rural wages, both to compensate for cost-of-living differentials and for urban workers to pay the differential opportunity costs of human capital if $h > h_a$. The relationship between h and h_a can be found by substituting in for $I, W_a, r,$ and p to get

$$h_a = h \frac{\theta_a}{\theta} \frac{(1-\theta-2\delta)}{1-\theta_a} \tag{14}$$

where we assume throughout that $\theta > \theta_a$, which is a sufficient condition here for $h > \tilde{h} > h_a$.

Full employment in the capital and labor market respectively require

$$n_a h_a + (N - n_a) h = \tilde{h} N \tag{15a}$$

$$n_a + mn = N. \tag{15b}$$

n_a is the agricultural labor force, m is the number of cities and N is the national labor force.

To close the model for an instantaneous equilibrium requires us to close one of the goods markets, in this case agriculture. That, in turn, requires a discussion of resources devoted to human capital production. Since we don't want to preclude a non-urban society but allow for human capital accumulation in agriculture, we could assume all human capital is produced with inputs of food. For this section instead we assume that human capital production in each sector is with goods from that sector where the same expenditure in each sector results in the same human capital additions. This is almost like devoting a fixed fraction of time (given a fixed savings rate) to human capital production. We assume savings at the rate s are from wage income net of rental costs or from $I - rh$ and $W_a - rh_a$, which magnitudes are equalized by migration.

---

and is chosen by the representative firm (not worker) in the CBD.

Thus in the food market, total production, $n_a D_a h_a^{\theta_a}$, equals consumption demand from (11), $N(p\gamma)^{\frac{1}{1-\gamma}}$ plus agricultural savings, or

$$n_a D_a h_a^{\theta_a} = N(p\gamma)^{\frac{1}{1-\gamma}} + sn_a(W_a - rh_a)$$

Substituting in for p from (12), h_a from (14), and for r yields the division of the national labor force into agriculture, which will be the key equation describing urbanization or the move out of agriculture:

$$\frac{n_a}{N} = Q_1 D_a^{\frac{\gamma}{1-\gamma}} D^{-\frac{1}{(1-2\delta)(1-\gamma)}} h^{\frac{\gamma\theta_a - \frac{\theta}{1-2\delta}}{1-\gamma}} \tag{16}$$

$$Q_1 \equiv \gamma^{\frac{1}{1-\gamma}}(1-s(1-\theta_a))^{-1}\left(\frac{\theta}{\theta_a}\left(\frac{(1-\theta_a)}{1-\theta-2\delta}\right)\right)^{\theta_a}\left[Q^{-1}\tau^{\frac{2\delta}{1-2\delta}}\left(\frac{1-\theta-2\delta}{1-\theta_a}\right)^{\theta_a-1}\left(\frac{\theta_a}{\theta}\right)^{\theta_a}\right]^{\frac{1}{1-\gamma}}$$

Finally we can link h and h_a to $\tilde{h}$ by substitution into (15), so

$$h\left[1 - Q_1\left(1 - \frac{\theta_a}{\theta}\left(\frac{1-\theta-2\delta}{1-\theta_a}\right)\right) D_a^{\frac{\gamma}{1-\gamma}} D^{-\frac{1}{(1-2\delta)(1-\gamma)}} h^{\frac{\gamma\theta_a - \frac{\theta}{1-2\delta}}{1-\gamma}}\right] = \tilde{h} \tag{17}$$

In (16) in order for there to be an urban sector (see below), $n_a / N < 1$. Therefore in (17) given $\theta > \theta_a$, the expression in the square brackets must be positive. For $\theta > \theta_a$, we then can show that h (and hence h_a) increases monotonically as $\tilde{h}$ increases.

2.3 <u>Urban Growth</u>.

We are now ready to derive the equations that underlie the empirical specification, the urban growth equations. For city size growth, from (8),

$$\frac{\dot{n}}{n} = \frac{2\theta}{1-2\delta}\frac{\dot{h}}{h} + \frac{2}{1-2\delta}\frac{\dot{D}}{D} \tag{18}$$

The main issue econometrically will be how to parameterize and measure changes in human capital, $\dot{h}/h$ and technology $\dot{D}/D$, as will as incorporate theoretically and empirically differences in individual city circumstances.

For changes in city numbers, we take the full employment constraint (15b) and differentiate to get

$\dot{m}/m = (N/mn)g - \dot{n}/n - (n_a/mn)\dot{n}_a/n_a$ where $g \equiv \dot{N}/N$ is treated as the exogenous national rate of population growth. We differentiate (16), where $\dot{n}_a/n_a = g + (\gamma\theta_a - \theta/(1-2\delta))/(1-\gamma)\dot{h}/h + \gamma/(1-\gamma)\dot{D}_a/D_a - (1-2\delta)^{-1}$ $(1-\gamma)^{-1}\dot{D}/D$. Substituting into $\dot{m}/m$ yields the key equation:

$$\frac{\dot{m}}{m} = g - \left(\frac{2\theta}{1-2\delta}\frac{\dot{h}}{h} + \frac{2}{1-2\delta}\frac{\dot{D}}{D}\right)$$

$$+\frac{\dfrac{\theta}{1-2\delta} - \gamma\theta_a}{1-\gamma}\left(\frac{n_a}{mn}\frac{\dot{h}}{h}\right) + \frac{1}{(1-2\delta)(1-\gamma)}\left(\frac{n_a}{mn}\frac{\dot{D}}{D}\right) \tag{19}$$

$$-\frac{\gamma}{1-\gamma}\left(\frac{n_a}{mn}\frac{\dot{D}_a}{D_a}\right)$$

On the RHS of (19), the first term tells us that city numbers grow at the national population growth, g, where that growth in city numbers is then adjusted for two considerations represented by the subsequent terms on the RHS. First, growth in city numbers is reduced by growth in individual city sizes, $\dot{n}/n$, represented by the second term which equals the RHS of equation (18). Secondly, city numbers also grow with rural-urban migration, or declines in n_a, as represented by the last three terms on the RHS. For the last three terms, first migration is speeded up by relatively greater urban human capital returns $(\theta/(1-2\delta) > \gamma\theta_a)$ which draw workers into the urban sector. Then migration is enhanced by technology improvements, D, in the urban sector and, finally, retarded by technology improvements, D_a, in the rural sector, which affect the relative price (eq. (12)) and demand for agricultural versus urban products. Note if agricultural land were a factor of production in our model, the main effect of that would be to retard migration since as n_a declines, the productivity of the remaining labor in agriculture rises.

Equations (18) and (19) ignore "lumpiness" problems in city size formation (the fact that m must be an integer), which disappear as m gets large. Econometrically, when we look at growth in city numbers, we will need to examine robustness of results to inclusion or exclusion of small economies.

2.4 Economic Growth

In a companion paper (Henderson and Wang (2003)), we delve into the details and dynamics of growth, in particular the move from a purely rural to an urbanizing economy. We explore steady-states with and without urbanization, how moving costs can give rise to multiple equilibria, OLG formulations, and the role of capital market imperfections, initial wealth endowments, and heterogeneity of workers. Here we summarize results for a simple benchmark case.

Exogenous Savings

For economic growth, in the benchmark case, we assume for technological change that $D_a = h_a^{\psi_a}$ and $D = h^{\psi}$. Human capital accumulation is a function of savings levels, where human capital does not depreciate. If H is total national human capital then $\dot{H} = s\left[(N - n_a)(I - rh) + n_a(W_a - rh_a)\right]$.[13] Given $\dot{H}$, we know $\dot{\tilde{h}}/\tilde{h} = \dot{H}/H - g$; and, given $I - rh = W_a - rh_a$, by substitution we have $\dot{\tilde{h}}/\tilde{h} =$ $-g + s\tilde{h}^{-1}(1-\theta_a)h^{\theta_a+\psi_a}\left(\frac{\theta_a}{\theta}\left(\frac{(1-\theta-2\delta)}{1-\theta_a}\right)\right)^{\theta_a+\psi_a}$. Substituting in (17) yields (given $D_a = h_a^{\psi_a}$ and $D = h^{\psi}$)

$$\dot{\tilde{h}}/\tilde{h} = -g + sQ_2 h^{\theta_a+\psi_a-1}\left[1 - Q_3 h^{\frac{\gamma(\theta_a+\psi_a)-\frac{\theta+\psi}{1-2\delta}}{1-\gamma}}\right]^{-1} \qquad (20)$$

$$Q_2 \equiv (1-\theta_a)\left(\frac{\theta_a}{\theta}\left(\frac{1-\theta-2\delta}{1-\theta_a}\right)\right)^{\theta_a+\psi_a}$$

$$Q_3 \equiv Q_1\left(1 - \frac{\theta_a}{\theta}\left(\frac{1-\theta-2\delta}{1-\theta_a}\right)\right)\left(\frac{\theta_a}{\theta}\left(\frac{1-\theta-2\delta}{1-\theta_a}\right)\right)^{\frac{\psi_a\gamma}{1-\gamma}} < Q_1$$

We assume just as $\theta > \theta_a$, $\theta + \psi > \theta_a + \psi_a$. In terms of growth, we have steady-state levels if $\theta_a + \psi_a < 1$ and we will approach steady-state growth if $\theta_a + \psi_a = 1$. But the analysis has several parts.

Eq. (16) for n_a / N defines a critical h and from (17) hence a critical $\tilde{h}$, $\tilde{h}_c$, below which $n_a / N = 1$ and we are in an agricultural world. Then $\dot{h}_a/h_a = \dot{\tilde{h}}/\tilde{h} = -g + s(1-\theta_a)h_a^{\psi_a+\theta_a-1}$. For $\psi_a + \theta_a < 1$, to have urbanization we need the steady-state level of $\tilde{h}$, $\tilde{h}_{ss}$, to exceed $\tilde{h}_c$. In that case human capital grows from some initial lower bound $\underline{h}$ in a pure agricultural world ("natural" human capital minimum) at a declining rate past $\tilde{h}_c$ to $\tilde{h}_{ss}$. In (20)

---

[13] It is easy to also augment $\dot{H}$ growth so the term in square rackets is multiplied by, say, $\tilde{h}^{\beta}$, as well as s with modest adjustment to the dynamics.

$\partial(\dot{\tilde{h}}/\tilde{h})/\partial h < 0$ for all $h \geq h_c$, so in the relevant region of parameter space a unique steady-state level exists. Note from (16) at h_c, the term in the square brackets in (20) is positive. If $\psi_a + \theta_a = 1$, we approach a steady-state growth rate. As h grows, $\dot{\tilde{h}}/\tilde{h}$ declines; but as h gets very large $\dot{\tilde{h}}/\tilde{h} \to -g + sQ_2$ given the term in the square brackets approaches one.

Extensions

Here we describe the basic approach with endogenous savings and no market for human capital. Following Black and Henderson (1999), consider a dynastic family model. A dynastic family maximizes $\int_0^\infty (X + a^\gamma)^\alpha \, e^{-(\rho-g)t} dt \quad \rho > g$ where ρ is the rate of discount and family size at time t is e^{gt}. Families allocate a proportion z of their members to cities and $(1-z)$ to agriculture. The human capital "market" involves intra-family transfers (urban remittances to agriculture), where investments and incomes differ between urban and rural areas. In equilibrium the z for each family will equal nm/N for the nation. The first constraint faced by the family, $H - ze^{gt}h - (1-z)\, e^{gt}\, h_a = 0$, allocates the family stock of human capital among members in the two sectors. The second is the equation of motion where net investment equals family income $e^{gt}\,(zI + (1-z)W_a)$ less family consumption (equal per person, which requires within family cross-sector transfers) so $\dot{H} = ze^{gt}I + (1-z)e^{gt}\, W_a - pX_c e^{gt} - a_c e^{gt}$. Human capital is produced from, say, just urban goods. For this model the solution to the city size growth equation is the same as above and the $\dot{m}/m$ equation has the same form as above, based on a redefined Q_1 in (16). In characterizing the form to $\dot{\tilde{h}}/\tilde{h}$ the problem now is that the time paths for $\dot{a}_c/a_c$ and $(dpX_c/dt)/pX_c$ are messy and $\dot{\tilde{h}}/\tilde{h}$ does not have a closed form expression. But the conditions to have solutions with steady-state levels $(\theta_a + \psi_a < 1)$ versus steady state growth are the same.

The paper assumes free migration from rural to urban areas. With exogenous savings, it is relatively easy to introduce migration costs, since implicitly consumers are naïve, or not forward-looking. So we could say that a fraction c of rural wages must be spent to migrate, so the free migration condition requires $W_a(1+c) + r(\tilde{h} - h_a) = I + r(\tilde{h} - h)$. While that affects the equilibrium n_a/N, it has no direct effect on growth rates in city sizes $\dot{n}/n$, and in numbers $\dot{m}/m$, but only an indirect effect through its impact on $\dot{\tilde{h}}/\tilde{h}$ and the analysis of human capital growth paths (see Henderson and Wang, 2003).[14]

2.5 <u>Extensions in Analyzing Urban Growth</u>

Econometric implementation in the real world requires a variety of modifications to the model. Here we outline the key ones to implement the country level equation describing growth in city numbers.

a) Multiple Types of Cities

Cities in an economy differ, as we saw with the wide, stable relative size distribution noted in Tables 2-4, as opposed to just one type of city. A basic model of multiple types of cities involves urban specialization, where there are different types of cities specialized in production of different export products and having different sizes. See Henderson (1974) and Duranton and Puga (2002) for static versions, as well as Duranton (2002) and Rossi-Hansberg and Wright (2004), for growth versions with stochastic processes yielding equilibrium distributions close to actual distributions, and not dissimilar to Zipf's Law. It is sufficient generally to have specialization if, for each city type, scale externalities in (2) are based on local <u>own</u> industry employment, so scale effects are $n_i^{\delta_i}$ (as opposed to n^{δ_i}).[15] Specialization minimizes local commuting/congestion costs relative to scale economy exploitation. Specialization of course is relative and can involve, for example, functional specialization where some cities are specialized in different production activities and others in having diverse business services produced as intermediate inputs, which are exported to production cities to combine with production and yield final goods for consumption.

Urban growth models with more than one type of city date to Henderson and Ioannides (1981), with an endogenous growth version in Black and Henderson (1999). In these models if there are v types of cities, under appropriate preferences, city growth is parallel so that $\frac{\dot{m}_1}{m_1} = \frac{\dot{m}_2}{m_2} = \ ... \ \frac{\dot{m}_v}{m_v}$ and $\frac{\dot{n}_1}{n_1} = \frac{\dot{n}_2}{n_2} = \ ... \ \frac{\dot{n}_v}{n_v}$. Each type of city

[14] A sophisticated treatment of migration costs requires forward-looking behavior and endogenous savings. Lucas (2002) has a single sector model of such migration. Migrants leave an old (rural) technology for producing the single economic good, to adopt a new (urban) technology. The growth of urban human capital improves efficiency of urban workers, and the growth rate of any person's urban human capital is an increasing function of the gap between that person's human capital and the highest skill urban worker (Eaton and Eckstein (1997)). Lucas has a neat story where new migrants to urban areas specialize in human capital investment until they catch-up to longstanding residents, with eventual depopulation of the rural (backward) sector. In this paper of course, agriculture does not remain backward and people always consume food, separate from urban goods.

[15] In fact specialization also occurs if the form in n^{δ_i} if δ_i's differ.

grows at the same rate in population size and in numbers. This is a basis for the relative cut on minimum city sizes we used in Figures 2-4 and Tables 3 and 5. If the cut in 1960 is at type i, by maintaining a cut at $n_i /(\sum_{j=1}^{n} n_j v_j / \sum_{j=1}^{n} v_j)$ in 2000 we are also cutting at i, with parallel growth. This formulation is also the basis for the empirical specification below which treats cities of different types as facing the same growth process, allowing us to estimate a single equation for growth in numbers of cities. The numbers of each city type grow at the same common rate, $\dot{m}/m$, and thus each and all grow at the national population growth rate adjusted for all the same forces in equation (19). That is, if we impose parallel growth and resolve (19) for multiple types of cities we get the exact same equation.

Relative to these urban growth models, ours adds in an agricultural sector, and urbanization. Adding in the agricultural sector, does not affect parallel growth in the urban sector if preferences have the form

$$V = (X_1^{\beta_1} X_2^{\beta_2} \ldots X_v^{\beta_v} + a^{\gamma})^{\alpha} \tag{21}$$

While p_i's may have different price paths but, with unitary price and income elasticities, relative city numbers and sizes will remain constant.

b) Institutions

A key issue in achieving the specified sizes and numbers of cities in the model concerns whether national land development markets operate to either create new cities or to transform villages into cities, in a timely fashion.[16] Becoming a city requires massive infrastructure investments, not specified in this model (cf. Henderson and Ioannides, (1981)), and mass relocation of industry and new housing construction. For the process to work as described above (Henderson and Becker, 2000), either land developers or autonomous local governments need to "set-up" and finance new city development, assembling the land for that purpose. Doing so requires well-defined property rights, functioning capital markets, and autonomous local governments. Lack of local government autonomy means infrastructure investment decisions are made centrally. In very restrictive environments, the

[16] A second issue may move city sizes and numbers in the opposite direction to the one discussed in the text. If cities do not subsidize for scale externalities, the point where the inverted $U-$shape I curve in Figure 6 achieves a maximum is shifted left, with the maximum occurring at a smaller city size. Specifically in eq. (7) $T=0$ and in eq. (8) $2^{\frac{2}{1-2\delta}}$ is replaced by $(4/3)^{\frac{2}{1-2\delta}}$. Given $1-2\delta<1$, the new n where I peaks is less than .44 of the old n. In this case, with active city development markets, economies would have smaller and more numerous cities; but growth rate equations are the same. However, a shift in regimes allowing economically and politically for internalization of externalities should lead in the short term to faster growth rates of cities and a reduction in the growth rate of numbers of cities. We don't see that effect empirically and have no way to measure whether locations of firms are subsidized or not.

creation of large cities may be de facto the decision of the central government. As such, the literature advances the notion that many countries favor national capitals with special resources, drawing migrants to those cities and retarding the development of alternative cities (Ades and Glaeser, 1995 and Davis and Henderson, 2003).

If national land markets are incomplete and there are no entrepreneurial forces encouraging population agglomerations, one model of city formation involves self-agglomeration, or self-organization. The problem is illustrated in Figure 6. There for the representative city an inverted $U-$ shape function of real income against city size is plotted, where real income from (2), (6), and (7), is $pDh^{\theta}n^{\delta} - p\tau bn^{1/2}$ (dividing all local land rents equally among residents). Given $\delta - 1/2 < 0,$ real income has a unique local and global maximum at $n = Q^{\frac{1}{\delta}}\ (Dh^{\theta}\tau^{-1})^{\frac{2}{1-2\delta}}$ and at that point real income, or $I = pQ(1-2\delta)(Dh^{\theta}\tau^{-2\delta})^{\frac{1}{1-2\delta}},$ the essence of eqs. (8) and (9). The “as if” land developer paradigm yields this size, although as noted earlier, usually the real life version of this involves city governments, which have the power to finance development, encourage in-migration, and then potentially limit city size.

What happens if there are no developers or city governments to help facilitate city formation. There, in a stark version in Figure 6, a new city only forms when existing cities (still illustrated by the representative city curve) become so oversized that any individual city resident would be better off leaving an existing city and forming her own single person city. This occurs at point $n_{\max}$ in Figure 6, for a representative city. In this "Malthusian version" of self-organizing formation of cities, existing cities are always growing towards $n_{\max}$, a "bifurcation" point where new cities form. This of course is an extreme representation. But it indicates a tendency for cities to be too large when it is difficult to do large scale developments, with institutions that promote the co-ordination of individual moving decisions. In many rapidly growing countries there are problems in assembling land, in obtaining development permission, and in financing urban infrastructure on the type of scale needed to encourage timely development of new metro areas.

Countries where cities lack local fiscal autonomy don’t have the ability to operate to facilitate the timely development of new cities. Thus, in such countries, we might expect to see in absolute terms fewer cities and ones of larger size, favored by the national government. While there could be fewer and larger cities in absolute number, the effect on growth rates of numbers of cities of a particular institutional setting is less clear. That is, as technology improves, ceteris paribus, $n_{\max}$ could shift at a faster or lower rate than n does in Figure 6, and with "bifurcation" the growth rate in numbers of cities is unclear. However the effect of a change in regime is not ambiguous. An increase in the degree of local fiscal autonomy resulting in greater ability of cities to govern themselves and restrict their sizes will result, ceteris paribus, in a short-term increase in the number of cities, as relative sizes of all cities move from away from the $n_{\max}$ point towards an n point. We will test for this effect.

3. Empirical Implementation and Results

In this Section we estimate versions of the equations for growth in numbers of cities (19) and sizes of cities (18). In both cases, we can't distinguish urban from rural human capital; but since $\dot{h}/h = \dot{h}_a/h_a$ and both increase monotonically with $\dot{\tilde{h}}/\tilde{h}$, the distinction is not important. The main concern is how to parameterize these changes and how to measure changes in technology, D and D_a. As a practical matter in estimation, $\dot{m}/m$ and $\dot{n}/n$ equations only respond directly to measures of human capital levels and not to measures of human capital changes, where our measure of human capital is the percent of adults completing high school from Barro and Lee (1996).[17] We believe this occurs because, as in Grossman and Helpman (1991) and Benhabib and Spiegel (1994), $\dot{D}_a/D_a$ and $\dot{D}/D$ (and potentially $\dot{\tau}/\tau$) are functions of human capital levels (e.g. $\dot{D}_a/D_a = f_a(\tilde{h}_a)$ and $\dot{D}/D = f(\tilde{h})$). So the focus in estimation becomes on how human capital levels spur innovation and technology growth, leading to increases in city size and rural-urban migration.

That still raises the question of why, apart from technology changes, changes in private human capital levels, $\dot{h}/h$, don't have measured effects in estimation. It could be that measures of changes, compared to levels, are noisier. It could also be that for city sizes, private human capital levels have little effect on city sizes. What drives city size growth is social human capital levels (knowledge spillovers) changing technology, and increasing D or lowering τ. Recall our discussion in the city size model, that if commuting costs are all time costs, D and h don't

[17] Of course human capital levels determine the growth of human capital, or changes in human capital. But signs reverse. As human capital levels rise, growth rates decline from (20). So higher human capital levels would work to reduce $\dot{\tilde{h}}/\tilde{h}$ in, for example (18), and reduce city sizes. Here increases in human capital levels raise city sizes, through we argue augmenting D.

affect city size. Then human capital levels driving technological improvements in τ would be the sole force increasing city sizes.

In presenting evidence we first start with the equation for numbers of cities. It has a simpler statistical specification and is more a direct test of the urbanization model, incorporating conditions of national population growth, rural-urban migration, and national technological progress. City size equations are also driven by changing individual city circumstances

3.1 <u>Numbers of Cities</u>

The object is to estimate a form of eq. (19) which shows how growth in city numbers is explained by the underlying forces of national population growth and national technological progress as it affects both city sizes and rural-urban migration. We also want to show that the econometric specification we use is a robust one. Given this, we then explore how the urbanization process is affected by politics.

Eq. (19) is implemented with data in 10- year intervals, measuring human capital levels by the percent of adults with high school education. In the equation, $\dot{D}/D = f(\tilde{h}) = \beta\tilde{h}$ (or $\delta\dot{\tau}/\tau = -\beta\tilde{h}$) and $\dot{D}_a/D_a = f_a(\tilde{h}) = \beta_a\tilde{h}$, allowing for differential knowledge spillover rates in urban versus agricultural areas. Under these assumptions, where $\dot{h}/h$ effects are not present, the equation for growth in numbers of cities in country j between t and $t-1$ in (19) becomes

$$\Delta \ln m_{jt} \equiv \ln m_{jt} - \ln m_{jt-1} = g_{jt} - \frac{2}{1-2\delta}\beta\tilde{h}_{jt-1} + \frac{\frac{\beta}{1-2\delta} - \gamma\beta_a}{1-\gamma}\left(\left(\frac{n_a}{mn}\right)_{jt-1}\tilde{h}_{jt-1}\right) + \varepsilon_{jt} \qquad \text{(19a)}$$

The first term on the RHS, as in (19), represents national population growth; the second represents growth of the representative city size from eq. (18); and the third, under our assumptions, collapses the last three terms of (19) into one, representing the effects of technology growth on rural-urban migration.

While in (19a) covariates representing technological change include just levels of human capital in the base $(t-1)$ period, $\tilde{h}_{t-1}$, we will experiment with adding to (19a) the terms in (19) representing changes in human capital, real income measures, and trade openness, where the last could encourage faster rates of technology adoption. g_{jt} is literally the national population growth rate which should have a coefficient of 1. $(n_a/mn)_{jt-1}$ will be that portion of the national population not in our sample, divided by the population in our cities. Details on data and variables are in Appendix B.

A key issue in estimation concerns error structure and endogeneity of covariates. A common way to proceed with panel data is to assume $\varepsilon_{jt} = u_j + d_t + \tilde{e}_{jt}$ where u_j is a fixed effect representing country time invariant unmeasured cultural-political, geographic, and institutional variables affecting city formation and growth in city numbers. These items could also affect covariates. d_t is a "world" technology level, and $\tilde{e}_{jt}$ represents a contemporaneous shock affecting growth in city numbers. However in our data, in the absence of fixed (or random) effects, error terms are not serially correlated. That is ε_{jt} is uncorrelated with ε_{jt-1} in (19a), so there is no persistence of errors over time. Correspondingly, there is no evidence of fixed or random effects in simple specifications. When we come to analyzing individual city size growth equations persistence in the error structure will play an important role in analysis, but here the evidence indicates a complete lack of persistence for numbers of cities in a country. Our final concern is endogeneity. Base period, $t-1$, variables may be pre-determined relative to $\tilde{e}_{jt}$, but growth rates in covariates between t and $t-1$ may not be. In particular, it could be that $(n_a / mn)_{t-1}$ is affected by the ε_{jt} that affects $\ln\ m_t - \ln\ m_{t-1}$. Moreover n_a / mn is approximated, or measured with error. The formulation that best satisfied specification tests is one where $(n_a / mn)_{t-1}$ is affected by shocks to $\ln\ m_t - \ln\ m_{t-1}$, while the rate of change of population growth and technological change are treated as exogenous.

a) Basic Results.

The basic results are in Table 6, which presents OLS and IV results. The presentation will focus on the IV results in column 1.[18] Results are second step GMM estimates allowing for within period heteroskedasticity (Arellano and Bond, 1991) with standard errors corrected for small sample bias (Windmeijer, 2000). Serial correlation of errors is rejected decisively (p-value of .96) and the Sargan test statistic is good (p-value of .503). Column 2 contains OLS results, which we will comment on below. A random effects and a fixed effects error structure are rejected in favor of OLS by standard Breusch-Pagan and $F-$ tests.

In the results in column (1), the coefficient on national population growth is 1.06, almost identical to the hypothesized coefficient of 1.0, and insignificantly different from it. Inclusion of a lagged dependent variable

---

[18] Instruments, over and above exogenous variables, are secondary educ * $n_a / nm\ (t-2)$, $n_a / nm\ (t-2)$, $\mathrm{n_a} / mn(t-3)$, and $\ln\ GDP_{pc}\ (t-2)$.

$(\ln\ m_{t-1})$, results in an insignificant coefficient of .014 in column (1). These two results are a nice affirmation of the steady-state growth model specification, where growth rates in city numbers have a stationary process (but absolute city numbers do not). Results where we exclude nine outlying observations in terms of small countries with high growth rates in a period are very similar to those in Table 6. Second for the full sample, plots of standardized residuals look very good, with the distribution conforming tightly to the expected normal in a QQ plot, except at the extreme upper tail.

Turning to the technological change variables, the negative coefficient on education represents the force of increasing city sizes reducing the need for more cities, while the positive coefficient on education interacted with n_a / nm represents the force of rural-urban migration pushing for more cities. Both forces are represented, as hypothesized. Note the coefficient of the covariate containing n_a / mn has an almost tenfold change in moving from OLS to IV estimation, with other coefficients unaffected. In net, the force of technological change depends on n_a / nm. As long as the relative rural population is not too big, so this ratio in column (1) is below 5.1 (.00304/.000599) which is roughly the mean in 1990, technological advances increasing existing city sizes reduce the need for more cities. At a low n_a / mn of, say, 0.5, a one-standard deviation increase (16) in education reduces the growth rate of new cities by .044, where the mean growth rate is .13. But at a higher n_a / mn of, say, 12 where rural-urban migration is very strong, a one-standard deviation increase in education increases the decade growth rate of numbers of cities by .067 as rural workers urbanize.[19] More new cities are needed to accommodate rural-urban migration.

Any formulations adding in, or using alone changes in human capital, the $\dot{h}/h$ and $(n_a / mn)\dot{h}/h$ terms in (19b) result in completely insignificant changes in human capital. As noted, this might suggest private returns to human capital as best we can measure them have a negligible effect on the city growth process.

b) Institutional Influences.

[19] The actual estimated coefficients can have implications in the model. Suppose we interpret (19a) literally -- that is h_{1t-1} does not proxy for $\dot{\tilde{h}}/\tilde{h}$ and $\dot{\tilde{h}}/\tilde{h}$ has no direct affect on the urban growth process. If the ratio of coefficients is r then γ=(r-2)/[r(1-2δ)($\beta_a / \beta) - 2$]. If $r > 2$, and here it equals 5.08, then for $\gamma < 1$, $\beta_a > \beta$. This may seem surprising. But the high value of knowledge spillovers in agriculture is well known, while knowledge spillovers in urban areas seem limited (Moretti (2003)).

This section examines the effect of institutions on the growth rate in city numbers. We experimented with various types of institutional measures. These include (1) whether a country is a (former) planned economy (Iron or Bamboo curtain), where migration restrictions are viewed as limiting city sizes, (2) British common law system (La Porta et al, 1998), (3) Kaufman et al's (2000) extent of rule of law and the IRIS measure of contract repudiation by governments as a proxies for well-defined property rights (Knack and Keefer, 1998), and (4) a measure of federalism for 48 larger countries in Davis and Henderson, (2003), as well as the Gastil (1978) measure of federalism.

In estimating our models, it became apparent that none of these variables in levels form have a consistent effect on growth in city numbers (or sizes). Nor from the discussion in Section 2.5b are such effects really expected, nor clear as to their direction. What we did argue in Section 2.5b is that a change in institutions would have at least a short-term effect on growth in city numbers. Earlier we argued that removing constraints on the degree of local autonomy could help the development of non-primate and non-capital cities, by increasing their ability to compete with primate and capital cities. Then a change in such institutions would lead to a jump in the number of cities. The only variable for which we have a long time series (back to 1960) for most countries in our sample which displays changes over time (as well as enormous cross-section variation) relates to democratization.

To measure the extent of democracy, we use the Polity IV index of democracy available through the University of Maryland website, which following Jaggers and Garr (1995) is the index of democracy minus the index of autocracy. It has values from -10 to +10, for which we have measures from 1960 on. Inherent in democracy is regional representation and increasing state and local autonomy; and measures of democracy and federalism are strongly correlated (Arzaghi and Henderson (2003)). We look at the effect of changes in democracy on growth in the numbers of cities. We treat such changes as exogenous. Results are in Table 7, column 1. There an increase in the democracy index leads to an increase in numbers of cities. The effect is strong. A one-standard deviation (7.9) increase in the democracy index increases the growth rate in numbers of cities by .047, from a mean of .13.[20] Democratization encourages new city development.

[20] We also experimented with examining for 2000, the absolute numbers of cities. The levels formulation doesn't directly come out of the model. We estimated an OLS equation of city numbers as a function of time varying covariates that affect growth in city numbers, plus land size of the country. Results other than country scale variables are weak. The democracy variable is positive, consistent with the growth results; but it is insignificant. We experimented also with 2SLS using instruments from the 1960's, but the small sample size renders results unstable (and subject to small sample bias).

We also look at openness, as a policy variable, inducing technology transfers and growth. In column (2) in Table 7 openness reduces significantly the growth in numbers of cities, on the basis that it leads to improved technology for cities per se and increased city sizes. Column (3) shows the results with both the openness and democratization variables included; coefficients are almost the same as in columns (1) and (2). In column (4) we allow for openness to affect also rural-urban migration which it should. The results are insignificant, but we could be asking too much from the data. We also looked at the effect of infrastructure investments in roads, as measured by kilometers of roads normalized by national land area. There for a smaller sample with overall weaker results, increased road investment does significantly increase the numbers of cities.[21]

3.2 Growth in City Sizes

In estimating the growth equation for city sizes, we turn to data on individual cities. As such, individual city growth is affected by city-specific conditions, as well as country conditions. We introduce the key time varying item— the prices a city receives for its output which affects the real income it can pay out. As we will detail momentarily, the prices of a city's output depends on economic geography, or the city's location relative to national markets. The effect of that on size is illustrated in Figure 6. Suppose most cities of a particular type are like the representative city of that type and face a similar price for their output. But a special city, with curve denoted by $\tilde{R}$, has a higher price and a shifted up real income curve in eq. (9). At any size, the special city can pay higher real incomes. This gap between potential income in the special city and other cities under free migration leads people to move into the special city until its size is $n_{\tilde{R}}$. Upton (1981) has a detailed treatment how sizes of cities (of the same type) differ when they face different circumstances, such as amenities or prices.

In empirical work, how do output prices vary by city location? If the country exports internationally from a single port at a fixed world price, the transport costs of shipping to the coast from different interior locations would affect the net received price of the city. That implies a world of heterogeneous geography where either "natural" or inhabitable urban sites vary randomly in location across the country and/or historical factors enter in determining locations of cities. All this is beyond a simple general equilibrium model, although stylized representations in static models exist (e.g. Davis, 2003).

[21] Estimation cuts the sample size from 216 observations for 75 countries to 144 observations for 63 countries. We estimate by GMM, with lagged road density values, as instruments for current log changes. While the road coefficient is positive and significant, other coefficients weaken with the loss in sample size.

From the trade literature, another way to introduce transport costs we utilize here, is to have iceberg transport costs for all urban exports in a differentiated products, monopolistic competition framework. Preferences would be respecified so in (21) $X_i^{\beta_i}$ is replaced by $\left(\sum_{j=1}^{m_i}\sum_{s=1}^{f_j}(X_i^{s_j})^{\rho}\right)^{\beta_i/\rho}$ where for type i cities, there are m_i of them, and any city j has f_j firms and varieties of product X_i (see Au and Henderson (2003) for a development of this model). Any variety is shipped from city j to all other cities in the country (and potentially all other countries), where cities have different sizes and locations. From trade theory (see Head and Mayer 2003, and Overman, Reading, and Venables, 2001, for a review), we can then obtain a measure of market potential, MP, for each city where the price facing any firm in city j (in a symmetrical equilibrium with respect to firms in city j) is

$$p_j = MP_j^{1-\rho}(x_j)^{-(1-\rho)}$$

for x_j the output of a firm in city j and MP_j, city $j's$ market potential, where

$$MP_j = \sum_k \frac{E_k\ I_k}{(1+\tau_{jk})^{\rho/(1-\rho)}} \qquad (22)$$

The sum is over all locations, E_k is disposable income at k, τ_{jk} is the shipping costs from j to k, and I_k is a price index. (For example $I_k = \sum_{\ell}(p_\ell f_\ell \tau_{k\ell})^{-\rho/(1-\rho)}$, for f_ℓ the number of firms in city ℓ, when there is only a single type of differentiated product in the economy). In next specifying an individual city growth equation, we adapt this equation.

The proposed estimating equation is

$$\Delta \ln n_{it} \equiv \ln n_{it} - \ln n_{it-1} = \ \alpha\Delta \ln\ (MP_{it}) + \frac{2}{1-2\delta}\beta\tilde{h}_{it-1} + \beta_t + v_{it} \qquad (18a)$$

In (18a) the second term on the RHS captures the same effect of technology growth on city sizes as in eq. (19a). As for city numbers, we formulate that $\dot{D}/D = \beta\tilde{h}.$ (and/or $\delta\,\dot{\tau}/\tau = -\beta\tilde{h}$). The first tem is changes in market potential which changes affect city received prices, real incomes, and sizes. The measure of market potential we use is

$$MP_{ij}(t) = \sum_{\substack{k\in j\\ i\neq k}}^{m_j(t)} \frac{n_{kj}(t)}{d_{ik}}$$

which is the measure for city i in country j. It equals the distance discounted sum of populations of all other cities, $m_j(t)$, in j at time t, excluding i. d_{ik} is the distance from city i to k in 100's of miles. In estimation we use $\ln(MP_{ij}(t))$. While this population measure is a crude representation of the extent of market demand for a city's products, compared to the proper definition of market potential as specified in eq. (22), it is viewed as generally performing well in trade econometrics (Head and Mayer, 2003). And, since we don't have the data to calculate (22), this seemed the best alternative to describe the potential demand for a city's product and the equilibrium prices it faces.

The main problem in implementing (18a) turned out to be estimation method. We believe we have come up with a formulation that resolves some puzzles in the literature about how to specify a city size growth model, in the face of strong persistence in individual growth rates. First, simple IV estimation of (18a) results in a failure of Sargan tests and an inability to reject serial correlation. One usual panel data strategy is to assume v_{it} has a city fixed effect component and then to difference (18a) to remove the fixed effect. Here differencing (18a) and instrumenting with lagged (predetermined) levels variables still results in second degree serial correlation and the specification fails Sargan tests, even though instruments are not weak. Alternatively, introducing a lagged dependent variable in (18a) doesn't eliminate serial correlation problems; and more particularly Sargan specification tests continue to fail.

Experimentation revealed that a simple AR (1) error structure solved these problems. We assume

$$v_{it} = \rho v_{it-1} + e_{it} \tag{23}$$

From (18a), by substituting for v_{it-1} from the equation for $\Delta \ln n_{it-1}$ in the equation for $\Delta \ln n_{it}$ (with v_{it} written as $\rho v_{it-1} + e_{it}$) gives the dynamic representation

$$\Delta \ln n_{it} = a_0 \Delta R_{it} + a_1 \tilde{h}_{it-1} + \rho \Delta \ln n_{it-1} - \rho a_0 \Delta R_{it-1} - \rho a_1 \tilde{h}_{it-2} + \beta_t^* + e_{it} \tag{24}$$

In (24) β_t^* represents the quasi-differenced time terms. We start by estimating

$$\Delta \ln n_{it} = B_0 \Delta R_{it} + B_1 \tilde{h}_{it-1} + B_2 \Delta \ln n_{it-1} + B_3 \Delta \ln R_{it-1} + B_4 \tilde{h}_{it-2} + B_t^* + e_{it} \tag{24a}$$

by instrumental variables by GMM as with (19a). For instruments, we simply use 1960 base period variables,[22] which are generally both strong instruments and improve Sargan values, relative to using recently lagged variables. Then we do a Wald test to check the specification to see if

$$\begin{aligned} B_0 B_2 + B_3 &= 0 \\ B_1 B_2 + B_4 &= 0 \end{aligned} \tag{25}$$

in (24a), as implied by (24). We then re-estimate the model, imposing the constraints in (25) by Chamberlain minimum distance methods (utilizing the covariance matrix from (24a) (e.g., Blundell and Bond, 1998a). Those are the constrained results we report (and they are almost identical to results for (24) estimated by non-linear 2SLS).

a. Basic Results

The basic results are given in Table 8, column 1. The top part of the column gives GMM results for the unconstrained model in (24a). The direct estimate of ρ or B_0 is .79. The relative market potential coefficients (B_3 / B_0) imply a ρ of .85, while those for high school (which are insignificant) also imply a ρ of .85. Given that, not surprisingly Wald tests on either the individual restrictions or on the set of two are far from rejecting the restrictions in (25). In the OLS version, column (2), the explanatory power of the model, for a growth rate equation is good, indicating a strong role for external factors to a city in explaining individual city growth rates. However relative to the model in the theory section, there is a strong role for persistence of unobservables in affecting city growth rates. That may reflect the fact that our model doesn't allow for "history", such as the role of non-malleable local physical infrastructure investments (cf., Henderson and Ioannides, 1981).

To improve efficiency, we examine coefficients or covariates for the constrained model. There ρ =.84. in the bottom panel of Table 8. In this dynamic model, the effects of schooling are very high. The percent high school in Table 8 has a coefficient of .0174, so one-standard deviation (16) increase in percent high school increases city growth by .28 (from a mean of .26). Note the coefficient on education in the bottom panel of Table 8 is larger than what is supposed to be the same coefficient from eq. (19a) in Table 6. That coefficient of the effect of education on city growth rates is indirectly inferred from the effect on growth in city numbers, while here it is a direct estimate. However, here in the unconstrained version of the model in top panel of Table 8 the coefficient is much smaller (.005), and similar to that in Table 6. For market potential, a one-standard deviation (.20) change in market potential

[22] 1960 instruments are % agriculture labor force, % high school, % agric lab. force * % high school, market potential, market potential * % agric lab. force, market potential * % high school, and city population.

raises city growth by .40 in the short term, a big impact as would be expected. There are only on-going effects, if changes in market potential continue indefinitely.

b. Institutional Effects

As for city numbers, for city sizes we look at the effect of institutions on city growth rates. Again in a growth context the only time varying institution for which we have measures is democracy. In the formulation in (24), changes in democracy do not have a significant effect on city size growth rates, although the sign patterns are "correct" (a negative effect of an increase in democracy on city growth rates). In general with this non-linear dynamic formulation, it is hard to identify effects beyond a parsimonious specification. Similarly for policy variables, openness and change in road density, results are insignificant.[23]

Finally, for the record, we also looked at city sizes in absolute terms in 2000, a specification which derives directly from eq. (8). We present OLS and 2SLS results in Table 9. Table 9 suggests technology variables (education and openness) have expected positive effects on city sizes. Market potential has a statistically weak effect in a levels formulation. Finally for political variables, capital cities in particular, given political favoritism, are much bigger than other cities (see Ades and Glaeser, 1995); and port cities are larger also. For democracy, earlier, growth in city numbers was enhanced by greater democracy and implied regional autonomy. Correspondingly, democracy works to reduce city sizes, presumably moving us away from $n_{\max}$ towards n in Figure 7. For planned economies, OLS results produce the expected negative sign, but the 2SLS results suggest no effect.

4. Conclusions

In this paper we developed a simple model of the urbanization process, and the split between development of new cities versus growth of existing cities. We show that the process, in particular city formation and the growth in numbers of cities can be explained in large part by a simple model and a few covariates -- national population growth, inferred technological change, and changes in institutions. For the last, increasing democratization, regional representation and local autonomy facilitate the formation of new cities. We show that there is no growing imbalance to urban hierarchies and no change to the relative size distribution of cities over time. We also show that much of urbanization is accommodated in smaller and medium size cities.

Table 1. World City Size Distribution, 2000

	count	mean	share[1)]
17,000,000 <= n2000	4	20,099,000	4.5
12,000,000 <= n2000 < 17,000,000	7	13,412,714	5.2
8,000,000 <= n2000 < 12,000,000	13	10,446,385	7.5
4,000,000 <= n2000 < 8,000,000	29	5,514,207	8.9
3,000,000 <= n2000 < 4,000,000	41	3,422,461	7.8
2,000,000 <= n2000 < 3,000,000	75	2,429,450	10.1
1,000,000 <= n2000 < 2,000,000	247	1,372,582	18.8
500,000 <= n2000 < 1,000,000	355	703,095	13.9
250,000 <= n2000 < 500,000	646	349,745	12.5
100,000 <= n2000 < 500,000	1,240	157,205	10.8
	2,657	678,218	100.0

1) a ratio of total population in the group to total population of cities with >=100,000

Table 2. World City Size Distribution, 1960

	count	mean	share[1)]
17,000,000 <= n1960	0	.	0.0
12,000,000 <= n1960 < 17,000,000	1	14,164,000	2.4
8,000,000 <= n1960 < 12,000,000	3	9,648,667	4.8
4,000,000 <= n1960 < 8,000,000	14	5,763,286	13.5
3,000,000 <= n1960 < 4,000,000	7	3,487,286	4.1
2,000,000 <= n1960 < 3,000,000	25	2,351,720	9.8
1,000,000 <= n1960 < 2,000,000	65	1,374,756	14.9
500,000 <= n1960 < 1,000,000	160	702,319	18.8
250,000 <= n1960 < 500,000	262	344,754	15.1
100,000 <= n1960 < 500,000	660	151,289	16.7
	1,197	500,340	100.0

1) a ratio of total population in the group to total population of cities with >=100,000

[23] For road density, the minimum distance coefficient is negative, as expected, but its $t-$statistic is just under one and the sample size drops to 2330 (846).

Table 3. Relative Size Distribution Comparisons: 1960 versus 2000

a) Counts and Sizes of Cities

cutoff[24]	variable	count	mean	Median	sd	min	Max
absolute	n1960	1,204	499,623	212,068	975,018	100,000	14,164,000
	n2000	2,684	673,317	264,743	1,462,224	100,009	26,444,000
relative	n1960	1,204	499,623	212,068	975,018	100,000	14,164,000
	n2000	1,673	995,639	465,161	1,785,332	199,355	26,444,000

a) Decomposition: Share of New Cities Since 1960 in 1960-2000 Growth of Total Population in Metro Areas.

cut-off	world	developed countries	Soviet bloc	all other countries
absolute	.31	.19	.47	.35
relative	.26	.12	.33	.29

Table 4. Spatial Inequality in 1960 versus 2000

	1960		2000	
	count	Gini coef.	Count	Gini coef.
World	1,197	0.5854	1,673	0.5642
Developed countries	523	0.6133	480	0.5825
Soviet bloc	179	0.5151	202	0.4465
All other countries	495	0.5660	991	0.5624
Brazil	24	.666	64	.654
China	108	.472	223	.425
India	95	.556	138	.582
Indonesia	22	.524	30	.614
Japan	106	.604	82	.656
USA	167	.577	197	.540
Russia	79	.538	91	.462

[24] absolute cutoff = 100,000; relative cutoff: minimum/mean size = .1999, given a 1960 minimum size of 100,000.

Table 5. Total Numbers of Cities and Sizes.

	1960	1970	1980	1990	2000
number of cities	969	1129	1353	1547	1568
mean size	556503	640874	699642	789348	943693
median size	252539	275749	304414	355660	423282
minimum size	100082	115181	126074	141896	169682

Table 6. Growth in City Numbers[a]
(standard errors in parentheses)

	(1)	(2)
	IV estimates	
	Level-growth	OLS
growth rate of national	1.06**	.963**
population (t)	(.143)	(.126)
% adults with secondary	-.00303**	-.00253**
educ ($t-1$)	(.00121)	(.00114)
% adults with sec. educ. *	.000599**	.0000643
(rural/urban pop.) ($t-1$)	(.000271)	(.000149)
time effects	yes	yes
N (countries)	216 [75]	216 [75]
R^2		.262
Sargan p-value	.507	

** Significant at 5% level. * Significant at 10% level. OLS errors are calculated allowing for country clustering.

[a] The mean and standard deviation of the national population growth rate, % adults with secondary education, and rural/urban population are .11 (.11), 26.6 [18.0], and 5.2 [6.9].

Table 7. Political and Policy Variables: Growth in City Numbers

	(1)	(2)	(3)	(4)
growth rate of national	1.11**	1.08**	1.20**	1.08**
pop (t)	(.182)	(.163)	(.210)	(.181)
% adults w/ secondary	-.00214*	-.00304**	-.00209**	-.00388**
education ($t-1$)	(.00115)	(.00104)	(.00105)	(.00176)
% adults w/ second. educ.	.00578**	.000663**	.000590**	.000914**
* (rural/urban pop.) ($t-1$)	(.000255)	(.000285)	(.000234)	(.000406)
change in democracy (t)	.00594**		.00607*	
	(.00302)		(.00326)	
openness ($t-1$)		-.00221**	-.00250**	-.00167
		(.00106)	(.00108)	(.00145)
openness * (rural/urban				-.000174
pop.) ($t-1$)				(.000271)
N (countries)	199 (74)	206 (75)	191 (74)	206 (75)
Sargan $p-$ value	.576	.397	.588	.362

Table 8. Growth Rates of City Sizes[25]

	(1)	(2)
	IV estimation	OLS
$\Delta \ln n_{ij}(t-1)$	.788**	.197**
	(.0726)	(.0417)
$\Delta \ln$ (market potential) $_{ij}(t)$	1.37**	.555**
	(.580)	(.0835)
$\Delta \ln$ (market potential) $_{ij}(t-1)$	-1.17**	.0497
	(.489)	(.0702)
percent high school $_j(t-1)$	.00510	-.000459
	(.00985)	(.000761)
percent high school $_j(t-2)$	-.00435	-.000141
	(.00823	(.000585)
time dummies	yes	yes
Sargan $p-$ value [R^2]	.847	[.403]
$\Delta \ln$ market potential (a_0)	2.01**	.732**
	(.425)	(.0818)
percent high school (a_1)	.0174**	-.000734
	(.0063)	(.000725)
ρ	.839**	.205
	(.0107)	
N (cities)	3394 (1180)	4232 (1629)

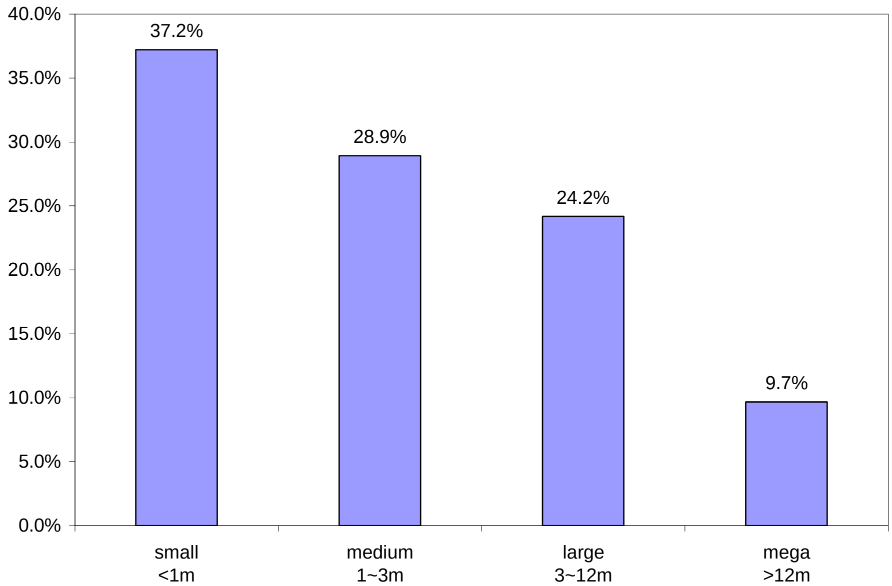
[25] OLS errors are clustered by country-year.

Table 9. Institutions and City Sizes[26]

	OLS	2SLS[27]
ln (market potential (2000))	-.0833	.0629
	(.0683)	(.115)
capital	2.28**	2.30**
	(.128)	(.167)
port	.275**	.253**
	(.0470)	(.067)
% adults w/ sec. educ. (1990)	.00835*	.0137**
	(.00456)	(.00338)
openness 1990	.00924**	.00553**
	(.00413)	(.0162)
ln(national pop (2000))	.263**	.340**
	(.0697)	(.0489)
ln(nat. land area)	.183**	.359**
	(.0477)	(.0623)
democracy (1990)	-.0130*	-.0318**
	(.00759)	(.0161)
planned economy	-.224*	.0110
	(.124)	(.149)
constant	8.15**	1.67
	(.753)	(2.76)
R^2	.369	
Sargan p – value		.405
N	1454	1237

[26] OLS errors are clustered by (57) countries. 2SLS are generalized 2SLS.
[27] Instruments are ln (nat. pop 1965), ln (land), ln (market potential 1960), ln (mark. pot. 1960) sq., % adults with sec. educ. (1965), planned economy, port, capital, democracy (1967), % workers in agric (1965), % adults with sec. educ (1965) * ln (market potential (1960)), and ln (distance from capital to nearest of Tokyo, NY or London).

Figure 1. Share by Size Category of World Population in Cities over 100,000 in the Year 2000



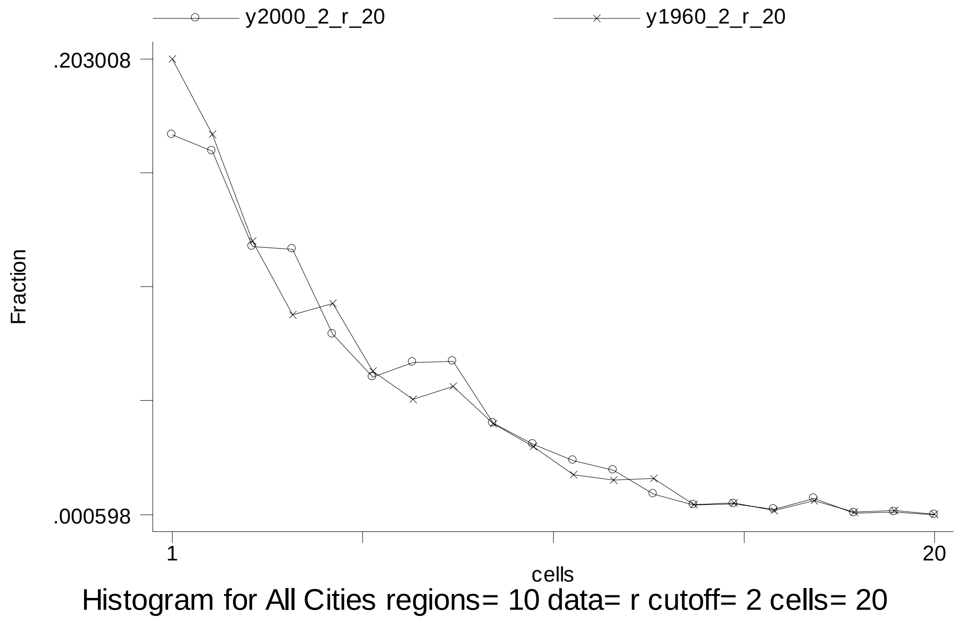


Figure 2. Relative Size Distribution for Cities in All Countries

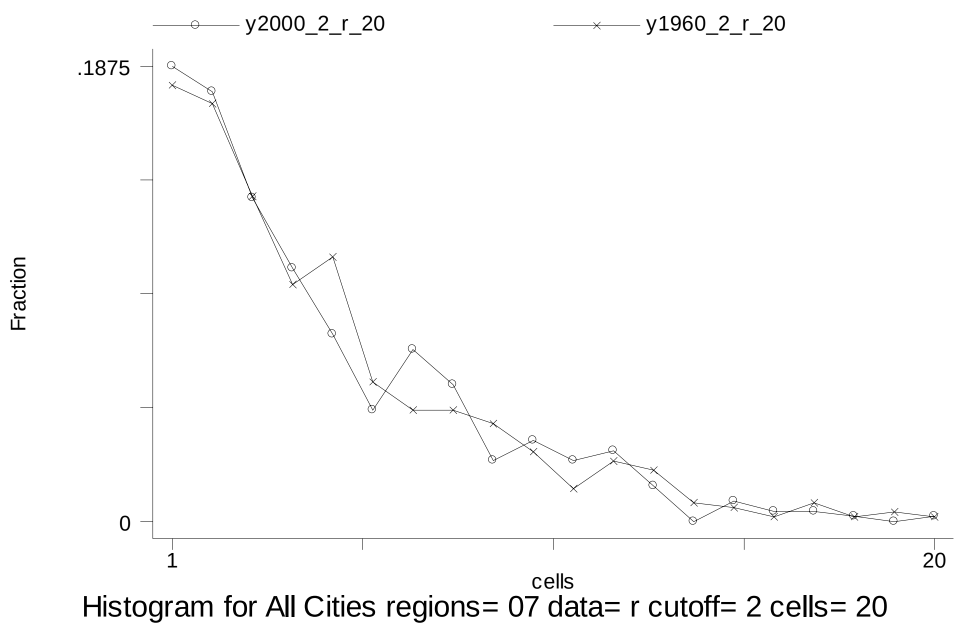


Figure 3. Relative Size Distribution of Cities in Developed Countries

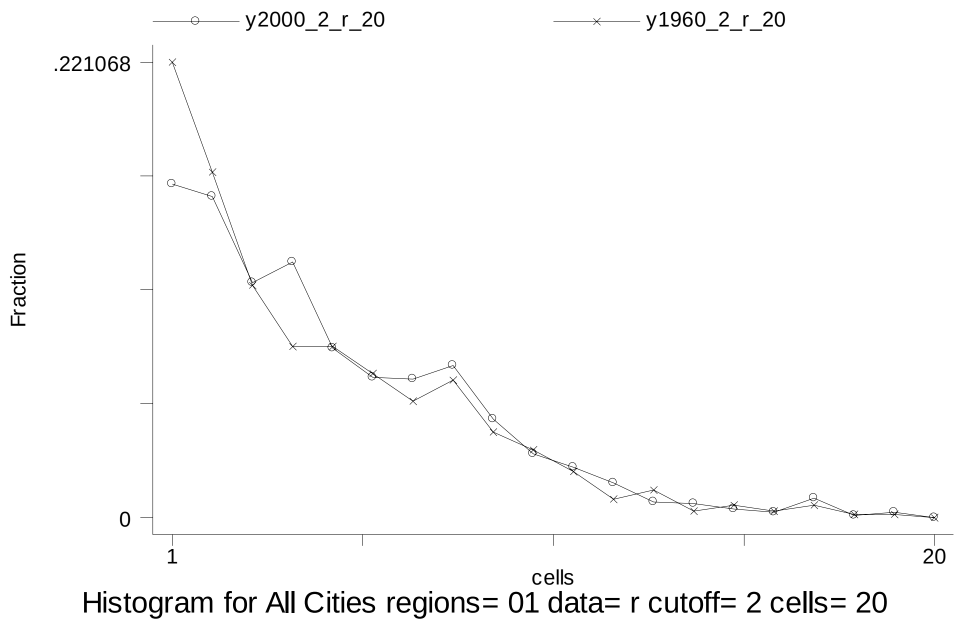


Figure 4. Relative Size Distribution of Cities in Developing and Transition Countries

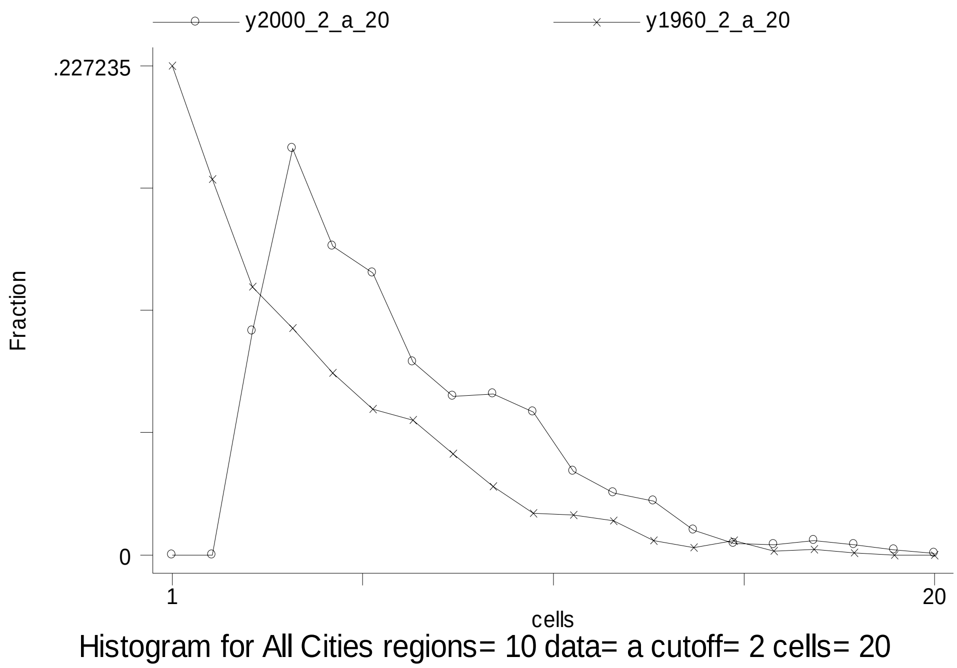


Figure 5. Absolute Size distribution of Cities for All Countries

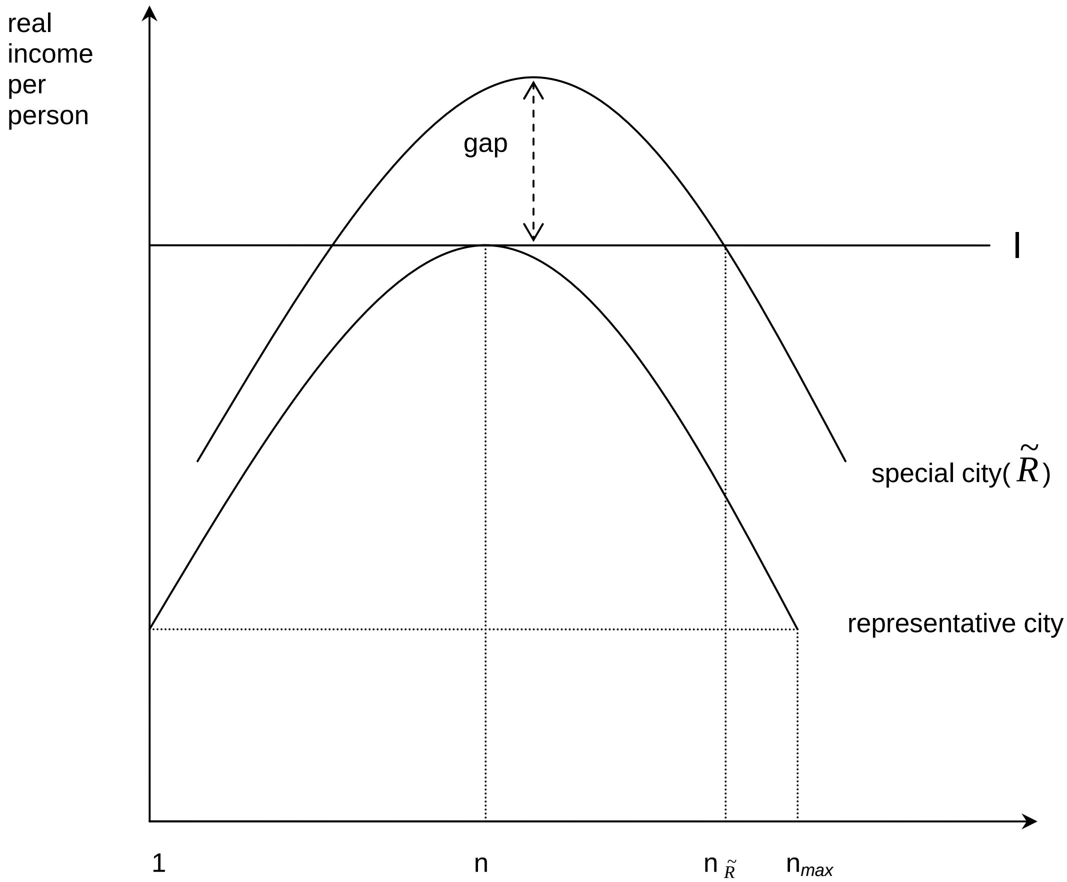


Figure 6. City Size Determination for One City Type

References

Ades, A. F. and E.L. Glaeser (1995), "Trade and Circuses: Explaining Urban Grants", Quarterly Journal of Economics, 10, 195-227.

Arellano, M. and S. Bond (1991), "Some tests of specification for panel data", Review of Economic Studies", 58, 277-297.

Arzaghi, M. and J.V. Henderson (2003), "Why Countries Are Fiscally Decentralizing", Journal of Public Economics, forthcoming.

Au, C-C. and J.V. Henderson (2003), "Estimating Net Urban Scale Economies, with an Application to Evaluating The Costs of China's Migration Restrictions", Brown University mimeo.

Barro, R.J. and J.W. Lee (1996), "International Measures of Schooling Years and Schooling Quality", American Economic Review, 86, 218-223.

Benhabib, J. and M. Spiegel (1994) "The Role of Human Capital in Economic Development", Journal of Monetary Economics, 34, 143-173.

Black, D. and J. V. Henderson (1999), "A Theory of Urban Growth," Journal of Political Economy, 107, 252-284.

______________________ (2003), "Urban Evolution in the USA," Journal of Economic Geography.

Blundell, R. and S. Bond (1998), "Initial conditions and moment restrictions in dynamic panel data models", Journal of Econometrics, 87, 115-143.

____________________ (1998a), "GMM Estimation with Persistent Panel Data: An Application to Production Functions", IFS Working Paper No. W99/4.

Cronon, W. (1991), Nature's Metropolis, Norton and Co.

Davis, J. (2003), "Headquarter Service and Factory Urban Specialization With Transport Costs", BRDC, CES mimeo (JDavis@nber.org).

Davis, J. and J.V. Henderson (2003), "Evidence on the Political Economy of the Urbanization Process", Journal of Urban Economics.

Dobkins, L. and Y. Ioannides (2001), "Dynamic Evolution of the US City Size Distribution in J-F Thisse et al., The Economics of Cities, Cambridge University Press.

Duranton, G. (2002), "A Zipf's Law Model" LSE mimeo, (http://158.143.98.51/~duranton/Papers/Zipf.pdf).

Duranton, G. and D. Puga (2000), "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of City Products", American Economic Review, 91, 1454-1463.

Eaton, J. and Z. Eckstein (1997), "Cities and Growth: Theory and Evidence from France and Japan," Regional Science and Urban Economics, 27, 443-474.

Flatters, F., J.V. Henderson and P. Mieszkowski (1974), "Public Goods, Efficiency, and Regional Fiscal Equalization", Journal of Public Economics, 3, 99-112.

Gabaix, X. (1999), "Zipf's Law for Cities: An Explanation", Quarterly Journal of Economics, 739-767.

Gabaix, X. and Y. Ioannides (2003), "The Evolution of City Size Distributions", Handbook of Urban and Regional Economics, Vol. 4, J.V. Henderson and J-F. Thisse (eds.), North Holland, forthcoming.

Gastil, R. (1978), Freedom in the World.

Glaeser, E., J. Scheinkman and A. Shleifer (1995), "Economic Growth in a Cross-Section of Cities," Journal of Monetary Economics, 36, 117-143.

Grossman, G. and E. Helpman (1991), Innovation and Growth in the Global Economy, MIT Press.

Hamilton, B. (1975), "Zoning and Property Taxation in a System of Local Governments", Urban Studies, 12, 205-11.

Head, K. and T. Mayer (2003), "The Empirics and Agglomeration and Trade", Handbook and Regional Economics, Vol. 4, J.V. Henderson and J-F Thisse (eds.), North Holland, forthcoming.

Henderson, J.V. (2003), "Urbanization and Economic Growth: The So-What Question", Journal of Economic Growth, in press.

Henderson, J.V. and Y. Ioannides (1981), "Aspects of Growth in a System of Cities", Journal of Urban Economics, 10, 117-139.

Henderson, J. V. and R. Becker (2000), "Political Economy of City Sizes and Formation," Journal of Urban Economics, 48, 453-484.

Henderson, J.V., T. Lee and J.Y. Lee (2001), "Scale Externalities in Korea," Journal of Urban Economics, 49, 479-504.

Henderson, J.V. and H.G. Wang (2003), "Aspects of the Rural-Urban Transformation of Countries", Brown University mimeo prepared for Wider Conference in Helsinki on May 29-30, 2003.

Jaggars, K. and T.R. Garr (1995), "Tracking Democracy's Third Wave in the Polity III Data", Journal of Peace Research, 32, 469-482.

Kaufman, D., A. Kraay, and P. Zoido-Lobaton (1999), "Aggregating Governance Indicators", World Bank Policy Research Working Paper #2195.

Krugman, P. (1991), "Increasing Returns and Economic Geography," Journal of Political Economy, 93, 483-499.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R. Vishny (1998), "The Quality of Government", NBER Working Paper #6727.

Lucas, R. (1988), "On the Mechanics of Economic Development", Journal of Monetary Economics, 22, 3-42.

Moretti, E. (2003), "Human Capital Externalities in Cities", Handbook of Urban and Regional Economics Vol. 4, J.V. Henderson and J-F Thisse (eds.) North Holland, forthcoming.

Puga, D. (1999), "The Rise and Fall of Regional Inequalities," European Economic Review, 43, 303-334.

Rappaport, J. (2000), "Why Are Population Flows So Persistent?," Federal Reserve Bank Kansas City mimeo.

Rossi-Hansberg, E.and M. Wright (2003). "Urban Structure and Growth", Stanford University mimeo.

Scotchmer S. (1986), "Local Public Goods in an Equilibrium", Regional Science and Urban Economics, 16, 463-81.

Stiglitz, J. (1977), "The Theory of Local Public Goods", in The Economics of Public Services, M.S. Feldstein and R.P. Inman (eds.) MacMillan Co.

Upton, C. (1981), "An Equilibrium Model of City Size", Journal of Urban Economics, IV, 15-36.

UN (1993), World Urbanization Prospects: The 1992 Revision, NY: United Nations.

Williamson, J. (1965), "Regional Inequality and the Process of National Development," Economic Development and Cultural Change, June, 3-45.

Windmeijer, F. (2000), "A finite sample correlation for the variance of linear two-step GMM estimators", IFS working paper No. W0019, London.

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

Appendix A. Table A1. 1960 versus 2000 World Cities

World

	1960			2000		
	count	mean	share[1)]	count	mean	share[1)]
12,000,000 <= x	1	14,200,000	2.36	11	15,800,000	9.67
4,000,000 <= x < 12,000,000	17	6,448,941	18.31	42	7,040,834	16.41
2,000,000 <= x < 4,000,000	32	2,600,125	13.89	116	2,780,428	17.90
1,000,000 <= x < 2,000,000	65	1,374,756	14.92	247	1,372,582	18.81
500,000 <= x < 1,000,000	160	702,319	18.76	355	703,095	13.85
250,000 <= x < 500,000	262	344,754	15.08	646	349,745	12.54
100,000 <= x < 250,000	660	151,289	16.67	1,240	157,205	10.82
total	1,197	500,340	100.00	2,657	678,218	100.00

1) a ratio of total population in the group to total population of cities with >=100,000

LAC

	1960			2000		
	count	mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	3	16,100,000	18.44
4,000,000 <= x < 12,000,000	4	5,453,500	38.15	5	6,804,200	12.95
2,000,000 <= x < 4,000,000	1	2,032,000	3.55	14	3,091,214	16.47
1,000,000 <= x < 2,000,000	7	1,292,571	15.82	29	1,416,187	15.63
500,000 <= x < 1,000,000	10	669,800	11.71	42	721,139	11.53
250,000 <= x < 500,000	21	356,140	13.08	85	332,659	10.76
100,000 <= x < 250,000	69	146,547	17.68	244	153,142	14.22
total	112	510,560	100.00	422	622,616	100.00

Sub-Saharan Africa

	1960			2000		
	count	mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	1	13,400,000	9.74
4,000,000 <= x < 12,000,000	0	.	0.00	1	5,064,000	3.67
2,000,000 <= x < 4,000,000	0	.	0.00	11	2,659,104	21.22
1,000,000 <= x < 2,000,000	1	1,147,000	8.45	21	1,449,632	22.09
500,000 <= x < 1,000,000	5	661,000	24.35	25	725,510	13.16
250,000 <= x < 500,000	10	337,390	24.85	44	350,925	11.20
100,000 <= x < 250,000	38	151,303	42.35	163	159,861	18.91
total	54	251,397	100.00	266	518,117	100.00

North Africa & Middle East

	1960			2000		
	count	mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	0	.	0.00
4,000,000 <= x < 12,000,000	0	.	0.00	4	6,671,750	24.91
2,000,000 <= x < 4,000,000	1	3,712,000	20.20	7	2,747,710	17.95
1,000,000 <= x < 2,000,000	3	1,465,333	23.93	13	1,377,569	16.72
500,000 <= x < 1,000,000	4	658,500	14.34	23	675,811	14.51
250,000 <= x < 500,000	5	321,462	8.75	43	368,558	14.79
100,000 <= x < 250,000	39	154,421	32.78	75	158,702	11.11
total	52	353,302	100.00	165	649,235	100.00

South Asia

	1960			2000		
	count	Mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	3	14,400,000	19.06
4,000,000 <= x < 12,000,000	2	4,780,000	21.38	7	7,534,286	23.22
2,000,000 <= x < 4,000,000	1	2,283,000	5.10	11	2,510,000	12.15
1,000,000 <= x < 2,000,000	6	1,402,500	18.82	29	1,308,207	16.70
500,000 <= x < 1,000,000	5	738,600	8.26	39	693,668	11.91
250,000 <= x < 500,000	27	360,845	21.79	78	341,140	11.71
100,000 <= x < 250,000	72	153,166	24.66	75	158,757	5.24
total	113	395,768	100.00	242	938,668	100.00

East Asia and Pacific

	1960			2000		
	count	mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	1	12,900,000	3.14
4,000,000 <= x < 12,000,000	2	7,554,000	13.71	11	7,561,091	20.26
2,000,000 <= x < 4,000,000	9	2,488,556	20.32	28	2,726,300	18.60
1,000,000 <= x < 2,000,000	10	1,291,900	11.72	78	1,390,749	26.43
500,000 <= x < 1,000,000	52	713,229	33.65	101	699,101	17.20
250,000 <= x < 500,000	42	359,528	13.70	101	376,109	9.25
100,000 <= x < 250,000	49	155,285	6.90	131	160,328	5.12
total	164	672,079	100.00	451	910,140	100.00

Soviet Bloc

	1960			2000		
	count	Mean	share	count	mean	share
12,000,000 <= x	0	.	0.00	0	.	0.00
4,000,000 <= x < 12,000,000	1	6,170,000	9.46	2	7,227,000	9.88
2,000,000 <= x < 4,000,000	2	2,905,000	8.91	5	2,525,600	8.63
1,000,000 <= x < 2,000,000	6	1,316,500	12.11	24	1,292,875	21.21
500,000 <= x < 1,000,000	25	708,992	27.18	33	656,763	14.82
250,000 <= x < 500,000	32	350,635	17.21	97	343,055	22.75
100,000 <= x < 250,000	113	144,952	25.12	216	153,740	22.70
total	179	364,267	100.00	377	387,980	100.00

Developed Countries

	1960			2000		
	count	mean	share	count	mean	share
12,000,000 <= x	1	14,200,000	4.89	3	18,700,000	11.01
4,000,000 <= x < 12,000,000	8	7,122,500	19.67	12	6,631,417	15.59
2,000,000 <= x < 4,000,000	18	2,609,445	16.22	40	2,854,854	22.37
1,000,000 <= x < 2,000,000	32	1,422,974	15.72	53	1,361,549	14.14
500,000 <= x < 1,000,000	59	698,786	14.23	92	720,587	12.99
250,000 <= x < 500,000	125	334,416	14.43	198	345,951	13.42
100,000 <= x < 250,000	280	153,395	14.83	336	159,196	10.48
total	523	553,786	100.00	734	695,421	100.00

Appendix B

Data and Sources

link to

http://www.econ.brown.edu/faculty/henderson/worldcities.html

Means and Standard Deviations[28]

	Mean	**Standard Deviation**
Growth in city numbers	.13	.30
Growth in city sizes	.26	.29
Growth in national population	.19	.12
% adults with secondary educ.	.19	16
Rural/urban population	7.7	11
openness	27.8	16
democracy	0.08	7.9
Change in democracy	1.1	5.2
Change in market potential	.28	.20

[28] Calculated for sample for growth rate in numbers of cities equation.